Volume 46 Number 2 June 2022 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

IoT-Enabled Remote Monitoring Techniques for Healthcare Applications – An Overview

Shajulin Benedict
E-mail: shajulin@iiitkottayam.ac.in or shajulinbenedict@mytum.de, www.sbenedictglobal.com
Indian Institute of Information Technology Kottayam, Kerala, India
Guest Professor, Technical University Munich, Garching, Germany

Overview paper

Keywords: healthcare, IoT, remote eHealth, social health

Received: January 15, 2022

IoT-enabled remote healthcare monitoring applications have surged against countless other technologies to assist patients. Recently, the healthcare sector has sought a devastating escalation in augmenting appropriate monitoring technologies for effective remote monitoring and improved diagnostics during the pandemic era – i.e., the existing methods need to be revamped with sophisticated technologies. In this paper, remote monitoring techniques proposed for healthcare applications and their challenges are surveyed. In addition, a healthcare architecture for effective remote monitoring is explored. It was observed that most of the solutions focused on the application of edge analytics and deep learning mechanisms. The review aimed at guiding healthcare practitioners and developers to understand the pitfalls of existing approaches and to innovate solutions in newer dimensions.

Povzetek: Podan je pregled IoT sistemov za pomoč v zdravstvu.

1 Introduction

As the COVID-19 pandemic goes ahead offering limited access to hospitals, there is a near certainty among patients that they would be proactively prevented from diseases. The pandemic created direct and indirect impacts on chronic patients who visit hospitals. However, it ignited innovations in the existing aid approaches – i.e., the growth in the healthcare industry has recently magnified.

Several healthcare applications evolved to address the needs of the patients by evaluating the sensor values mounted on the body, in the home, or in environments. For instance, mobile applications that advise proper diet or telemedicine [76, 10], IoT-enabled sensor applications that monitor the health status of patients or remind appropriate medicines [23, 48, 70], AI-assisted applications that check the sleep issues of patients, online counseling, and so forth have marked upgraded functionalities in the recent past.

In spite of inviting improvements and enabling strong technical requirements, the ICT solutions of the healthcare sector have not become mature to experiment with them on patients/individuals. A few notable challenges that exist in the domain include:

1. *Patient Data security:* Data theft, hacking, and the other security breaches in the patient data [5, 41] hamper the realization of remote patient monitoring.
2. *Availability:* Most remote patient monitoring IoT-enabled applications utilize cloud databases or service-oriented infrastructures. Availability of these compute resources is challenged due to the emergence of modern execution models such as serverless compute environments.
3. *Device inter-operability:* There are several proprietary sensor devices involved in accomplishing healthcare applications. Managing the data transfers and connecting them using appropriate communication protocols are challenging aspects for researchers and industry practitioners.
4. *Cost/Performance efficiency:* The cost and performance of applications grow into a tradeoff that needs to be elegantly handled in healthcare applications. For instance, scaling applications, enhancing privacy [27], or improving the reliability of applications could directly or indirectly influence the operational costs of applications.

A few researchers have implemented healthcare applications using IoT, blockchain, and hierarchical-computing technologies. However, there exist several limitations and research gaps to be fulfilled for improving the diagnostics for real-world scenarios.

The major contributions of this article are listed as follows:

1. To illustrate a remote IoT-enabled health monitoring architecture that details the inner functional details;
2. To critically review on the existing remote monitoring healthcare techniques; and,

3. To study the impact of datasets, devices, and IoT-enabled applications.

The rest of the article is organized as follows: in Section 2, survey works related to the healthcare domain are discussed; in Section 3, a generic IoT-enabled healthcare application architecture is revealed; in Section 4, a few available remote monitoring techniques and their challenges are presented; Section 5 reveals a few applications that are widely utilized in the market; and, finally, research directions and conclusions are expressed in Section 6, the last part of the article.

2 Related work

Remote monitoring of the health status of patients has heightened the utility rate of healthcare applications since the inception of technologies such as IoT, serverless, edge, blockchain, and so forth. For instance, an increase in the number of filing patents and publications manifests the growth pace of healthcare applications [22, 42, 10].

The techniques involved in the remote monitoring of patients or individuals have expanded in various implementations – for example, telemedicine, clinical trials, online counseling, psychological monitoring [43], and so forth, have diverged the markets using wearables, mobile phones, or implantable sensor units. Particularly, the focus on the remote monitoring of the health status of patients/individuals involved elderly [61, 75], chronic patients, or preventive care individuals such as working professionals. Unique methods need to be adopted for the efficient handling of healthcare applications.

In the past, a few survey works were carried out by researchers to study healthcare applications. For instance, authors of [35] have surveyed the healthcare applications such as emergency monitoring, gesture determination using mobile phones, knowledge-based decision support systems, and so forth. Similarly, researchers of [9] have explored the available communication protocols and technologies for healthcare applications. However, these survey works did not focus on remote monitoring health applications.

The need for delving into the remote monitoring techniques that have been practiced in recent years is multifaceted:

1. To suggest the directions of research and the possible improvements in the existing monitoring techniques;

2. To promote newer insights while designing remote monitoring healthcare architectures incorporating technologies such as serverless or AI methods;

3. To provide the amount of works/developments or competitors before investigating time for their innovations; and, so forth.

This work focuses on investigating the remote IoT-enabled health-monitoring techniques and available solutions.

3 IoT-enabled healthcare architecture

This section explains the generic remote health monitoring healthcare architecture and the significance of the components involved in it. Figure 1 illustrates the architecture and the functions of components. In addition, the section highlights a few notable existing healthcare monitoring architectures/frameworks.

3.1 Major components

The major components involved in the architecture and their functionalities are listed in the following paragraphs.

3.1.1 User-interface

Users, mostly patients and their well-wishers, are prompted with ease-to-use interfaces. The major features that an interface includes are:

– *Seamless Responsive Designs*: The users or patients prefer to utilize multi-size screens of varying gadgets such as mobile devices, laptops, or servers. The GUI design of remote monitoring systems, in general, includes a unique design with minimal variations to have a flexible layout of visibility features at the end devices.

  In doing so, the contents and visual representation of patient information are scalable with respect to the contents and screen size of the gadgets involved in the application.

– *Automated Bots*: The web designs of the remote monitoring healthcare applications involve automated software robots named bots. Bots are, in general, a piece of software instance that performs automated execution of tasks [63]. In healthcare applications such as smart e-consulting or e-counseling, AI-assisted software robots are implemented to simplify the processes involved in performing tasks. These bots quicken the processes such that registration of patients to appropriate hospitals and suitable available doctors happen in a short span of time.

  In the past, authors of [67] have studied the importance of chatbots in assisting patients. These authors have studied the input data formats such as voice, text, or video of chatbots while interacting with patients; also, they have identified the requirements of natural language processing, reasoning, and so forth for elegantly automating the assisting processes.

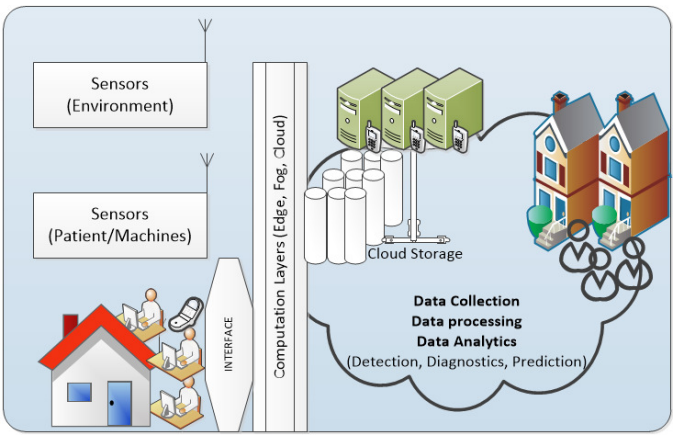


Figure 1: Generic IoT-enabled Healthcare Monitoring Architecture.

– *Multi-factor Authentication*: The privacy of medical data or personal records is crucial in healthcare applications [27, 65]. This feature prevents users' data from being protected from hackers or malicious suspect devices. It is very unlikely that multiple pieces of programs that are responsible for independent authentication mechanisms such as username/password, iris recognition, biometric sensor units, and so forth, will be trapped by the hackers.

  In a few IoT-enabled remote monitoring healthcare applications, a single sign on (SSO) feature is implemented to improve the user-authentication experiences by seamlessly connecting with multiple cloud-enabled healthcare services.

– *Virtual Reality*: The user interface of healthcare applications needs support for virtual reality. In short, virtual reality incorporates human senses such as touch, sight, hearing, taste, and smell into programs and supportive hardware devices to provide a virtual environment. Applications such as remote monitoring and counseling require a realistic environment of patient diagnostics involving doctors from varying locations. The user interface of such applications, therefore, shall include these features for robustness and liveliness.

3.1.2 End devices

Increasing the number of end devices and sensor units for continuous medical diagnostics or disease detection has prompted for delivery of superlative care by doctors to patients in recent years.

In fact, the detection of diseases evolves based on the screening processes of healthy or non-healthy individuals. It requires skillful listening to abnormalities if any. The major purpose of detecting diseases is to identify the risk indicators concerning diseases or probabilistic diseases. By doing so, individuals could reduce the long-term risk factor that might arise in the future. Whereas, diagnostics is the confirmation of the availability/non-availability of a particular disease.

IoT-enabled systems require real-time sensed data from different patients/individuals to detect diseases. By this, the spread of the diseases is prevented. More importantly, in order to perform diagnostics of diseases, IoT devices require more prominent authenticated accurate data for confirming the diseases that are often represented in a layered structure[69]. In the past, researchers have utilized biomarkers to dictate the confirmation of diseases.

Recently, researchers have coined the term Internet of Medical Things (IoMT) to establish connected IoT medical devices by extending the internet to tiny sensing objects. The end devices that are utilized in the healthcare application domain for remote monitoring the health status of individuals can be classified into 6 categories as discussed below:

1. *In-body devices*: Implanted sensors or medical things constantly monitor and add value to remote monitoring health applications. Many actuating things are often implanted in human bodies to stimulate near-failure organs. For instance, devices such as neuro-stimulators, cardiac defibrillators, insulin pumps, cochlear stimulators, and so forth are implanted to the patients for stimulating chemical actions for medical benefits. In recent years, the utilization of insulin pumps has tremendously increased as diabetic patients are prone to reach an uncontrolled augmentation of insulin.

2. *On-body devices*: The majority of the devices under the category of IoMT are wearable devices. These devices are attached to a human body in the form of wristwatches, dresses, rings, or adorable devices to frame wireless body area network (WBAN) [26]. These devices follow communication protocols such as IEEE802.15.4, which has a short-range communication medium, to emit less energy. Most commonly available on-body devices include accelerometers which are designed to measure acceleration

forces; gyroscopes which measure the angular rate or orientation angles; GPS sensors that sense the latitude/longitude of locations; heart-rate sensors which measure the heart rate of humans; and pedometers that count the steps taken by patients or athletes while running or walking.

3. *Portable devices*: Medical things that are mobile in nature are termed portable devices. These devices are often directly connected with cloud server instances or fog nodes. Devices that constantly monitor the blood pressure or insulin level of patients are examples of portable sensor devices.

4. *Static devices*: Devices such as temperature sensors or air pollution sensor nodes [64] that are attached to environments such as smart homes/cities are denoted as static nodes. For instance, sending alarms or short messages to doctors/relatives of patients for detecting pneumothorax in X-rays is fixed to x-ray devices. The connectivity of these devices is much more reliable when compared to portable devices. Besides, the inherent accuracy of these devices is stable in healthcare applications.

5. *Ambulatory devices*: IoT-enabled devices fitted to mobile vehicles such as ambulances are peculiar as they have to consider the real-time delivery of data to the hospital clouds servers. In spite of being error-prone due to the frequent hops in communication channels, the devices should deliberately handle switching-on stations without errors. Reliable measures need to be practiced so that patients could be saved due to the right medical prescriptions based on the sensor data.

6. *Hospital devices*: IoT has been utilized in hospitals not only to detect and monitor patients' temperature, blood pressure, and so forth, but also to locate medical kits. For instance, a 1000 bed hospital usually has over 200 wheelchairs. It is also mandatory to locate patients considering the safety measures of patients – i.e., an asthma patient needs pollution-free environment; or, patients wanting for medical equipment such as defibrillators, nebulizers, oxygen pumps, and so forth, have to be guided to the nearest available wards in the hospital. Clearly, IoMT is beneficial to quickly identify the location of medical devices and serve patients/hospitals. Figure 2 pictorially represents the devices utilized in the remote healthcare applications.

The most commonly available IoT-enabled sensor devices for measuring health-related parameters include i) glucose monitors, ii) temperature sensors, iii) heart-rate monitors, iv) oxygen pulse monitors, v) electromyography sensors (ECG) for heart-care checks, vi) wheeze anomaly detection sensors, vii) movement disorder checks, viii) stress indicator, ix) posture indicator, x) lung status indicator, and so forth.

3.1.3 Edge/fog layers

Edge and Fog nodes increase the capability of user experience in the healthcare sector. These networked nodes provide text, video, or image analytics with the help of robust AI mechanisms and sensor devices. For instance, monitoring patients in ambulances using edge devices or mobile devices assist technicians to deliver first-aid services with the advice of remotely available doctors before the patient was reached in hospitals; virtual reality enabled operating rooms are becoming a new normal for practicing edge supported operations to patients.

Edge and fog nodes reduce latency when compared to cloud-level analytics. This real-time delivery of findings is one of the major characteristics of healthcare applications.

3.1.4 Cloud services

A large volume of scalable computations and analytics of healthcare data is possible in cloud infrastructures. In fact, the recent newer cloud computing execution model such as serverless clouds has manifested the feasibility of providing reduced cost to the users. Typically, in healthcare applications, periodic monitoring of health-check devices is practiced. The frequency of remote monitoring is lower in some applications where the patients involved in the processes are quite normal. If a serverless execution model is not provided for such applications, the user could lead to huge costs due to the utilization of cloud server instances.

In addition, the sensor data could be of higher size ranging from terabytes to zettabytes in healthcare applications, especially when video analytics of operations were involved. Obviously, there is a dire need for an automated scalable environment for healthcare applications.

It could be noticed that to improve the prediction accuracy in remote health monitoring applications cloud services requires long-term analytics obtained from a larger dataset. For instance, bots increase accuracy by improving text analysis strategies; analytics of images is required for classifying, learning, or predicting health-related symptoms such as tumor analysis, cancer analysis [50, 21], eye-retinal failure detections, and so forth; analytics of videos is required for learning the emotions of patients, neurological disorders, and sleep disorders.

3.1.5 Hospitals and doctors

Remote healthcare monitoring architectures enable an active involvement of doctors or hospitals. For instance, mobile phones engage doctors and individuals; and, associated servers involve hospital authorities for decision-making processes. These mobile devices and gadgets connect doctors/hospitals in a remote fashion to strengthen the assurance of patients.

Observing the most existing remote monitoring applications, doctors and hospitals are connected for providing online prescriptions/diagnostics [18], scheduling doc-

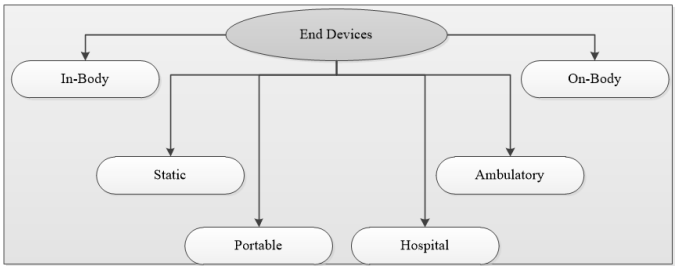


Figure 2: Devices Utilized in Remote Health-care Applications.

tor appointments, sharing files/videos, communicating effectively, and so forth.

3.2 Existing architectures and future insights

Most of the existing architectures/frameworks developed for monitoring patients has shown similarities with the generic architecture. Table 3.2 reveals the key components of generic architecture adapted in the existing architectures and the uniqueness found in them. For instance, authors of [49] developed healthcare monitoring system for Saskatchewan, a specific region in Canada; authors of [62, 16] included alerts, notifications, or early warning systems for agile operations. In the table, "Y" resembles "YES". A few architectures submitted medical data to cloud services using publish/subscribe approaches ([56, 29]). There exist a few works in the healthcare monitoring domain that improved the security aspects of medical data using blockchains ([59]); and, the data analytics of healthcare data using novel deep learning algorithms ([33]). Additionally, frameworks that improve the human behavior/attitude by assessing the mental states had opened a novel research dimension in the healthcare sector.

Apart from the existing healthcare monitoring frameworks/architectures, the following modules or techniques could be incorporated in the near future architectures of the remote healthcare monitoring systems:

1. Serverless-based cloud execution model – It is a known fact that cloud services or the associated resources are not required throughout the entire period of healthcare applications. For instance, a few applications might only update the health status of remote patients on cloud storage devices once in a week. Considering the infrequent utilization of the services, it is sufficient to adopt serverless-based based cloud execution models for these applications;

2. Blockchain-enabled medical data protection systems – Tampering medical data or records is not tolerated in many countries or policies. Blockchain-enabled frameworks/architectures could solve these underlying issues by involving multi-party stakeholders;

3. End-to-End auditing framework – A very few efforts have been attained in the past to audit the functions or applications at end-to-end level. The remote monitoring healthcare applications execute logics that are computed on multiple nodes or servers of varied computational capabilities. By diligently auditing the services, the optimization of resources or efficiency of computations could be improved.

4 Remote monitoring techniques – taxonomy

The remote monitoring techniques could be classified depending on several factors, as follows: i) based on the utilization of different computing systems, ii) based on importance to serve tasks (priorities), iii) based on communications involved, iv) based on accessibility features, v) based on level of intelligence, vi) based on specified metrics, and vii) based on focused delivery. This section explores the most available IoT-enabled remote monitoring techniques and their characteristics in detail.

4.1 Computing continuum assisted techniques

Remotely monitoring the health-related parameters of healthcare applications involves varied computing devices surpassing from battery-operated sensor devices to scalable clouds. Depending on the utility pattern of computing nodes, the remote healthcare monitoring applications could be either performance-efficient or cost-efficient.

In most healthcare applications, edge nodes are predominantly applied for the immediate analysis of sensor data. Authors of [14] have applied edge devices to register patients, authenticate them, and raise alarms using blockchains. Similarly, in [51], authors have designed a BodyEdge platform that connects most of the patients' sensor data with edge nodes for improving the scalability; authors of [47, 73] have studied the impact of energy efficiency, and so forth. These edge devices are often battery-operated with limited processing capabilities. Edge nodes, in fact, deal with decentralized medical data obtained from the nearby sensor modules.

In recent years, a few implementations including containers such as docker containers in edge nodes have been

Table 1: A Comparison of Existing Healthcare Monitoring Architectures.

Components	[49]	[62]	[56]	[16]	[50]	[29]	[73]	[59]	[33]	[38]
UI	Y	Integration Units	Y	Y	Integration Units	Y	-	Integration Units	Y	-
End Device	Y	Y	Sensors (Wearables)	Sensors (Wearables)	Y	Sensors (Wearables)	Y	Y	Y	Y
Edge/Fog	Hierarchical (Mesh)	Fog	-	Fog	Y	-	-	Y	Edge	-
Cloud services	Y	Y	MQTT	Y	Y	MQTT	Y	Y	Y	Y
Hospital/ Doctors	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Uniqueness	Saskatchewan (Canada)	Alerts	Publish/ Subscribe	Early Warning	Cancer care	Elderly	ECG Compression	Blockchain security	Deep Learning	Persuasive Technology

adopted to improve the lightweight migrations and executions. For instance, authors of [34] have utilized raspberry pi-based architecture to execute containers on edge nodes for enabling the visual representations of healthcare applications. Additionally, RFID clusters placed in the edge layer improve the connectivity of sensor nodes in some healthcare applications. Authors of [3] have utilized active RFID devices to establish RFID clusters and monitor patients at home.

Compared to Edge nodes, Fog nodes handle a large volume of data collected from multiple sensor modules of the vicinity [8]. Typically, fog nodes are responsible to process sensitive data and autonomously decide on performing certain actions without the knowledge of the entire status of applications available in the cloud.

Clouds, on contrary, are often utilized by many healthcare applications. They are highly scalable with the capability to adapt to the increasing data size and computational requirements – i.e., the storage space for handling a large volume of data in the cloud is very high. Clouds are also capable of quickly processing/transferring the analysis observed from the large sensor data. In the recent past, authors of [52] have applied a cloud ecosystem to provide health diagnoses among students. Similarly, authors of [15] highlighted the importance of cloud services for analyzing health data in a hierarchical fashion. Obviously, clouds, due to the capability of their larger processing power, remote diagnostics within a limited time frame is possible. However, clouds are powered on for the entire duration of the remote monitoring healthcare applications.

Overriding the continuous execution of cloud instances, serverless cloud environments spawn compute nodes based on the trigger received from sensor nodes. Serverless doesn't mean that the computations are performed without servers. Rather, it is an execution model of clouds where the users are benefited from the limited utility of servers. Depending on the availability of servers, the serverless framework could either power on and boot up the machines or it could utilize the available idle server instances.

The scalability component of the different continuum of computational units involved in the remote monitoring healthcare applications varies – i.e., increasing the number of server instances and cloud-enabled services, scales the spanning of applications across the globe. Similarly, it could be noticed that the cost of including only serverless functions for healthcare applications would drastically reduce their operational/infrastructural costs of them. On contrary, this method reduces the serviceability when compared to involving the combination of edge, fog, cloud, and serverless environments.

4.2 Priority-enabled remote monitoring

Remotely monitoring the health status of individuals is clouted by the priority in attending tasks/requirements. For instance, a healthcare monitoring application should prioritize the monitoring features involving healthcare workers such as doctors and monitoring patients – i.e., in a healthcare remote monitoring system, monitoring the health status of healthcare workers, especially in isolated wards, is quite important than serving the other devices or individuals.

Also, cloud services have to be prioritized with sufficient intelligence in an automated fashion. In general, the requests are served on a first-come-first-serve basis. However, policies and protocols need to be provided in the remote healthcare monitoring applications so that notified prioritized activities are executed.

Depending on the tasks specified in Figure 4.7, priorities could be set up for remotely monitoring the healthcare applications, as follows:

1. *Critical Tasks* Availing a secured data channel to express the health status of patients to hospitals/doctors is one of the crucial tasks in remote monitoring healthcare applications. Healthcare applications have to prioritize these tasks and execute them within the limited computational node availability.

2. *Periodic Tasks* There exist tasks in the remote healthcare sector to pursue routine checks and monitoring mechanisms. Routine checking is not only specific to the patients but also to the machines involved in sending data. For instance, monitoring the patient's temperature, pressure, glucose, and so forth is a common practice in the health sector domain. However, monitoring the health status of the robotic machines involved in the operations or learning the failure cases of machines using IoT-enabled systems are crucial tasks.

3. *Preventive Tasks* A few healthcare tasks are preventive in nature. Mechanisms involved in preventing medi-

cal failures or diseases require lots of artificial intelligence embedded in the system [28, 46]. For instance, learning algorithms such as random forests, support vector machines, linear regressions, deep learning models, and so forth could be applied to preventing the patients from severity in the impacts. A few algorithms such as convolutional neural networks are applied to early detect cancer symptoms from images relating to skin, lungs, and brain.

4. *Federated Tasks* In remote monitoring, collaborative engagement of multiple peers or medical things is involved. Examples such as analyzing the patient similarity or emotion analysis require collaborative learning involving sensors from various hospitals and organizations. A federated learning mechanism needs to be incorporated into these systems. In federated learning processes, the global learning models are refined by the distributed local learning models. This way, a robust process of targeting multiple machines and organizations is possible for healthcare applications. Applications such as collaborative drug discovery with the involvement of sensor data and machines from multiple organizations are carried out using federated mechanisms.

5. *Alarming/Actuating Tasks* Healthcare monitoring also involves tasks to send notifications or alarms by actuating appropriate actuators. Hospital beds might have to be tilted to 60 degrees to comfort patients, especially for patients with Parkinson's disease. Similarly, eye-related operations are supported with laser positioning equipment in eye hospitals. Based on the feedback given by the doctor, the position of the lasers has to be accurately organized in real-time.

6. *Educational Tasks* Providing training or awareness often prevents the spread of diseases such as malaria, COVID-19 [54], and so forth. Tasks relating to awareness have increased due to the inclusion of mobile devices and mobile health applications in recent years. Although these tasks are not crucial to execute in real-time, they have to get a wider reach with minimal costs and failure rates. Additionally, these tasks have to support multiple formats such as audio, video, or texts that choose the most appropriate protocols. These tasks have to increase the public reach in the form of successful campaigns incorporated in wearable gadgets.

Table 2 reveals the comparison of tasks in the remote monitoring healthcare domain. The importance of certain characteristics is specified in the star ratings for tasks. It could be observed that the periodic tasks have to be cost-efficient whereas critical tasks must be completed within a specific time-bound. The critical tasks have to provide more priority to performance efficiency than cost efficiency as they are crucial life-saving tasks. Similarly, educational tasks or alarming tasks aim at reaching a larger mass of connected people or automated systems for delivering messages.

4.3 Communication-specific RM

The remote monitoring of health applications requires an apt selection of communication protocol standards for connecting sensors. IoMT devices are connected using WIFI, Bluetooth4.0, GPS, infrared, Zigbee, and so forth (see Figure 4.7).

High bandwidth requirements are often the scenario in healthcare-related applications, especially when operations were held by doctors using videos or robotic engines. To compensate the high-speed bandwidth requirements, 5G, NB-IoT [12, 57], or similar connections are often set up in the modern hospital premises. The utility of 5G is immense: for instance, authors of [11] have studied the importance of 5G and the design criteria such as channel selection options, bandwidth efficiency, and so forth, in the healthcare applications. Particularly, narrow-band communications for healthcare have profoundly been utilized in the healthcare sector. The main reason is that the NB-IoT communication medium could provide low-energy support to sensors. However, due to the limited bandwidth support and security breaches [59], 5G overrides NB-IoT. There exist a few works relating to NB-IoT in the past. For instance, the authors of [31] studied the uplink/downlink transmission efficiencies of healthcare sensor nodes when NB-IoT was applied as communication links. A few researchers have improved the transmission efficiency of communication links by designing software-defined networks. In the work by Farag et al. [58], the authors have designed an SDN approach to channelize sensor nodes for efficient communications with limited traffic delays.

In addition, ISO/IEEE 11073 protocols are imposed on healthcare devices for establishing efficient secure communications and the delivery of data over these secured connections. Authors of [79] implemented the IEEE 11073 protocol to integrate tiny biosensors to the cloud using Constrained Application Protocol (CoAP).

The important features of the communications involved in the remote health monitoring applications include:

1. *Unique identification*: Several devices might intend to connect to each other or the cloud using various means of connectivity options such as wifi, Bluetooth, or infrared. A unique identification mechanism is required in healthcare applications to avoid the wrong delivery of insights to medical practitioners or users.

2. *Interoperability feature*: The devices manufactured from different vendors such as Samsung, OPPO, cisco, intel, and so forth, might need robust interoperability features [55] so that the healthcare applications are effectively laid in a vicinity. To ensure interoperability, the exchange of data in a specified format and the conversion of protocols are mandatory processes. There exist a few protocol conversion mech-

Table 2: Comparison of Priority-Enabled Tasks – Characteristics.

Tasks	Time/Speed	Cost Efficient	Performance Efficient	Wider Reach
Critical	*****	*	****	***
Periodic	*	****	**	**
Preventive	**	*	****	***
Federated	*	***	***	****
Alarming	***	**	***	*****
Educational	**	****	**	****

anisms which could be incorporated into the gateway devices or edge devices for solving the interoperability issues.

3. *Cooperative Aspect*: In some cases, cooperative decision-making features and collaborative involvement of devices are required to accomplish the remote healthcare monitoring of patients. For instance, studying the behavior of patients could be analyzed by sensing data across hospitals. Here, an inter-hospital communication system that augments the collaborative learning of patients is desired.

   Similarly, sharing the findings of patient information across multi-specialty hospitals also claims for a cooperative mechanism in the healthcare sector.

4.4 Accessibility-specific RM

Access to healthcare services is bound to financial capabilities. Accessing international medical equipment or hospital facilities requires uninterrupted services with the ability to audit the necessity of facilities/equipment. In addition, healthcare services should automatically assign scalable computational units or storage units for the delivery of medical assistance.

Depending on the accessibility feature of remote monitoring assistance, the healthcare services could be classified into local or global remotely accessible monitoring services. In local accessibility setup, the underlying medical services comprise intra-hospital medical things such as oximeter measurements, blood pressure monitoring services, and so forth. In a global accessibility set up, the services involve inter-hospital services and doctors for monitoring the health status of patients. In general, the global accessibility services are meant for collecting the health status of pandemics such as COVID-19, examining infectious diseases, analyzing the causes of deaths across countries, and so forth.

4.5 Intelligent level specific RM

Associating artificial intelligence to the internet of medical things is quite important for automating processes and equipping with more accurate results – an option to deliver proactive healthcare to patients/users. An array of applications have evolved in the recent past with the inclusion of AI methods for assisting healthcare services. For instance, predicting the hospital risks in a remote fashion or developing AI-powered robots to assist patients have been well appreciated among the healthcare communities.

Technologies such as computer vision, which embodies AI into it, obtain information from images or video frames to provide meaningful insights. For instance, performing remote clinical trials could be effectuated using computer vision. Similarly, AI-driven big data processing of medical machines [33] or patients encourages multi-specialty hospitals or concerned doctors to proactively strengthen the prescriptions with utmost accuracy.

Multi-label classification-based AI methods, which utilize robust learning algorithms, improve the decision support systems. These approaches promote automation in remote health monitoring applications.

Additionally, tailor-made solutions could be adopted using ontologies [2] and semantic technologies [68, 78] for enabling health successes. For instance, suggesting the timing and duration of regular exercises, closely monitoring the performance of healthcare, and so forth.

The incorporation of AI in remote health monitoring applications increases the revenue due to the involvement of many doctors or hospitals. In addition, it improves the cost and accuracy efficiency of patients/users. For instance, authors of [72] provided an approach to handle the trade-off between sensing health details and advising remedies irrespective of the less availability of health datasets using adversarial sensing method.

4.6 Performance metric-specific RM

Remote monitoring of patients can be fine-tuned based on the metrics applied. For instance, the metrics such as makespan of healthcare applications, availability of service requests, scheduling features of applications, and so forth, could be improved by skillfully choosing the metrics in algorithms or implementations.

The most common metrics that typically improve the performance of healthcare applications are listed as follows:

1. *Round-Time*: This metric determines the time taken for delivering the service requests. This round-time metric involves time to select the doctor, choose the right services, and so forth. In [36], authors have simulated a case study to manifest the importance of responsiveness metrics while designing healthcare applications in cloud environments. The authors revealed the scalable feature of clouds which improved

the round-time of sensor applications. Also, authors of [45] discussed how low-latency could be improved in healthcare applications when edge devices were included in the monitoring system.

2. *Reliability*: Reliability metric checks the availability of services or medical things for processing applications. It includes features for evaluating the availability of services. A few researchers have studied the reliability feature of healthcare applications. Notably, authors of [44] designed oneM2M protocol that included a fault-tolerant algorithm for increasing the reliability in the gateway layer of sensor networks.

3. *Energy consumption*: As similar to the performance of services, this metric, if fine-tuned, attempts to reduce the energy consumption of applications. In fact, the energy consumption of applications has to be diligently handled, especially for executing applications in power scarce locations or power-constraint devices. This metric is one of the most crucial ones for assessing healthcare applications. Authors of [18] proposed energy-efficient optimization-based clustering approach for connecting sensor nodes of the patient monitoring system. Their results, when applied with the Particle Swarm Optimization (PSO) technique have manifested to diagnose diseases with a minimum energy value.

4. *End-to-end encryption*: Providing end-to-end encryption for healthcare applications could improve their security and privacy features of them. Blockchain features are incorporated in some healthcare applications to improve security aspects of them [30, 40]. A few authorization approaches are studied in [74]. However, a tradeoff exists between the performance and security aspects of applications.

5. *Inter-operability count*, The inter-operability feature specifies the number of sinking devices of a medical thing. A higher number of interoperability metrics reveals the capabilities of the device to connect to more healthcare applications.

4.7 Focused delivery-specific RM

The mechanisms applied for healthcare applications could be focused on solving specified objectives – i) Emergency purposes [39], ii) Elderly Care, iii) Disability Care [60], and iv) Behavioral/Mental Care [37, 38, 81]. The mechanisms implemented to solve these specified objectives are unique. For instance, focused delivery of healthcare services to elderly people requires specialized care for bathing, fall detection, medicine reminders, personal hygiene, and emotional care. Most preferably, these elderly people staying in old-age homes are depressed due to emotional avoidance from family or relatives. IoT-enabled systems could drive them to play comforting music or their preferred games, protect them from more-likely falls by actuating airbags, warm up them during winter seasons using heat-actuated jackets, and so forth. Researchers have started to work on these focused deliveries of medical assistance in recent years. For instance, authors of [29] proposed a three-tier framework involving medical centres to care the elderly people; authors of [13] investigated the procedure to handle older adults considering scheduled medical consultations.

Authors of [37, 38] introduced a novel persuasive technology approach for improving the behaviors or mental states of human beings. The persuasive technology, associated with IoT-enabled wearable gadgets, could improve the cost involved in hiring mental healthcare professionals. For instance, AI-assisted chat bots could control the suicide attempts of young minds if they are semantically designed to tailor the needs of defaulters.

5 Remote monitoring healthcare applications – discussions

Remote monitoring healthcare applications are drastically shaping the future, especially during the post COVID-19 pandemic era. These healthcare applications modernize the approach of living standards as appropriate management procedures, sensor nodes, communication methodologies, detection/diagnostic algorithms, tools, and datasets have been evolved in the recent past.

5.1 Applications – categories

Establishing an intuitive understanding of the characteristics or requirements of the existing remote monitoring healthcare applications is an initial primordial step to deliver newer techniques and innovations. In fact, a few orthogonal research dimensions were identified based on i) applying managerial techniques, ii) addressing specific health concerns, and iii) utilizing assistive technologies.

5.1.1 Management-oriented

The involvement of IoT sensors and associated technologies has fixated in a few healthcare applications for remotely managing hospital premises or knowledge acquisition. Mechanisms need to be channelized to manage sectorial growth in healthcare.

Hospital management Clinical trials, especially during the post-COVID era, have seen a shift of focus in managing hospital premises with the advent of IoT technologies. The most diverse remote monitoring which manages hospital premises involve:

1. organizing medical equipment and devices based on the knowledge shared by the IoT gadgets;

2. optimizing the location layout of medical equipment such as wheelchairs, oxygen defibrillators, and

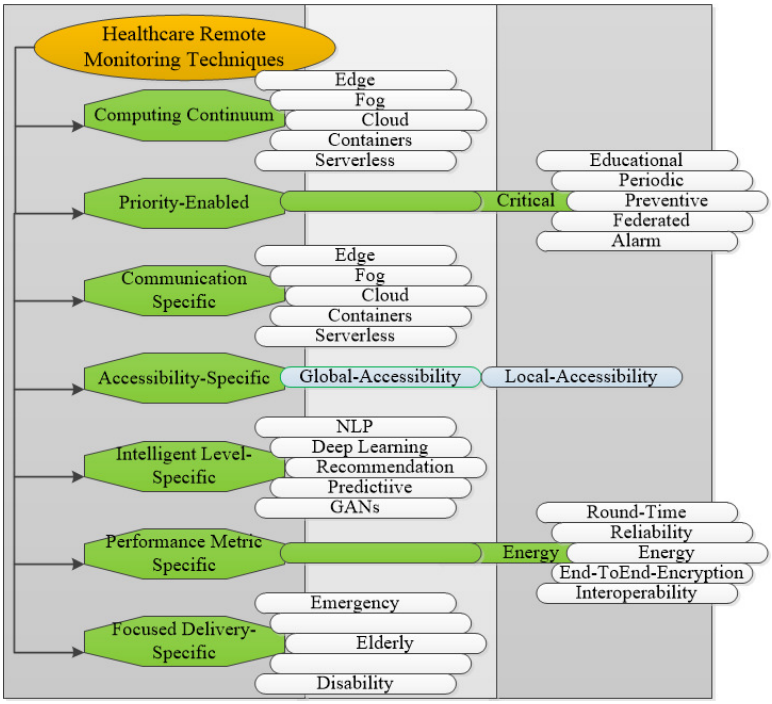


Figure 3: Taxonomy of Remote Healthcare Monitoring.

so forth, for distributing them across the hospital premises;

3. designating manpower such as doctors, nurses, cleaning staff, and security guards considering their willingness to serve wards and overseeing the need of departments;

4. automatizing the collaborative involvement of machines and sensors to coordinate the out/in patients during the stay in hospital premises;

5. managing assets in pharmaceutical department of hospitals, and, so forth.

Drug management In addition to managing hospital premises, IoT sensors and remote monitoring solutions are often utilized to trigger the well-being of patients by managing the drug delivery processes. There exist a few drug management solutions as listed below:

1. Drug Reminders – Solutions and medical kits were developed to constantly remind patients of the intake of drugs. In doing so, chronic patients are benefited at large. There have been several cases of deaths or medical emergencies as patients such as diabetes consume overdose or low dose insulin owing to repressing the consumption of drugs.

2. Drug Identification – In some situations, it is required to identify the complications due to drugs. Particularly, the previous history of patients with respect to the drugs has to be evaluated based on IoT systems or medical records. Investigating the real-time effect of drugs using IoT-enabled applications has bolstered the reduction of side effects to patients.

3. Drug Governance – Pharmaceutical companies and drug delivery units of countries require a machine-assisted system that senses and administers optimal supply chain management. Appropriate communication technologies and processing of data have to be handled in these machine-enabled applications for providing better logistics.

Data management One of the key points to converge IoT and healthcare applications is data management – i.e., a clear plan to perform data acquisition from sensor nodes considering geographical locations or large-scale data. Healthcare applications need to consider the continuum of computing devices located in edge, fog, or cloud environments [49]. Handling data in edge nodes based on the latency and real-time requirement of applications has to be automatized in the data-related IoT healthcare applications. There exist a few data management tools for healthcare applications such as Enterprise Data Warehouses (EDWs) for processing data collected from diverse sensor nodes in real-time. A few companies such as SnapLogic have delivered a platform-as-a-service model to elastically scale data integration services.

5.1.2 Health-specific

IoT-enabled healthcare applications that target addressing health issues could be classified depending on i) chronic

diseases, ii) Disease-specific or iii) Preventive care diseases.

Chronic Diseases Remotely monitoring chronic diseases with the intention of reducing the death rate and improving cost efficiency is one of the most awaited purposes of applications in this pandemic era. Chronic patients such as diabetic Mellitus, cardiovascular diseases, respiratory diseases, malignant neoplasms, cancer patients, and so forth, have to periodically visit hospitals if remote monitoring was not extended to the patients. It is to be noticed that over 85 percent of elderly people are prone to at least one of the above-mentioned chronic diseases. Obviously, the costs involved in the patient monitoring processes have been reduced due to the IoT-enabled healthcare system in such situations.

Common Diseases The majority of the diseases relating to general infections or deficiencies could be remotely assisted using IoT-enabled healthcare systems. The solutions are most welcomed by individuals, including transnational healthcare aspirants owing to the following reasons:

1. Avoidance of hospital appointments,
2. Mobile-assisted interactions,
3. Frequent virtual meetings with doctors or hospital administrations, and,
4. Tailor-made solutions, including robotic solutions.

In fact, home assistance in a remote fashion plays a vital role in specific health-conscious diseases. In the past, a home nurse would be appointed to monitor the vital signs and administer medicines. With the advent of IoT-enabled mechanisms, home-based medical assistance is enthused for these specific-category patients. For instance, patients belonging to traumatic brain injury (TBI) or spinal cord injury prefer home assistance as traveling to hospitals for visiting the doctors is sometimes dangerous, especially for patients living in crowded societies. Tailor-made robots for home-assisted medical delivery are often practiced to provide therapy such as speech therapy or physiotherapy.

Preventive Care Preventive healthcare or prophylaxis is introduced decades ago to avoid potential risk factors of the living. IoT-enabled systems are designed and innovated to consider the sector of people requiring preventive care through appropriate remote healthcare monitoring mechanisms.

It is also important to advise concerned patients to avoid their routine habits such as improper dietary, chewing tobacco, smoking, consuming alcohol, and so forth, by understanding the level of glucose, blood pressure, and body mass. In succinct, IoT-enabled devices and frameworks could automatically guide the patients and prevent them from many of the diseases with limited visits to the concerned hospitals/doctors.

5.2 Tools, libraries, and datasets – remote monitoring mechanisms

Connected IoT devices and hospital premises have to include sophisticated tools, including AI tools, to collect the right sensor data and prevent or diagnose diseases in a remote fashion. The most commonly utilized tools, libraries, and datasets for predicting diseases are discussed based on the remote monitoring techniques discussed earlier.

5.2.1 Tools for computing continuum

The computing continuum consisting of edge, fog, cloud or serverless containers has most commonly utilized in IoT-enabled healthcare applications. It is often crucial to understanding the most available tools that enable the right computing continuum for executing applications. For instance, offloading sensor data from medical things to edge or cloud or fog-connected cloud should be automatized in the healthcare application framework.

The important functions for establishing similar tools to enable the computing continuum of IoT healthcare applications are listed as follows:

1. *Dynamic Configurations* – Setting configurations that switch between edge or fog or cloud for executing healthcare tasks is required for incorporating computing continuum in applications. For instance, the configurations could be set up using YAML files and processed at runtime using programs written in nodejs or golang.

2. *Flow-specific Representations* – The tool that provides computing continuum in an automatic fashion needs specific tools to represent tasks and subtasks in a flow model. Obviously, workflow-specific tools such as kubeflow or workflow-description language-based tools will be utilized for specifying the flow of tasks.

3. *Reliability Features* – One of the crucial ingredient for developing tools that promote a continuum of computing in healthcare tasks is to provide reliability features. Reliability measures ensure the consistency delivery of computing resources to complete the assigned tasks.

4. *Statefulness* – Stateful services, including the health status database, are important for completing workflow tasks or for recreating container instances during failures. In IoT-enabled healthcare applications, there is a high possibility that the applications fail due to connectivity failures or the non-availability of appropriate resources. Most importantly, a healthcare application that requires a cold-start computing instance could take a long execution time than executing it in a warm-start instance.

5.2.2 Tools for prioritizing tasks

Healthcare applications could easily lead to delayed responses owing to the increasing number of remote monitoring requirements and decreasing number of available resources, including hospital accessibility or doctor availability. Patient prioritization is, therefore, a crucial component of healthcare tools. There exist tools to prioritize tasks for delivering a fair prioritization of tasks. For instance, Clinical Priority Assessment Criteria (CPAC) is followed in New Zealand, Australia, and a few other countries to prioritize the need of outpatients. This approach is transformed to IoT-enabled healthcare applications for servicing the requests of online patients based on the availability of resources, including doctors.

5.2.3 Tools for communications

As discussed earlier, several sensor nodes or wearables are involved in submitting the sensor data over a secured channel to edge, fog, or cloud compute instances. These sensor nodes communicate to the higher-level device or peer nodes using various communication protocols such as WIFI, Bluetooth4.0, LoRA, Infrared, LiDR, and so forth. 4G/5G/NarrowBand technologies are involved to connect the communication medium of devices.

Tools are often required to automatically suggest connecting to devices within hospital premises and outside hospital premises. For instance, the data rate outside hospitals could be very less when compared to the hospital premises. Obviously, the teleconference approach, especially in video mode, to remotely monitor the health status of individuals has to be diligently handled by applications. Similarly, the devices are prone to electromagnetic interference. Hence, tools need to assist the application to utilize appropriate communication links.

5.2.4 Accessibility features

Accessibility in healthcare applications has two dimensions:

1. *Interface* – The framework should involve interfaces to provide feasible and effective communications to patients/individuals. For instance, a patient might not have the strength to type in available gadgets to communicate to doctors; the patient might find it difficulty to infer the suggestions delivered by doctors. In such a scenario, multiple assisting technologies should be combined to enable efficient communication and accessibility of end devices or doctors/hospitals.

2. *Device Accessibility* – In addition, accessibility of different connected medical things that follow different standards or protocols to collect data or deliver data is a challenge. Tools should enable the best possible approach to efficiently access devices or actuators belonging to the application.

5.2.5 AI-enabled tools

The inclusion of AI techniques in healthcare increases the detection accuracy or assisting nature of applications. However, appropriate tools and techniques have to be integrated into the system for achieving better results and experiences. The most notable approaches of utilizing AI techniques for healthcare applications could be listed as follows:

1. *NLP-assited interfaces*: Natural Language Processing (NLP) methods are required in healthcare to deal with the large volume of healthcare records or datasets. NLP algorithms could parse the text messages from a large volume of data and extract the significance of the context for representing them through interfaces. The programs or algorithms associated with these methods have to become robust and error-free – i.e., careful handling of the dataset is required for the NLP logic to represent the intended content of doctors or patients.

2. *Deep-Learning or Models for inferences*: Deep learning or similar neural network-based learning models have manifested in the healthcare sector for their fast predictions or inferences without compromising the expected accuracy.

   The applications of deep learning models in healthcare are immense. For instance, deep learning models are utilized to detect anomalies in X-ray images, CT scans, MRI scans, and so forth; to analyze documents or health records and predict diseases; and, so forth. For instance, authors of [15] have proposed deep learning-based hierarchical learning approach to improve the response time of sensor-enabled classification problems; also, authors of [53] applied neural networks to classify diseases.

3. *Recommendation systems*: AI is most commonly utilized as recommendation system in healthcare applications. Several recommendations for providing appropriate diets or awareness about the spread of diseases are carried out using recommendation systems. In addition, drug recommendation is also carried out from the perspective of doctors to patients. These recommendation systems have to collect sensed data in a secure fashion [19] from various sensors or medical records and reason recommendations depending on the presented contexts. Recently, in [24], authors have proposed a word similarity measure method using learning algorithms which improves the recommendations of selecting online IoT-enabled medical services.

4. *Predictions and Classifications*: AI embodies several statistical and optimization techniques to predict health conditions or classify diseases. There exist several prediction and classification problems in the past: for instance, the application of decision trees, support

Table 3: Remote Monitoring Techniques in Healthcare Applications.

No	**Focused**	**Perf. Metrics**	**Computing**	**Priority-level**	**Comm. Protocols**	**AI/ML**
[77]	Tele ECG (Heart Rate)	*	Cloud	Periodic	WIFI	*
[71]	Elderly	*	*	Periodic	WIFI	*
[16]	*	Exec. Time CPU Load	Fog	*	WIFI	*
[25]	Disability (Speech Disorder)	*	Mobile/ Edge	Periodic	WIFI	NLP
[20]	Heart disease	*	Android/ Edge	Continuous	WIFI, Bluetooth	*
[31]	Heart disease	Throughput	*	*	NBIoT	*
[50]	Cancer care	*	Cloud, Hadoop	*	WBAN, WIFI, Zigbee	ML/DL
[32]	Cancer care	Sensitivity	Cloud	*	WIFI	ANN/PSO
[1]	Diabetes care	Accuracy	Cloud/ Fog/Edge	Periodic	WBAN	Decision Support
[6]	*	Energy	Simulation	Critical	RFID, WBAN IEEE802.15.4	*
[4]	Cancer care	Security	*	Critical	WIFI	Kernels/ ML
[15]	*	Response Time	Edge, VMs	*	Hierarchical	DL
[34]	*	*	Container Edge	Periodic	WIFI, Zigbee, Bluetooth	*
[80]	Chronic Diseases	Energy harvesting	Edge, Cloud	Periodic	Bluetooth4.0	*
[3]	*	Round Time Energy	WBN cluster Simulation	*	RFID, WBN	*
[7]	*	Reliability, Response Time	Fog, Edge	*	5G, Fibre/xDSL	*
[17]	Non Focused	Latency, Computation Cost	Cloud	Periodic	WIFI, 5G	RS
[39]	Emergency	*	Hadoop, Cloud	Critical	6LoWPAN, RFID	*

vector machines, K-nearest neighbor algorithms, linear regressions, and so forth, were quite common in the healthcare domain. In addition, several unsupervised learning algorithms such as reward-based agent systems and reinforcement learning methods were incorporated in the applications.

In a few medical diagnosis approaches, researchers have proposed federated learning mechanisms to protect the privacy of patients by protecting the data within local compute resources. By incorporating federated learning models, fine-tuned local learning models were updated in global models for fast predictions. In fact, these learning mechanisms are beneficial for mobile-assisted learning systems [66] where the local models are executed in battery-powered mobile devices.

5. *Generative Adversarial Networks (GANs)*: In a few cases, the electronic medical records could not be exposed to the public for exploration which keeps the supervised learning models such as Convolutional Neural Networks (CNNs) a failure. GANs, an unsupervised AI learning model, support the learning process in such cases. GANs, in general, involve both the generator model and discriminator model for the learning dataset to increase the positivity of the predictions. For instance, the electronic medical records for Zika, a mosquito-borne virus infection, in Kerala are very rare. In such situations, GANs could produce some additional datasets without compromising the accuracy of predictions.

5.3 Comparison of remote monitoring applications

Table 3 illustrates the remote monitoring healthcare applications that were developed by researchers and practitioners in the past. It highlights the target groups such as the elderly, common diseases addressed, and so forth; the performance metrics analyzed in the applications such as execution time, CPU load, energy, round-time, and so forth; the computational units involved in the applications such as containers, VMs, Edge, Fog, Cloud, serverless, or hierarchical compute instances; the priority-level initiated in addressing the tasks; the communication protocols which connected sensors with gateways or healthcare applications/services; and the intelligence level incorporated in the applications.

The notable observations from the existing remote monitoring healthcare applications are listed as follows:

1. Majority of the applications utilized WIFI-based communication protocol to connect sensor nodes or hospital devices;

2. Among the available AI-assisted healthcare applications, deep-learning algorithms, and their variants were incorporated in the solutions for better accuracy and resolutions;

3. Almost all applications that were not simulated included a compendium of computing instances such as edge, fog, cloud, or VMs in a hierarchical fashion for pursuing the detection of diseases or providing remote medical assistance.

6 Conclusions

The prevalence of IoT-enabled healthcare applications has urged the inclusion of novelties to support individuals and doctors during the pandemic era. Investigating the remote monitoring techniques in the healthcare sector is a vital role in delivering superlative care by doctors to patients. In this paper, the most predominantly applied remote monitoring techniques were studied. Based on the findings, a generic remote monitoring healthcare architecture was elaborated and the taxonomy of remote monitoring techniques was illustrated. Additionally, techniques practiced in the existing remote monitoring healthcare applications and associated research directions were enlightened for the researchers.

Compliance with ethical standards

– Funding: This study was funded by IIIT-Kottayam Faculty Research Fund.

– Conflict of interest/Competing interests: The author declares that he has no conflict of interest.

– Ethics approval: This article does not contain any studies with human participants or animals performed by any of the authors.

– Consent to participate: Yes

– Consent for publication: Yes

– Authors' contributions: The corresponding author did the entire survey work.

– Informed consent: Informed consent was obtained from all individual participants included in the study.

References

[1] Abdel-Basset M., Manogaran G., A. Gamal and V. Chang, A Novel Intelligent Medical Decision Support Model Based on Soft Computing and IoT, in IEEE Internet of Things Journal, 7(5), pp. 4160–4170, 2020. https://doi.org/10.1109/jiot.2019.2931647

[2] Abdullah A., Ontology Middleware for Integration of IoT Healthcare Information Systems in EHR Systems., Computers. 7(51), pp. 1–15, 2018. https://doi.org/10.3390/computers7040051

[3] Abuelkhail, A., Baroudi, U., Raad, M. et al. Internet of things for healthcare monitoring applications based on RFID clustering scheme. Wireless Netw, Vol. 27, pp. 747–763, 2021. https://doi.org/10.1007/s11276-020-02482-1

[4] Ali Burak Unal and Mete Akgun and Nico Pfeifer, ESCAPED: Efficient Secure and Private Dot Product Framework for Kernel-based Machine Learning Algorithms with Applications in Healthcare, arXiV, 2012.02688, pp. 9982-9995, 2020.

[5] Alromaihi S., W. Elmedany and C. Balakrishna, Cyber Security Challenges of Deploying IoT in Smart Cities for Healthcare Applications, 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 140-145, 2018. https://doi.org/10.1109/w-ficloud.2018.00028

[6] Aktas F., Ceken C. and Erdemli Y.E., IoTBased Healthcare Framework for Biomedical Applications, J. Med. Biol. Eng. Vol. 38, pp. 966–979, 2018. https://doi.org/10.1007/s40846-017-0349-7

[7] Akrivopoulos O., I. Chatzigiannakis, C. Tselios and A. Antoniou, On the Deployment of Healthcare Applications over Fog Computing infrastructure, IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), pp. 288-293, 2017. https://doi.org/10.1109/compsac.2017.178

[8] Aladwani T., Scheduling IoT Healthcare Tasks in Fog Computing Based on their Importance, in Procedia Computer Science, Vol.163, pp. 560–569, 2019. https://doi.org/10.1016/j.procs.2019.12.138

[9] Alam A.A., H. Malik, M. I. Khan, T. Pardy, Kuusik and Y. Le Moullec, A Survey on the Roles of Communication Technologies in IoT-Based Personalized Healthcare Applications, in IEEE Access, Vol. 6, pp. 36611-36631, 2018. https://doi.org/10.1109/access.2018.2853148

[10] A.S. Albahri and Jwan K. Alwan and Zahraa K. Taha and Sura F. Ismail and Rula A. Hamid and A.A. Zaidan and O.S. Albahri and B.B. Zaidan and A.H. Alamoodi and M.A. Alsalem, IoTbased telemedicine for disease prevention and health promotion: State-ofthe-Art, in JNCA, 173(102873), pp. 1–52, 2021. https://doi.org/10.1016/j.jnca.2020.102873

[11] Amin Ul Haq, Victor Ejianya, Jalaluddin Khan, Asif Khan, Mudassir Khalil, Amjad Ali, Ghufran A. khan, Mohammad Shahid, Bilal Ahamad, Amit Yadav et al., A New Approach for Enhancing the Services of the 5G Mobile Network and IOT-Related Communication Devices Using Wavelet-OFDM and Its Applications in Healthcare, Vol. 2020, pp. 1–13, 2020. https://doi.org/10.1155/2020/3204695

[12] Anand S. and S. K. Routray, Issues and challenges in healthcare narrowband IoT, 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 486-489, 2017. https://doi.org/10.1109/icicct.2017.7975247

[13] Ara´ujo I´talo Linhares De, Costa Junior Evilasio, Duarte Paulo, Santos Ismayle De Sousa, Oliveira Pedro Almir Martins De, Mendes C´ıcero Marcelo Oliveira, Andrade Rossana Maria De Castro, Towards a Taxonomy for the Development of Older Adults Healthcare Applications, in Proc. of the 53rd Hawaii International Conference on System Sciences, 2020. https://doi.org/10.24251/hicss.2020.430

[14] Aujla G.S. and A. Jindal, A Decoupled Blockchain Approach for Edge-Envisioned IoT-Based Healthcare Monitoring, in IEEE Journal on Selected Areas in Communications, 39(2), pp. 491-499, Feb. 2021. https://doi.org/10.1109/jsac.2020.3020655

[15] Azimi I., J. Takalo-Mattila, A. Anzanpour, A. M. Rahmani, J. Soininen and P. Liljeberg, Empowering Healthcare IoT Systems with Hierarchical Edge-Based Deep Learning, 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 63-68, 2018. https://doi.org/10.1145/3278576.3278597

[16] Bahar Farahani and Farshad Firouzi and Victor Chang and Mustafa Badaroglu and Nicholas Constant and Kunal Mankodiya, Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare, in Future Generation Computer Systems, Vol. 78, pp. 659-676, 2018. https://doi.org/10.1016/j.future.2017.04.036

[17] Bhattacharya P., S. Tanwar, U. Bodkhe, S. Tyagi and N. Kumar, BinDaaS: BlockchainBased Deep-Learning as-a-Service in Healthcare 4.0 Applications, in IEEE Transactions on Network Science and Engineering, 8(2), pp. 1242-1255, 2021. https://doi.org/10.1109/tnse.2019.2961932

[18] Bharathi R., T. Abirami, S. Dhanasekaran, Deepak Gupta, Ashish Khanna, Mohamed Elhoseny, K. Shankar, Energy efficient clustering with disease diagnosis model for IoT based sustainable healthcare systems, Sustainable Computing: Informatics and Systems, 28(100453), 2020. https://doi.org/10.1016/j.suscom.2020.100453

[19] Cansu Eken, and Hanim E, Security Threats and Recommendation in IoT Healthcare, doi. 10.3384/ecp17142369, pp. 369-374, 2016.

[20] Chao Li and Xiangpei Hu and Lili Zhang, The IoT-based heart disease monitoring system for

pervasive healthcare service, in Procedia Computer Science, Vol. 112, pp. 2328-2334, 2017. https://doi.org/10.1016/j.procs.2017.08.265

[21] Danish Jamil, Diagnosis of Gastric Cancer Using Machine Learning Techniques in Healthcare Sector – A Survey, in Informatica – An Int. J. of Computing and Informatics (Slovenia), 45(7), pp. 147–166, 2021. https://doi.org/10.31449/inf.v45i7.3633

[22] David Charlas Amos, Remote patient monitoring, patent, https://lens.org/010-983-959-342-733, 2017.

[23] De Morais Barroca Filho I., De Aquino Junior G.S., IoT-Based Healthcare Applications: A Review. In: Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2017. ICCSA 2017. Lecture Notes in Computer Science, Vol. 10409, 2017. https://doi.org/10.1007/978-3-319-62407-5_4

[24] Dehai Zhang, Xiaoqiang Xia, Yun Yang, PoYang, Cheng Xie, Menglong Cui, Qing Liu, A novel word similarity measure method for IoT-enabled Healthcare applications, Future Generation Computer Systems, Vol. 114, pp. 209-218, 2021. https://doi.org/10.1016/j.future.2020.07.053

[25] Dubey H., J. C. Goldberg, K. Makodiya, and L. Mahler, A multi-smartwatch system for assessing speech characteristics of people with dysarthria in group settings, in Proceedings e-Health Networking, Applications https://doi.org/10.1109/healthcom.2015.7454559

[26] Elhayatmy G., Dey N., Ashour A.S., Internet of Things Based Wireless Body Area Network in Healthcare. In: Dey N., Hassanien A., Bhatt C., Ashour A., Satapathy S. (eds) Internet of Things and Big Data Analytics Toward Next-Generation Intelligence. Studies in Big Data, vol 30. Springer, Cham. 2018. https://doi.org/10.1007/978-3-319-60435-0_1

[27] Faris A. Almalki and Ben Othman Soufiene, EPPDA: An Efficient and PrivacyPreserving Data Aggregation Scheme with Authentication and Authorization for IoT-Based Healthcare Applications, Wireless Communications and Mobile Computing, 2021(5594159), pp. 1–18, 2021. https://doi.org/10.1155/2021/5594159

[28] Fouad H., Azza S. Hassanein, Ahmed M. Soliman, Haytham Al-Feel, Analyzing patient health information based on IoT sensor with AI for improving patient assistance in the future direction, in Measurement, 159(107757), pp. 1–11, 2020. https://doi.org/10.1016/j.measurement.2020.107757

[29] Guizani K. and S. Guizani, IoT Healthcare Monitoring Systems Overview for Elderly Population, 2020 International Wireless Communications and Mobile Computing (IWCMC), pp. 2005-2009, 2020. https://doi.org/10.1109/iwcmc48107.2020.9148446

[30] Haoxiang Wang, IoT based Clinical Sensor Data Management and Transfer using Blockchain Technology, in Journal of ISMAC, 2(3), pp. 154-159, 2020. https://doi.org/10.36548/jismac.2020.3.003

[31] Hassan Malik and Muhammad Mahtab Alam and Yannick Le Moullec and Alar Kuusik, NarrowBand-IoT Performance Analysis for Healthcare Applications, in Procedia Computer Science, Vol. 130, pp. 1077-1083, 2018. https://doi.org/10.1016/j.procs.2018.04.156

[32] Indra P., Manikandan M., Multilevel Tetrolet transform based breast cancer classifier and diagnosis system for healthcare applications. J Ambient Intell Human Comput Vol. 12, pp. 3969–3978, 2021. https://doi.org/10.1007/s12652-020-01755-z

[33] Ifrim Claudia, and Pintilie, Andreea-Mihaela and, Apostol, Elena Simona and, Dobre, Ciprian and Pop, Florin, The Art of Advanced Healthcare Applications in Big Data and IoT Systems, 2017. https://doi.org/10.1007/978-3-319-45145-9_6

[34] Jaiswal K., S. Sobhanayak, A. K. Turuk, S. L. Bibhudatta, B. K. Mohanta and D. Jena, An IoT-Cloud Based Smart Healthcare Monitoring System Using Container Based Virtual Environment in Edge Device, International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), pp. 1–7, 2018. https://doi.org/10.1109/icetietr.2018.8529141

[35] Jun Qi, Po Yang, Geyong Min, Oliver Amft, Feng Dong, Lida Xu, Advanced internet of things for personalised healthcare systems: A survey, Pervasive and Mobile Computing, Vol. 41, pp. 132-149, 2017. https://doi.org/10.1016/j.pmcj.2017.06.018

[36] K. Kadarla, S. C. Sharma, T. Bhardwaj and A. Chaudhary, A Simulation Study of Response Times in Cloud Environment for IoT-Based Healthcare Workloads, 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), pp. 678-683, 2017. https://doi.org/10.1109/mass.2017.65

[37] Kolenik T, Gams M. Increasing Mental Health Care Access with Persuasive Technology for Social Good, in IJCAI 2021 Workshop on AI for Social Good, 2021.

[38] Kolenik T. and M. Gams, Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality?, in IEEE Technology and Society Magazine, 40(1), pp. 80-86, March 2021, https://doi.org/10.1109/mts.2021.3056288

[39] Lakkis S.I., and M. Elshakankiri, IoT based emergency and operational services in medical care systems, 2017 Internet of Things Business Models, Users, and Networks, pp. 1-5, 2017. https://doi.org/10.1109/ctte.2017.8260983

[40] Leila Ismail, Huned Materwala, and Alain Hennebelle, A Scoping Review of Integrated Blockchain-Cloud (BcC) Architecture for Healthcare: Applications, Challenges and Solutions, in Sensors, 21(3753), pp. 1–23, 2021. https://doi.org/10.3390/s21113753

[41] Manogaran G., Thota C., Lopez D., Sundarasekar R., Big Data Security Intelligence for Healthcare Industry 4.0. In: Thames L., Schaefer D. (eds) Cybersecurity for Industry 4.0. Springer Series in Advanced Manufacturing. Springer, Cham. 2017. https://doi.org/10.1007/978-3-319-50660-9_5

[42] Matthew McGrath and Evan Alexander Dewhirst, Method Of Medical Imaging Using Multiple Arrays, patent, https://lens.org/041-801-178-563-669, 2021.

[43] Martin Gjoreski, Bhargavi Mahesh, Tine Kolenik, Jens UWE-Garbas, Dominik Seuss, Hristijan Gjoreski, Mitja Lustrek, Matjaz Gams, and Veljko Pejovic, Cognitive Load Monitoring With Wearables–Lessons Learned From a Machine Learning Challenge, in IEEE Access, Vol. 9, pp. 103325–103336, 2021, https://doi.org/10.1109/access.2021.3093216

[44] Min Woo Woo and JongWhi Lee and Kee-Hyun Park, A reliable IoT system for Personal Healthcare Devices, in Future Generation Computer Systems, Vol. 78, pp. 626-640, 2018. https://doi.org/10.1016/j.future.2017.04.004

[45] Muneeb Ejaz, Tanesh Kumar, Ivana Kovacevic, Mika Ylianttila and Erkki Harjula, Health-BlockEdge: BlockchainEdge Framework for Reliable LowLatency Digital Healthcare Applications, in Sensors, 21(2502), 2021. https://doi.org/10.3390/s21072502

[46] Naghshvarianjahromi N., S. Majumder, S. Kumar,and M. J. Deen, Natural BrainInspired Intelligence for Screening in Healthcare Applications, in IEEE Access, Vol. 9, pp. 67957-67973, 2021. https://doi.org/10.1109/access.2021.3077529

[47] Nidhya R., Karthik S., Smilarubavathy G., An End-to-End Secure and EnergyAware Routing Mechanism for IoT-Based Modern Health Care System. In: Wang J., Reddy G., Prasad V., Reddy V. (eds) Soft Computing and Signal Processing. Advances in Intelligent Systems and Computing, vol 900, 2019. https://doi.org/10.1007/978-981-13-3600-3_35

[48] Noemi S., Pieroni A., Luca D.N., Francesca F., E-health-IoT universe: A review. International Journal on Advanced Science, Engineering and Information Technology, pp. 2328-2336, 2017.

[49] Onasanya, A., Lakkis, S. and Elshakankiri, M., Implementing IoT/WSN based smart Saskatchewan Healthcare Systems, in Wireless Networks, Vol. 25, pp. 3999–4020, 2019. https://doi.org/10.1007/s11276-018-01931-2

[50] Onasanya, A. and Maher E., Smart integrated IoT healthcare system for cancer care, in Wireless Networks, doi: https://doi.org/10.1007/s11276-018-01932-1, 2019.

[51] Pace P., G. Aloi, R. Gravina, G. Caliciuri, G. Fortino and A. Liotta, An Edge-Based Architecture to Support Efficient Applications for Healthcare Industry 4.0, in IEEE Transactions on Industrial Informatics, 15(1), pp. 481-489, Jan. 2019. https://doi.org/10.1109/tii.2018.2843169

[52] Prabal Verma, Sandeep K. Sood, Cloud-centric IoT based disease diagnosis healthcare framework, Journal of Parallel and Distributed Computing, Vol. 116, pp. 27-38, 2018. https://doi.org/10.1016/j.jpdc.2017.11.018

[53] Priyan Malarvizhi Kumar, S. Lokesh, R. Varatharajan, Gokulnath Chandra Babu, P. Parthasarathy, Cloud and IoT based disease prediction and diagnosis system for healthcare using Fuzzy neural classifier, Future Generation Computer Systems, Vol. 86, pp. 527-534, 2018. https://doi.org/10.1016/j.future.2018.04.036

[54] Ravi Pratap Singh, Mohd Javaid, Abid Haleem, Rajiv Suman, Internet of things (IoT) applications to fight against COVID19 pandemic, in Diabetes Metabolic Syndrome: Clinical Research Reviews, 14(4), pp. 521–524, 2020. https://doi.org/10.1016/j.dsx.2020.04.041

[55] Redowan Mahmud, Fernando Luiz Koch, and Rajkumar Buyya, Cloud-Fog Interoperability in IoT-enabled Healthcare Solutions, In Proceedings of the 19th International Conference on Distributed Computing and Networking (ICDCN '18). Association for Computing Machinery, pp. 1–10, 2018. https://doi.org/10.1145/3154273.3154347

[56] Rita Zgheib, Emmanuel C., R´emi B., Engineering IoT Healthcare Applications: Towards a Semantic Data Driven Sustainable Architecture, EAI International Conference on Ambient Assisted Living Technologies based on Internet of Things, Budapest, JUNE 14–15, 2016. https://doi.org/10.1007/978-3-319-49655-9_49

[57] Routray S.K. and S. Anand, Narrowband IoT for healthcare, 2017 International Conference on Information Communication and

Embedded Systems (ICICES), pp. 1-4, 2017. https://doi.org/10.1109/icices.2017.8070747

[58] Sallabi F., F. Naeem, M. Awad and K. Shuaib, Managing IoT-Based Smart Healthcare Systems Traffic with Software Defined Networks, International Symposium on Networks, Computers and Communications (ISNCC), pp. 1–6, 2018. https://doi.org/10.1109/isncc.2018.8530920

[59] Salahuddin M.A., A. Al-Fuqaha, M. Guizani, K. Shuaib and F. Sallabi, Softwarization of Internet of Things Infrastructure for Secure and Smart Healthcare, in Computer, 50(7), pp. 74-79, 2017. https://doi.org/10.1109/mc.2017.195

[60] Santiago Meli´a, Shahabadin Nasabeh, Sergio Luj´an-Mora, MoSIoT: Modeling and Simulating IoT Healthcare-Monitoring Systems for People with Disabilities, in International Journal of Environmental Research and Public Health and Cristina Cachero, 18(6357), pp. 1–25, 2021. https://doi.org/10.3390/ijerph18126357

[61] Sebbak F., Benhammadi, F. Majorityconsensus fusion approach for elderly IoT-based healthcare applications, Ann. Telecommun, Vol. 72, pp. 157–171, 2017. https://doi.org/10.1007/s12243-016-0550-7

[62] Selvaraj S., Sundaravaradhan S., Challenges and opportunities in IoT healthcare systems: a systematic review, SN Appl. Sci., 2(139), 2020. https://doi.org/10.1007/s42452-019-1925-y

[63] Shajulin Benedict, Badami, R., Bhagyalakshmi, M., APM Bots: An Automated Presentation Maker for Tourists/Corporates Using NLP-Assisted Web Scraping Technique, in proceedings of Advanced Network Technologies and Intelligent Computing. ANTIC 2021, Communications in Computer and Information Science, vol 1534, doi: https://doi.org/10.1007/978-3-030-96040-7_49

[64] Shajulin Benedict, Revenue oriented air quality prediction microservices for smart cities, 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 437-442, https://doi.org/10.1109/icacci.2017.8125879

[65] S. Sharma, K. Chen and A. Sheth, Toward Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems, Remote Monitoring Techniques, in IEEE Internet Computing. 22(2), pp. 42-51, 2018. https://doi.org/10.1109/mic.2018.112102519

[66] Shah Nazir, Yasir Ali, Naeem Ullah, and Ivn Garcıa-Magarino, Internet of Things for Healthcare Using Effects of Mobile Computing: A Systematic Literature Review, in Wireless Communications and Mobile Computing, 2019(5931315), pp.1–20. 2021. https://doi.org/10.1155/2019/5931315

[67] Sheth A., H. Y. Yip and S. Shekarpour, Extending Patient-Chatbot Experience with Internet-of-Things and Background Knowledge: Case Studies with Healthcare Applications, in IEEE Intelligent Systems, 34(4), pp.24-30, 2019. https://doi.org/10.1109/mis.2019.2905748

[68] Sivadi B., Thirumaran, M., Vijender Kumar S., IoT Sensor Data Integration in Healthcare using Semantics and Machine Learning Approaches, in A Handbook of Internet of Things in Biomedical and Cyber Physical System, pp.275300, 2020. https://doi.org/10.1007/978-3-030-23983-1_11

[69] Somayeh Nasiri, Farahnaz Sadoughi, Afsaneh Dehnad, Mohammad Hesam Tadayon, Hossein Ahmadi: Layered Architecture for Internet of Things-based Healthcare System: A Systematic Literature Review. in Informatica – An Int. J. of Computing and Informatics (Slovenia), 45(4), pp. 543–562, 2021. https://doi.org/10.31449/inf.v45i4.3601

[70] Subramanian, Balaji, Nathani, Karan, Rathnasamy, Santhakumar, IoT Technology, Applications and Challenges: A Contemporary Survey. Wireless Personal Communications, 2019. https://doi.org/10.1007/s11277-019-06407-w

[71] Tan B. et al., Wi-Fi based passive human motion sensing for in-home healthcare applications, 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), pp. 609-614, 2015. https://doi.org/10.1109/wf-iot.2015.7389123

[72] Tang F., L. Zeng, F. Wang and J. Zhou, Adversarial Precision Sensing with Healthcare Applications, 2020 IEEE International Conference on Data Mining (ICDM), pp. 521-530, 2020. https://doi.org/10.1109/icdm50108.2020.00061

[73] Tekeste T., H. Saleh, B. Mohammad and M. Ismail, Ultra-Low Power QRS Detection and ECG Compression Architecture for IoT Healthcare Devices, in IEEE Transactions on Circuits and Systems I: Regular Papers, 66(2), pp. 669-679, Feb. 2019. https://doi.org/10.1109/tcsi.2018.2867746

[74] Tsu-Yang Wu, Tao Wang, Yu-Qi Lee, Weimin Zheng, Saru Kumari, and Sachin Kumar, Improved Authenticated Key Agreement Scheme for Fog-Driven IoT Healthcare System, in Security and Communication Networks, 2021(6658041), pp. 1–16, 2021. https://doi.org/10.1155/2021/6658041

[75] Tun SYY, Madanian S, Mirza F. Internet of things (IoT) applications for elderly care: a reflective review, Aging Clin Exp Res, 33(4), pp.855-867, 2020. https://doi.org/10.1007/s40520-020-01545-9

[76] Ud Din, A. Almogren, M. Guizani and M. Zuair, A Decade of Internet of Things: Analysis in the Light of Healthcare Applications, in IEEE Access, vol. 7, pp. 89967-89979, 2019. https://doi.org/10.1109/access.2019.2927082

[77] Ungurean L. and A. Brezulianu, An internet of things framework for remote monitoring of the healthcare parameters, Adv. Electr. Comput. Eng., 17(2), pp. 11–16, 2017. https://doi.org/10.4316/aece.2017.02002

[78] Venceslau A., V. Vidal and R. Andrade, Context-driven Abnormal Semantic Event Recognition for Healthcare Applications, 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 434-435, 2021. https://doi.org/10.1109/percom workshops51409.2021.9431117

[79] Wei Li, Cheolwoo Jung and Jongtae Park, IoT Healthcare Communication System for IEEE 11073 PHD and IHE PCD-01 Integration Using CoAP, KSII Trans. on Internet and Information Systems, 12(4), pp. 1396–1414, 2018. https://doi.org/10.3837/tiis.2018.04.001

[80] Wu T., J. Redout´e and M. R. Yuce, A Wireless Implantable Sensor Design With Subcutaneous Energy Harvesting for LongTerm IoT Healthcare Applications, in IEEE Access, vol. 6, pp. 35801-35808, 2018. https://doi.org/10.1109/access.2018.2851940

[81] Xainglan Peng, Research on Emotion Recognition Based on Deep Learning for Mental Health, in Informatica – An Int. J. of Computing and Informatics (Slovenia), 45(1), pp. 127 – 132, 2021. https://doi.org/10.31449/inf.v45i1.3424

Unsupervised Deep Learning: Taxonomy and Algorithms

Aida Chefrour[1,2,*] and Labiba Souici-Meslati[2]
E-mail: aida.chefrour@univ-soukahras.dz, souici_labiba@yahoo.fr
[1] Computer Science Department, Mohamed Cherif Messaadia University, Souk Ahras, Algeria
[2] LISCO Laboratory, Computer Science Department, Badji Mokhtar University, B.P-12 Annaba, 23000, Algeria

Overview paper

Keywords: clustering, deep learning, autoencoder, taxonomy

Received: November 10, 2021

Clustering is a fundamental challenge in many data-driven application fields and machine learning techniques. The data distribution determines the quality of the outcomes, which has a significant impact on clustering performance. As a result, deep neural networks can be used to learn more accurate data representations for clustering. Many recent studies have focused on employing deep neural networks to develop a clustering-friendly representation, which has resulted in a significant improvement in clustering performance. We present a systematic survey of clustering with deep learning in this study. Then, a taxonomy of deep clustering is proposed, as well as some sample algorithms for our overview. Finally, we discuss some exciting future possibilities for clustering using deep learning and offer some remarks.

Povzetek: Ta članek opisuje metode globokega združevanja v skupine in predlaga taksonomijo globokega združevanja v skupine.

1 Introduction

Clustering is one of the most important aspects of unsupervised machine learning. Its main goal is to separate a data set into subsets or clusters so that data values in the same cluster have some common characteristics or attributes. It aims to divide the data into groups (clusters) of similar objects. The objects in the same cluster are more identical to each other than to those in other clusters. Clustering is widely used in Artificial Intelligence, pattern recognition, statistics, and other information processing fields. The input of a cluster analysis system is a set of samples and a measure of similarity (or dissimilarity) between two samples. The output is a set of clusters that form a partition, or a structure of partitions of the data set. Generally, finding clusters is not a simple task and the current clustering algorithms take a long time when they are applied to large databases [1].

In addition, the transformation of input data into a feature space where separation is easier concerning the problem's context, dimensionality reduction, and representation learning has been widely applied to clustering, because the similarity measurements utilized in these procedures are ineffective.

Existing data transformation methods generally include linear transformations such as Principal Component Analysis (PCA) and non-linear transformations such as kernel approaches and spectral methods [2].

To solve this problem, Deep Neural Networks (DNNs) are used to train non-linear mappings that allow the data to be transformed into clustering-friendly representations because they have a significant non-linear transformation feature. In this paper, we refer to clustering approaches involving deep learning as deep clustering for simplicity.

In our research, we focus on Deep Clustering, which represents a family of clustering algorithms that adopt deep neural networks to learn cluster-oriented features [3]. Deep clustering has recently become popular as a method for data classification and feature representation discovery, a solution for large-scale and high-dimensional learning problems [4,5]

We were particularly interested in the studies conducted in deep clustering for image recognition. We give an overview of deep clustering to review most methods and implementations in this field.

The main contributions treated in this paper are:

- use of a deep autoencoder for embedding the data into a lower-dimensional space;
- integrate the extracting intermediate features phase and the performing phase of the traditional clustering algorithm;
- employ the similarity of the representation features if they are assigned to the same cluster;
- add dimensionality reduction and temporal clustering into a single unsupervised learning framework;

* Corresponding author

- apply the impressive ability to deal with unsupervised learning for structure analysis of high-dimensional visual data;
- find a solution to the problem of subspace clustering by partitioning data drawn from a union of multiple subspaces.

The contribution of this study is (1) to provide an overview of various deep learning-based clustering algorithms. It includes an explanation of the most recent improvements in unsupervised clustering; (2) propose a taxonomy of methods that use deep learning for clustering.

We chose to synthesize studies published in the previous 3-4 years since they used deep learning to increase unsupervised clustering performance. On the MNIST dataset, several algorithms achieve more than 96% accuracy without using a single labeled datapoint. However, for more difficult datasets like CIFAR-10 and ImageNet, they are still a long way from achieving good accuracy.

We'll go over all of the most recent deep learning-based clustering approaches in this article. The aim of most of these strategies is to discover feature representation and solve the problem of large-scale, high-dimensional learning, as well as to respond to the contributions mentioned above.

The rest of the paper is organized as follows. In the next section, we survey in brief the literature on deep clustering overviews. We present the most recent works using unsupervised deep learning in section 3, with a synthesis of all of this work in section 4. In section 5, we describe the proposed taxonomy of clustering with deep learning algorithms and we introduce some representative methods. Section 6 includes a conclusion and proposals for further research.

2 Related work

Several custom taxonomies for clustering with deep learning have been proposed in the literature. In this section, we outline the best known and most recent ones:

[6] focus on a review of deep learning for multimodal data fusion, which provides readers with the fundamentals of the multimodal deep learning fusion methods. This study summarizes the representative architectures—DBN, SAE, CNN, and RNN—which are fundamental to understanding multimodal deep learning fusion models. This work summarizes the pioneering multimodal deep learning fusion models from the task, model framework, and data set perspectives, and groups them by the deep learning architecture used.

[2] divide deep clustering algorithms into four categories: AE-based (Autoencoder), CDNN-based (Clustering DNN), VAE-based (Variational encoder), and GAN-based deep clustering (Generative Adversarial Network). Each category has some representative methods as well.

- (a) AE-based has a (1) Deep Clustering Network (DCN), which combines an autoencoder with the k-means algorithm; (2) Deep Embedding Network (DEN) to extract effective representations for clustering, which utilizes a deep autoencoder; (3) Deep Subspace Clustering Networks (DSC-Nets) which introduces a novel autoencoder architecture; (4) Deep Multi-Manifold Clustering (DMC); (5) Deep Embedded Regularized Clustering (DEPICT); and (6) Deep Continuous Clustering (DCC);
- (b) CDNN-based deep clustering algorithms can be divided into three categories according to the way of network initialization, i.e., unsupervised pre-trained (Deep Nonparametric Clustering (DNC), Deep Embedded Clustering (DEC), Discriminatively Boosted Clustering (DBC)), supervised pre-trained (Clustering Convolutional Neural Network (CCNN)), randomly initialized (non-pre-trained) (Information Maximizing Self-Augmented Training (IMSAT), Joint Unsupervised Learning (JULE) and Deep Adaptive Image Clustering (DAC));
- (c) VAE-based deep clustering, which can be considered as a generative variant of AE. It presents two algorithms: (1) Variational Deep Embedding (VaDE) and (2) Gaussian Mixture VAE (GMVAE);
- (d) GAN-based deep clustering contains a (1) Deep Adversarial Clustering (DAC), (2) Categorial Generative Adversarial Network (CatGAN), and (3) Information Maximizing Generative Adversarial Network (InfoGAN).

[7] propose a taxonomy of clustering algorithms that employ deep learning. Their taxonomy helps the user to see what methods are available and to create new ones by combining the best characteristics of existing methods in a simple context. This taxonomy's main principle is representation learning with DNNs and using these representations as input to a specific clustering approach. Every method is divided into the following parts, each of which has a variety of options: (1) Architecture of the main neural network branch (Multilayer perceptron (MLP), Convolutional neural network (CNN) and Deep Belief Network (DBN)); (2) Set of deep features used for clustering (one layer, several layers); (3) Non- clustering loss (No non-clustering loss, Autoencoder reconstruction loss); (4) Clustering loss (No clustering loss, k-Means loss, Cluster assignment hardening, Balanced assignments loss, Locality-preserving loss, Group sparsity loss, Cluster classification loss, and Agglomerative clustering loss); (5) Method to combine the losses (Pre-training, fine-tuning, Joint training and Variable schedule); (6) Cluster updates (Jointly updated with the network model, and Alternatingly updated with the network model); (7) After network training (Clustering a similar dataset and Obtaining better results). The methods which use this taxonomy are Deep Embedded Clustering (DEC), Deep Clustering Network (DCN), Discriminatively Boosted Clustering (DBC), Joint Unsupervised Learning of Deep Representations and Image Clusters (JULE), and Clustering CNN (CCNN).

[8] propose a simplified taxonomy based on deep clustering algorithms' overall procedural structure or design. Deep Clustering techniques can be classified into three broad families according to this taxonomy: (a) Sequential multistep Deep Clustering approaches: these approaches have two basic steps. The first stage involves learning richer deep (also known as latent) representation of the input data, followed by clustering on this deep or latent representation in the second step; (b) Joint Deep Clustering approaches: Instead of two independent processes for representation learning and clustering, this family of approaches includes a step where the representation learning is intimately associated with the clustering. Tight coupling is usually achieved by optimizing a combined or joint loss function that promotes good reconstruction while accounting for some sort of data grouping, clustering, or codebook representation; (c) Closed-loop multistep Deep Clustering approaches: Similar to the first family (sequential multistep Deep Clustering), this family of algorithms has two key phases that alternate in an iterative loop rather than being conducted in a single feedforward linear approach.

3 Contributions of deep clustering

In recent years, we have noticed that there are many applications in the field of deep learning using unsupervised learning algorithms for image recognition. We now discuss some of the most common deep clustering approaches.

[9] find that existing deep clustering algorithms either do not take advantage of convolutional neural networks well enough or do not preserve the local structure of data-generating distribution in the learned feature space sufficiently. In this research, they suggest a deep convolutional embedded clustering method as a solution to this problem. They create a convolutional autoencoder structure to learn embedded features from start to finish. Then, using embedded features, a clustering-oriented loss is created to accomplish feature refinement and cluster assignment simultaneously. They keep the decoder, which can preserve the local structure of data in feature space, to avoid feature space being affected by clustering loss. In summary, they minimize both the reconstruction and clustering losses of convolutional autoencoders. Mini-batch stochastic gradient descent with back-propagation can effectively solve the resulting optimization issue. Experiments on benchmark datasets (MNIST-full, MNIST-test, and USPS) empirically verify the usefulness of local structure preservation and the power of

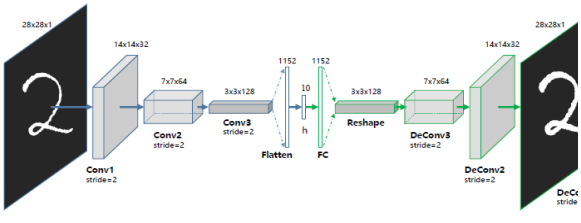


Figure 1: The structure of proposed Convolutional AutoEncoders (CAE) for MNIST [9].

convolutional autoencoders for feature learning in terms of accuracy (acc) and the normalized mutual information (NMI).

DeepCluster [10] is a clustering algorithm developed by the authors that learn both the parameters of a neural network and the cluster assignments of the generated features. DeepCluster uses a typical clustering technique, k-means, to iteratively group the features and uses the following assignments as supervision to update the network's weights. They use DeepCluster to train convolutional neural networks unsupervised on big datasets like ImageNet and YFCC100M, using accuracy criteria evaluation. On all typical benchmarks, the generated model exceeds the present state of the art by a significant margin.

This study's [11] concern is that data representation affects the performance of subspace clustering. Subspace clustering data representation translates data from one space to another with higher separability. In recent years, a slew of new data visualization techniques has emerged. Low-rank representation (LRR) and an autoencoder are two examples. LRR is a low-rank constraint linear representation method that captures the global structure of data. An autoencoder, on the other hand, uses a neural network to nonlinearly map data into a latent space by minimizing the difference between the reconstruction and the output. The authors of this work suggest a unique data representation approach for subspace clustering that combines the benefits of an LRR (globality) and an autoencoder (self-supervision-based locality). The low-rank constrained autoencoder (LRAE) method introduced in this research causes the neural network's latent representation to be of low rank, and the low-rank constraint is derived as a prior from the input space. One of the most significant advantages of the LRAE is that the learned data representation not only preserves the data's local properties but also serves as a precursor to the underlying low-rank global structure. Extensive subspace clustering tests were carried out on a variety of datasets (MNIST, COIL-100, and ORL), using ACC, NMI, and the adjusted rand index (ARI). They showed that the suggested LRAE outperformed state-of-the-art subspace clustering approaches significantly.

The researchers in this paper [12] created a hybrid autoencoder (BAE) model for image clustering by combining three AE-based models: the convolutional autoencoder (CAE), adversarial autoencoder (AAE), and stacking autoencoder (SAE). The MNIST and CIFAR-10 datasets are used to test the suggested models' results and compare them to those of other researchers. The proposed models outperform others in the numerical experiment, according to the clustering criteria: ACC, NMI, and ARI.

GANs have demonstrated great performance in a variety of unsupervised learning problems, and clustering is unquestionably an important unsupervised learning challenge. While the latent-space back-projection in GANs could be used to cluster, they show that the cluster structure is not preserved in the GAN latent space. ClusterGAN is a new mechanism for clustering using GANs proposed by the authors in this study [13]. They achieve clustering in the latent space by sampling latent

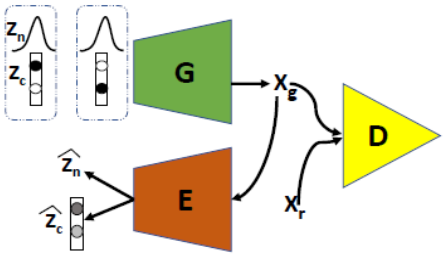


Figure 2: ClusterGAN Architecture [13].

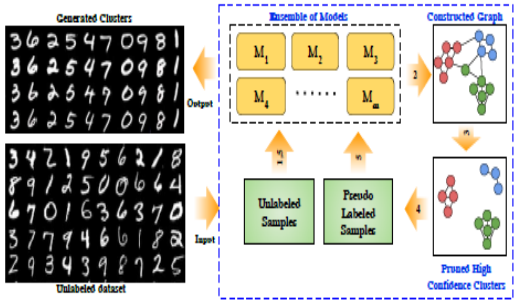


Figure 3: Kingdra overview . They train all the models using the unlabeled samples, in step 1. In step 2, they construct a graph modeling the pairwise agreement of the models. In step 3, they get k high confidence clusters by pruning out data points for which the models do not agree. In step 4 they take the high confidence clusters and generate pseudo labels. In step 5 they train the models using both unlabeled samples and pseudo labeled samples. They iterate from step 2 to step 5 and final clusters are generated [16].

variables from a mixture of one-hot encoded variables and continuous latent variables, together with an inverse network (which projects the data to the latent space) trained jointly with a clustering specific loss. GANs can maintain latent space interpolation across categories, even though the discriminator is never exposed to such vectors, according to their findings. They compared their results to a variety of clustering benchmarks (MNIST, Synthetic, Fashion-10,6 Fashion-5, 10x_73k, and Pendigits) and showed that they outperformed them on both synthetic and real-world datasets according to the following evaluation criteria: ACC, NMI, and ARI.

This work [14] proposes a new approach to this study, in which the embedding is performed using a differentiable model such as a deep neural network. They create a fully differentiable loss function that can be minimized concerning both the embedding parameters and the cluster parameters via stochastic gradient descent by rewriting the k-means clustering method as an optimal transport problem and adding an entropic regularization. They show that by including limits on cluster sizes, this new formulation generalizes a previously suggested state-of-the-art soft-k-means technique. According to empirical evaluations of image classification benchmarks (MNIST, CIFAR-10), their optimum transport-based technique provides greater unsupervised accuracy and does not require a pre-training step when compared to state-of-the-art methods.

The researchers of this work [15] present a deep Generative Adversarial Clustering Network (ClusterGAN) in this publication, which addresses the challenges of unsupervised deep clustering model training. ClusterGAN is made up of three networks that include a discriminator, a generator, and a clustered (i.e. a clustering network). They use an adversarial game between these three players to use the generator to synthesize actual samples given discriminative latent variables, and the clustered to learn the inverse mapping of the real samples to the discriminative embedding space. Furthermore, they use a conditional entropy minimization loss to increase/decrease Intra/inter-cluster sample similarity. Because the ground-truth similarities in the clustering task are unknown, they offer a new balanced self-paced learning algorithm for gradually incorporating data into training from simple to tough while taking into account the diversity of selected samples from all clusters. Their unsupervised learning approach allows them to train clusters with a lot of depth quickly. On numerous datasets (MNIST, USPS, FRGC, CIFAR-10, and STL-10), ClusterGAN produces competitive outcomes when compared to state-of-the-art models, according to experimental results, using accuracy criteria evaluation Acc and NMI.

The main topic of this work [3] is that deep clustering outperforms conventional clustering by combining feature learning and cluster assignment. Although several deep clustering algorithms have been developed for various purposes, the majority of them fail to learn robust cluster-oriented features, resulting in poor final clustering performance. The authors suggest a two-stage deep clustering technique (ASPC-DA) that incorporates data augmentation and self-paced learning to overcome this challenge. They discover robust features in the first stage by training an autoencoder with examples that have been enhanced by random shifting and rotating the clean instances. Then, in the second stage, they alternate between finetuning the encoder with augmented examples and modifying the cluster assignments of the clean examples to encourage the learned features to be cluster-oriented. The center of the cluster to which the clean example is assigned is the target of each augmented example in the loss function during finetuning of the encoder. The targets could be computed improperly, and the encoder network could be misled by instances of inaccurate targets. They use adaptive self-paced learning to select the most confident instances in each iteration to stabilize the network training. Extensive testing shows that their algorithm outperforms the competition on four image datasets (MNIST-full, MNIST-test, USPS, and Fashion) in terms of ACC and NMI.

The authors of this study [16] present a system for improving unsupervised clustering performance using semi-supervised models called Kingdra. To use semi-supervised models, they must first create pseudo-labels, which are automatically generated labels. Prior approaches to creating pseudo-labels have been found to degrade clustering performance due to their low precision.

Instead, they generate a similarity graph using an ensemble of deep networks, from which they extract high-accuracy pseudo labels. The method of employing ensembles to find high-quality pseudo-labels and training the semi-supervised model is iterated, resulting in continual improvement. For numerous image and text datasets, they show that their approach beats state-of-the-art clustering findings. To evaluate their method, they used the accuracy evaluation criteria and five datasets (MNIST, STL, CIFAR10, Reuters, and 20news). They reached 54.6 % accuracy for CIFAR-10 and 43.9 % for 20 news.

In [17], discriminative models are the most common in the literature, and they produce the best results. These algorithms learn a deep discriminative neural network classifier with latent labels. As is common in supervised learning, they typically use multinomial logistic regression posteriors and parameter regularization. Discriminative objective functions (e.g., those based on mutual information or KL divergence) are generally thought to be more flexible than generative approaches (e.g., K-means) in that they make fewer assumptions about data distributions and, as a result, produce much better unsupervised deep learning results. Several contemporary discriminative models may appear to be unrelated to K-means at first glance. Under mild conditions, these models are similar to K-means, common posterior models, and parameter regularization, as demonstrated in this paper. The authors show that maximizing the L2 regularized mutual information via an approximate alternating direction method (MI-ADM) for commonly used logistic regression posteriors is comparable to minimizing a soft and regularized K-means loss. Their theoretical study not only ties numerous recent state-of-the-art discriminative models directly to K-means but also leads to a novel soft and regularized deep K-means algorithm that performs well on a variety of image clustering benchmarks. They used the accuracy and normalized mutual information criteria for the evaluation of five datasets: USPS, MNIST, YTF, CMU-PIE, and FRGC.

The researchers [18] introduced a new clustering objective that develops a neural network classifier from the start using only unlabeled input samples. In eight unsupervised clustering benchmarks spanning image classification and segmentation, the model discovers clusters that accurately match semantic classes, delivering state-of-the-art performance. These include STL10, an unsupervised ImageNet variation, and CIFAR10, which outperformed their closest competitors by 6.6 and 9.5 absolute percentage points, respectively. The strategy isn't limited to computer vision and can be applied to any paired dataset sample; in their studies, they used random transforms to generate a pair from each image. Instead of high-dimensional representations that require further processing to be useable for semantic clustering, the trained network outputs semantic labels directly. The goal is simple: to maximize the mutual information between each pair's class assignments. It's simple to use and is firmly rooted in information theory, so it easily avoids the degenerate solutions that other clustering algorithms are prone to. The experiments used four datasets: STL10, CIFAR10, CIFAR 100-20, and MNIST. They examine two semi-supervised settings in addition to the unsupervised mode. The first achieves a global state-of-the-art of 88.8% accuracy in STL10 classification, surpassing all current approaches (whether supervised, semi-supervised or unsupervised). The second reveals that it can withstand 90 percent reductions in label coverage, which is useful for applications that just need a few labels.

In [19], the authors of this paper discuss a variant of variationally-oriented autoencoders where the superstructure of latent variables is on top of the features of the autoencoders. Their model is based on a tree structure that consists of multiple super latent variables. When there is only one active variable in the superstructure, it generates a model that assumes the latent features of that variable are generated by the Gaussian mixture model. The model, known as the Latent Tree Variational AutomaticEncoder (LTVAE) learns by creating multiple partitions of data, each containing a super latent variable. It is a type of deep learning method that produces multiple partitions of data. This method allows us to partition high-dimensional data into multiple ways. To evaluate this model, they used four datasets: MNIST, STL-10, Reuters, and HHAR, the criteria for clustering accuracy.

In [20], to resolve the problem of high-dimensional dataset clustering difficulties, the authors of this paper describe a clustering approach that simultaneously conducts nonlinear dimensionality reduction and clustering. A deep autoencoder embeds the data in a lower-dimensional space. As part of the clustering process, the autoencoder is optimized. The resulting network generates data that is clustered. The proposed method, Deep Continuous Clustering (DCC) does not rely on knowing the number of ground-truth clusters in advance. The optimization of a global continuous objective is used to combine nonlinear dimensionality reduction and clustering. As a result, they avoid the discrete reconfigurations of the objective that previous clustering algorithms are known for. Experiments on six datasets (MNIST, Coil100, YTF, YaleB, Reuters, and RCV1) using the accuracy evaluation criteria (AMI) show that the proposed approach outperforms current clustering approaches, including deep network-based approaches like k-means, DBSCAN, AC-W, SEC, LDMGI, GDL, and RCC.

Deep clustering through a Gaussian-mixture variational autoencoder (VAE) with Graph embedding is proposed by the authors in [21]. They use the Gaussian mixture model (GMM) as the prior in VAE to make clustering easier. They use graph embedding to handle data with a complicated spread. Their hypothesis is that graph data, which captures local data structures, is a great complement to deep GMM. When they're combined, the network can develop more powerful representations that adhere to global models and local structural restrictions. As a result, their method unites model-based and similarity-based clustering methodologies. They propose a novel stochastic extension of graph embedding to combine graph embedding with probabilistic deep GMM: they consider samples as nodes on a graph and minimize

the weighted distance between their posterior distributions. The distance is calculated using the Jenson-Shannon divergence. They integrate the deep GMM's divergence minimization and log-likelihood maximization. They came up with formulations to achieve a unified goal that allows deep representation learning and clustering to happen at the same time. Their findings on four datasets (MNIST, STL-10, Reuters, and HHAR) in terms of accuracy reveal that their suggested DGG outperforms recent deep Gaussian mixture approaches (model-based) and deep spectral clustering techniques (similarity-based). The benefits of integrating model-based and similarity-based clustering, as advocated in this paper, are highlighted by their findings.

The authors [22] present a shared learning paradigm for discriminative embedding and spectral clustering in this research. To embed the inputs into a latent space for clustering, they first build a dual autoencoder network that enforces the reconstruction requirement for the latent representations and their noisy variants. As a result, the learned latent representations may be more noise-resistant. Then, to give more discriminative information from the inputs, mutual information estimation is used. Furthermore, a deep spectral clustering method is used to embed the latent representations in the eigenspace and then cluster them, allowing for optimal clustering outcomes by fully exploiting the link between inputs. Experiments on benchmark datasets (MNIST-full, MNIST-test, USPS, Fashion-10, and YTF) reveal that their strategy outperforms state-of-the-art clustering algorithms significantly (k-means, NMF,...) in terms of ACC and NMI.

The researchers [23] offer a unique clustering framework called deep comprehensive correlation mining (DCCM) in this paper for analyzing and exploiting various types of correlations behind unlabeled data from three perspectives: 1) Pseudo-label supervision is presented as an alternative to employing only pair-wise information to examine category information and develop discriminative features. 2) The resilience of the features to picture alteration in the input space is completely studied, which aids network learning and boosts performance greatly. 3) For the clustering problem, triplet mutual information among features is introduced to lift the recently discovered instance-level deep mutual information to a triplet-level formation, which aids in the learning of more discriminative features. Extensive tests on a variety of tough datasets (CIFAR-10, CIFAR-100, STL-10, ImageNet-10, Imagenet-dog-15, and Tiny-ImageNet) in terms of ACC, NMI, and adjusted rand index (ARI) reveal that their method works well, with 62.3 % clustering accuracy on CIFAR-10, which is 10.1 % better than the state-of-the-art results (k-means, AE,...).

By jointly maximizing a clustering loss and a non-clustering loss, deep clustering algorithms combine representation learning with clustering. In such systems, a deep neural network is combined with a clustering network to learn representations. Rather than using this framework to increase clustering performance, the researchers [24] offer a simpler method of maximizing the entanglement of an autoencoder's learned latent code representation. They define entanglement as the distance between pairs of points belonging to the same class or structure and pairs of points belonging to different classes or structures. They employ the soft closest neighbor loss and expand it by adding an annealing temperature factor to assess the entanglement of data points. The test clustering accuracy was 96.2% on the MNIST dataset, 85.6% on the Fashion-MNIST dataset, and 79.2% on the EMNIST Balanced dataset when they used their proposed approach, beating their baseline models.

The Matching Priors and Conditionals for Clustering (MPCC) is a GAN-based model featuring an encoder for inferring latent variables and cluster categories from data and a flexible decoder for generating samples from a conditional latent space, according to the researchers of [25]. They show via MPCC that a deep generative model may compete/outperform discriminative approaches in clustering tasks, outperforming the state of the art across a variety of benchmark datasets (MNIST, CIFAR10). In CIFAR10, their tests show that adding a learnable prior and increasing the number of encoder updates improves the quality of the generated samples, resulting in an inception score of 9,49± 0,15 and a 46,9% improvement in the Fréchet inception distance above the state of the art.

The researchers of [26] show that greedy or local methods of maximizing mutual information (such as stochastic gradient optimization) identify local optimal for the mutual information criterion; as a result, the resulting representations are less-than-ideal for complex downstream tasks. This problem has not been identified or addressed in previous research. They introduced deep hierarchical object grouping (DHOG), which generates representations that better optimize the mutual information objective by computing many separate discrete representations of pictures in a hierarchical sequence. They also discovered that these representations are more suited to the task of grouping objects into underlying object classes. They put DHOG to the test on unsupervised clustering, which is a natural downstream test given that the target representation is discrete data labeling. They produced new state-of-the-art scores on the three key benchmarks (CIFAR-100-20, STL-10, and SVHN) without any of the pre-filtering or Sobel-edge detection that many earlier approaches needed to work. They obtained accuracy improvements of 4,3% on CIFAR-10, 1,5% on CIFAR-100-20, and 7,2% on SVHN.

The researchers in this work [27] tackle the problem of Federated Learning (FL), where users are spread and partitioned into clusters. This configuration represents scenarios in which separate groups of users have their own goals (learning tasks), but by aggregating their data with those of others in the same cluster (same learning task), they can take advantage of the power of numbers to execute more efficient Federated Learning. They present the Iterative Federated Clustering Algorithm (IFCA), a new framework that uses gradient descent to estimate user cluster identities and improve model parameters for user clusters. They investigated the algorithm's convergence rate in a linear model with squared loss, as well as for generic strongly convex and smooth loss functions. They demonstrate that IFCA converges at an exponential rate in

both scenarios with good initialization, and they explain the statistical error rate's optimality. They propose training the models by combining IFCA with the weight sharing strategy in multi-task learning when the clustering structure is uncertain. They show that our technique can succeed even if we reduce the initialization criteria by using random initialization and repeated restarts in the tests. They also offer practical data demonstrating the efficiency of our technique in non-convex problems like neural networks. On numerous clustered FL benchmarks (Rotated MNIST, Rotated CIFAR), they show how IFCA outperforms the baselines in terms of precision.

The problem with this work [28] is that unsupervised image classification is a difficult computer vision task. Deep learning-based algorithms have produced excellent results, with the most recent technique using uniform embedding and class assignment losses. Because these processes have distinct goals fundamentally, improving them together may result in a suboptimal solution. To overcome this problem, the researchers suggest the IIC model (Invariant Information Clustering), a novel two-stage approach in which a pretraining embedding module is followed by a refining module that does both embedding and class assignment simultaneously. When evaluated with different datasets (CIFAR-10, CIFAR-100-20, and STL-10), their model outperforms SOTA in unsupervised tasks, with an accuracy of 81.0% for the CIFAR-10 dataset (an increase of 19.3% points), 35.3 % for CIFAR-100-20 (9.6 pp), and 66.5 % for STL-10 (6.9 pp).

Deep clustering has demonstrated an excellent ability to deal with unsupervised learning for structure analysis of high-dimensional visual data by learning visual features and data grouping at the same time. Local learning constraints based on inter-sample relations and/or self-estimated pseudo labels are commonly used in existing deep clustering algorithms. This is vulnerable to unavoidable errors that spread throughout the neighborhood, as well as to error propagation during training. Based on the observation that assigning samples from the same semantic categories into different clusters

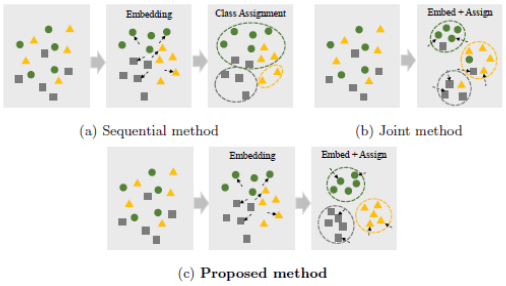


Figure 4: Methods for unsupervised image classification. (a) The sequential method embeds and assigns data points to classes one by one, whereas (b) the joint technique embeds and organizes data points into classes all at once. (c) The proposed technique performs embedding learning as a pretraining step to determine suitable initialization, then optimizes the embedding and class assignment processes simultaneously. During the pretraining stage of their two-stage design, they experience distinctive losses [28].

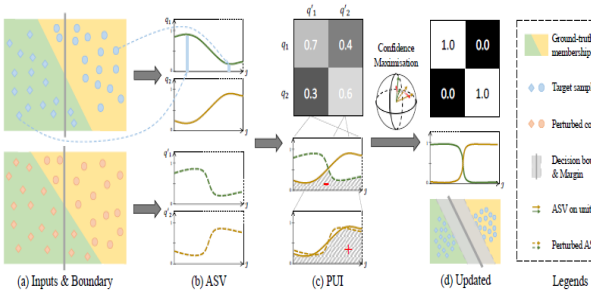


Figure 5: Unsupervised deep clustering using the proposed PartItion Confidence mAximisation (PICA) approach. (a) Given the input data as well as the CNN model's decision bounds, (b) Using a mini-batch of data and its randomly perturbed copy, PICA computes the cluster-wise Assignment Statistics Vector (ASV) in the forward pass. (c) To reduce the partition uncertainty index as much as possible (PUI), (d) PICA is taught to use a specific objective loss function to distinguish the ASV of all clusters on the hypersphere to discover the most confident and potentially promising clustering solution [29].

reduces both intra-cluster compactness and inter-cluster diversity, i.e. lower partition confidence, the authors of [29] propose to solve this problem by learning the most confident clustering solution from all possible separations. In particular, they present PartItion Confidence MAximisation, a unique deep clustering method (PICA). It is based on the principle of learning the most semantically plausible data separation, in which all clusters may be mapped one-to-one to the ground-truth classes, by increasing the "global" partition confidence of the clustering solution. This is accomplished by introducing a differentiable partition uncertainty index and its stochastic approximation, as well as a principled objective loss function that minimizes such an index, all of which, when combined, allow for direct application of traditional deep networks and mini-batch based model training. Extensive testing on six frequently used clustering benchmarks (CIFAR-10, CIFAR-100, STL-10, imageNet-10, ImageNet-dogs, and Tiny-ImageNet) demonstrates that their model outperforms a wide range of state-of-the-art techniques in terms of ACC, NMI, and ARI.

The challenge with this study [30] is that there is no obvious easy-cost function that can capture the major variables of differences and similarities in unsupervised learning. Because natural systems feature smooth dynamics, if an unsupervised objective function remains static during the training process, an opportunity is missed. Smooth dynamics should be introduced in the absence of concrete monitoring. Dynamic goal functions, as opposed to static cost functions, enable greater use of the progressive and unpredictable knowledge gained through pseudo supervision. In this study, they present Dynamic Autoencoder (DynAE), a new deep clustering model that eliminates the clustering reconstruction trade-off by gradually and seamlessly removing the reconstruction objective function in favor of a

construction one. In comparison to the most relevant deep clustering algorithms, experimental evaluations on benchmark datasets (MNIST-full, MNIST-test, USPS, and Fashion-MNIST) reveal that our methodology achieves state-of-the-art outcomes in terms of ACC and NMI.

The problem addressed in this paper [31] is: Clustering with deep autoencoders has received a lot of attention in recent years. Current methods rely on learning embedded features and clustering data points in the latent space at the same time. Although many deep clustering algorithms beat shallow models in achieving good findings on a variety of high-semantic datasets, a major flaw in such models has gone unnoticed. The embedded clustering objective function may distort the latent space by learning from faulty pseudo-labels in the absence of concrete supervisory signals. As a result, the network can learn non-representative features, lowering its discriminative ability and resulting in inferior pseudo-labels. Modern autoencoder-based clustering articles advocate using the reconstruction loss for pretraining and as a regularizer during the clustering phase to mitigate the effect of random discriminative features. Feature Drift can, however, be caused by a clustering reconstruction trade-off. The authors suggest ADEC (Adversarial Deep Embedded Clustering), a novel autoencoder-based clustering model that uses adversarial training to handle a dual problem, namely, Feature Randomness and Feature Drift. They use benchmark real datasets (MNIST-full, MNIST-test, USPS, Fashion-MNIST, Reuters-10K, and Mice Protein) to empirically illustrate the applicability of their model for dealing with these difficulties. The researchers' model outperforms state-of-the-art autoencoder-based clustering approaches in terms of ACC and NMI.

For image clustering, the authors of [32] suggest a self-supervised Gaussian ATtention network (GATCluster). GATCluster delivers semantic cluster labels without further post-processing, rather than extracting intermediate features first and then conducting the standard clustering technique. The Label Feature Theorem is used to ensure that the learned features are one-hot encoded vectors and that trivial solution are avoided. They created four self-learning tasks with the restrictions of transformation invariance, separability maximization, entropy analysis, and attention mapping to train the GATCluster unsupervised. The transformation invariance and separability maximization tasks, in particular, are used to understand the relationships between sample pairs. The goal of the entropy analysis task is to avoid finding simple solutions. They created a self-supervised attention method that incorporates a parameterized attention module and a soft attention loss to capture object-oriented semantics. During the training process, all of the clustering guiding signals are self-generated. Furthermore, they create a memory-efficient two-step learning approach for grouping large-size images. Extensive trials show that their suggested method outperforms the current state-of-the-art image clustering benchmarks (CIFAR-10, CIFAR-100, STL-10, imageNet-10, ImageNet-dogs, and Tiny-ImageNet) in terms of ACC, NMI, and ARI.

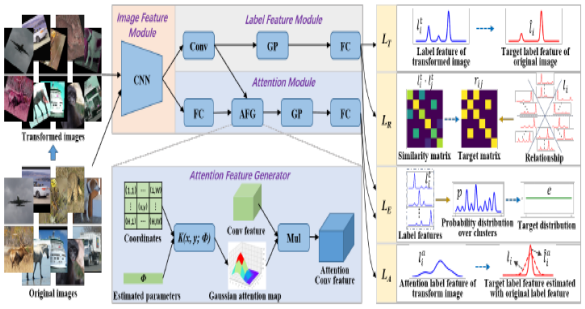


Figure 6: GATCluster framework. CNN is a convolutional neural network, GP means global pooling, Mul represents channel-independent multiplication, Conv is a convolution layer, FC is a fully connected layer, and AFG represents an attention feature generator [32].

Deep learning has recently demonstrated its ability to learn strong feature representations for images. The work of image clustering necessitates appropriate feature representations to capture the data distribution and, as a result, distinguish data points from one another. Often, these two aspects are dealt with independently, and thus, traditional feature learning alone does not suffice in partitioning the data meaningfully. Variational Autoencoders (VAEs) naturally lend themselves to learning data distributions in a latent space. The authors [33] suggest a method based on VAEs that uses a Gaussian Mixture before helping cluster the images appropriately since they seek to efficiently differentiate between distinct clusters in the data. They learn the parameters of both the prior and posterior distributions at the same time. Their method represents a true Gaussian Mixture VAE. In this way, their system learns a prior that captures the latent distribution of the images as well as a posterior that aids in data point discrimination. They also suggest a new reparametrization of the latent space that includes both discrete and continuous variables. One important takeaway is that, unlike existing methods, their method generalizes well across diverse datasets without the use of pre-training or learned models, allowing it to be trained from scratch in an end-to-end manner. They demonstrate our efficacy and generalizability in the lab by achieving state-of-the-art outcomes on a variety of datasets using unsupervised approaches. To the best of their knowledge, they are the first to use VAEs for image clustering on real image datasets (MNIST, Fashion-MNIST, STL-10, CIFAR10, CIFAR100, and FRGCv2) in an unsupervised manner and the accuracy evaluation criteria.

The authors of this research [34] deviate from current work by advocating the SCAN method (Semantic Clustering by Adopting Nearest neighbors), a two-step strategy in which feature learning and clustering are separated. To obtain semantically relevant features, a self-supervised task from representation learning is used first. Second, in a learnable clustering strategy, they employ the collected features as a prior. They accomplish so by removing cluster learning's capacity to rely on low-level

features, which are present in existing end-to-end learning systems. In terms of classification accuracy, they surpass state-of-the-art approaches by substantial margins, with +26,6 % on CIFAR10, +25,0 % on CIFAR100-20, and +21,3 % on STL10, respectively. Furthermore, their technology is the first to successfully classify images on a large-scale dataset.

In this paper [35], the authors offer a new deep image clustering framework for learning a category-style latent representation (Deep Clustering with Category-Style representation (DCCS) for unsupervised image clustering), in which the category information is decoupled from the image style and may be used directly for cluster assignment. Mutual information maximization is used to embed relevant information in the latent representation to achieve this goal. Furthermore, the augmentation-invariant loss is used to separate the representation into two parts: category and style. Last but not least, the latent representation is given a prior distribution to ensure that the elements of the category vector can be used as probabilities over clusters. Extensive tests show that the suggested method significantly outperforms state-of-the-art approaches on a variety of public datasets (MNIST and Fashion-MNIST) in terms of ACC, NMI, and ARI.

The study's authors [36] proposed Deep Robust Clustering (DRC). Unlike existing methods, DRC approaches deep clustering from two perspectives: semantic clustering assignment and representation features, which can simultaneously improve inter-class and intra-class diversities. Furthermore, by examining the internal relationship between mutual information and contrastive learning, they established a generic framework that may change maximizing mutual information into minimizing contrastive loss. They used it to learn invariant features and robust clusters in DRC with great success. Extensive tests on six widely used deep clustering benchmarks (CIFAR-10, CIFAR-100, STL-10, imageNet-10, ImageNet-dogs, and Tiny-ImageNet) show that DRC outperforms them in terms of both stability and accuracy. For example, on CIFAR-10, they achieved a mean accuracy of 71.6%, which is 7.1% higher than current values.

In this research [37], they introduced Contrastive Clustering (CC), a one-stage online clustering algorithm that performs explicit instance-and cluster-level contrastive learning. To be more exact, the positive and negative instance pairs for a given dataset are created using data augmentation and then projected into a feature space. In this case, instance- and cluster-level contrastive learning are carried out in the row and column space, respectively, by maximizing positive pair similarities while minimizing negative pair similarities. Their main finding is that the feature matrix's rows can be thought of as soft labels, for instance, and the columns can be thought of as cluster representations. The model learns representations and cluster assignments in an end-to-end way by maximizing the instance- and cluster-level contrastive loss at the same time. On six challenging image benchmarks (CIFAR-10, CIFAR-100, STL-10, imageNet-10, ImageNet-dogs, and Tiny-ImageNet),

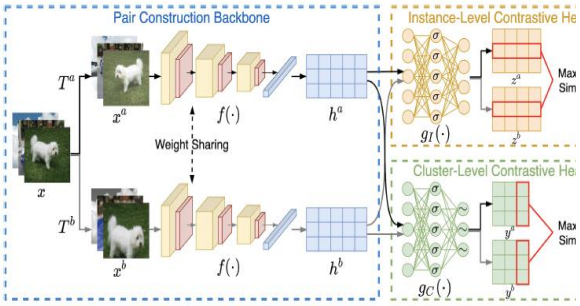


Figure 7: Contrastive Clustering framework. Two data augmentations are used to create data pairs. One shared deep neural network is utilized to extract features from distinct augmentations given data pairs. To project the features into the row and column space, two distinct MLPs (denotes the ReLU activation and denotes the Softmax operation to produce soft labels) are utilized to undertake instance- and cluster-level contrastive learning, respectively [37].

extensive experimental data shows that CC beats 17 competitive clustering approaches. On the CIFAR-10 (CIFAR-100) dataset, in particular, CC obtains an NMI of 0.705 (0.431), which is a performance gain of up to 19% (39%) above the best baseline.

The authors of this paper [38] propose learning an autoencoder embedding and then searching for the underlying manifold using it. They then cluster this using a shallow clustering technique rather than a deeper network for simplicity. They investigated a variety of local and global manifold learning methods on both raw data and autoencoder embeddings, concluding that UMAP in their framework is capable of determining the optimal clusterable manifold of the embedding. This shows that using local manifold learning on an autoencoder embedding to find higher-quality clusters is a good idea. They show numerically that their method outperforms the existing state-of-the-art on a variety of image and time-series datasets (MNIST, MNIST-test, USPS, Fashion, Pendigits, and HAR) including outperforming the current state-of-the-art on numerous in terms of ACC and NMI. They believe these findings point to a viable research direction in deep clustering.

SPICE, a Semantic Pseudo-labeling framework for Image ClustEring, is presented in this work [39]. SPICE generates pseudo-labels by self-learning and directly employs the pseudo-label-based classification loss to train a deep clustering network, rather than requiring indirect loss functions as required by the recently proposed approaches. The core idea behind SPICE is to use a semantically-driven paradigm to improve the clustering network by combining the discrepancy between semantic clusters, similarity across instance samples, and semantic consistency of local samples in an embedding space. To train a clustering network by unsupervised representation learning, a semantic-similarity-based pseudo-labeling approach was presented initially. A local semantic consistency principle is employed to pick a set of consistently labeled samples based on the initial clustering results, and a semi-pseudo-labeling technique (SPICE-

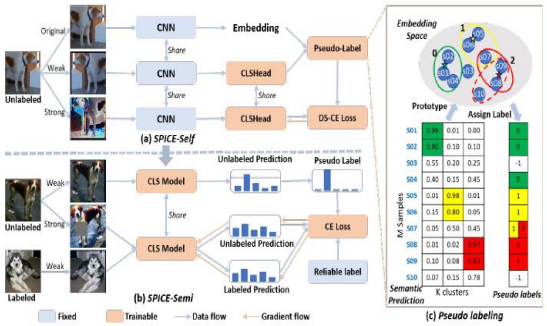


Figure 8: Representation of the SPICE framework. (a) SPICE-Self uses pseudo labeling to train a classification model, with CNN fixed after pretraining using representation learning. (b) SPICE-Semi retrains the classification model by semi-pseudo-labeling, in which reliable labels are chosen from the SPICE-Self findings based on the local consistency of nearby samples. (c) A simple example of pseudo labeling, with red, green, and blue indicating different clusters [39].

Semi) is adopted for performance boosting. On six typical benchmark datasets, including STL10, Cifar10, Cifar100-20, ImageNet-10, ImageNet-Dog, and Tiny ImageNet, extensive studies show that SPICE outperforms existing approaches. In terms of adjusted rand index, normalized mutual information, and clustering accuracy, the proposed SPICE technique improves the existing best results by roughly 10% on average.

Unsupervised image clustering approaches are prone to incorrect predictions and overconfident outcomes since they use alternate objectives to indirectly train the model. To address these issues, the current study [40] provides a new RUC model that is based on resilient learning. RUC is unique in that it uses the pseudo-labels of existing picture clustering algorithms as a noisy dataset with potentially misclassified samples. Its retraining method can correct mismatched knowledge and reduce the problem of overconfidence in forecasts. The model's flexible structure allows it to be used as an add-on module to existing clustering algorithms, allowing them to perform better on a variety of datasets (CIFAR-10, CIFAR-20, STL-10). Extensive studies show that the suggested approach can improve model confidence and gain additional robustness against adversarial noise by properly calibrating it. RUC is a module that may be added to any off-the-shelf unsupervised learning method to improve its performance. RUC is motivated by a desire to learn more. It separates clustered data points into clean and noisy sets before fine-tuning the clustering results. SCAN and TSUC, two state-of-the-art unsupervised clustering algorithms, exhibited considerable performance increases with RUC. (STL-10 : 86.7 %, CIFAR-10 : 90.3 %, CIFAR-20 : 54.3 %).

In the research [41], the authors use instance discrimination and feature decorrelation to propose a clustering-friendly representation learning approach. The principles of classical spectral clustering inspired their deep-learning-based representation learning method. Instance discrimination discovers data commonalities, whereas feature decorrelation eliminates redundant correlation between features. They employ a method of instance discrimination in which knowing individual instance classes leads to learning similarities between examples. They show that the methodology may be extended to learning a latent space for clustering through comprehensive experimentation and examination of the benchmark datasets (CIFAR-10, CIFAR-100, STL-10, ImageNet-10, and ImageNet-Dog). For learning, they create new softmax-formulated decorrelation constraints. Their method achieves an accuracy of 81,5% and 95,4% in image clustering tests using CIFAR-10 and ImageNet-10, respectively. They also demonstrate that the softmax-formulated constraints work with a variety of neural networks.

The authors of this study [42] introduced Mixture of Contrastive Experts (MiCE), a unified probabilistic clustering approach that concurrently uses contrastive learning's discriminative representations and a latent mixture model's semantic structures. MiCE uses a gating function to partition an unlabeled dataset into subsets according to latent semantics and numerous experts to differentiate separate subsets of instances allotted to them in a contrastive learning method, which is motivated by the mixing of experts. They designed a scalable form of the Expectation-Maximization (EM) algorithm for MiCE and showed proof of convergence to overcome the nontrivial inference and learning challenges caused by latent variables. They tested MiCE's clustering performance empirically on four frequently used natural image datasets (CIFAR-10, CIFAR-100, STL10, and ImageNet-Dog). MiCE outperforms a variety of earlier approaches and provides a strong contrastive learning baseline using the criteria ACC, NMI, and ARI.

The problem with this study [43] is that, as measured by curated class-balanced datasets, unsupervised feature learning has made significant progress with contrastive learning based on instance discrimination and invariant mapping. Natural data, on the other hand, maybe highly linked and skewed. The supposed instance distinction clashes with natural between-instance similarity, resulting in inconsistency in training and poor performance. The goal is to identify and integrate between-instance similarity into contrastive learning via cross-level discrimination (CLD) between instances and local instance groups rather than instance grouping directly. While attraction inside each instance's augmented perspectives forces invariant mapping, between-instance similarity comes via common repulsion against instance groupings. The batch-wise and cross-view comparisons also help to increase contrastive learning's positive/negative sample ratio and produce improved invariant mapping. We impose both grouping and discrimination objectives on characteristics obtained separately from a shared representation to achieve both goals. For the first time, they also present normalized projection heads and unsupervised hyper-parameter adjustment. CLD is a lean and powerful add-on to existing methods (e.g., NPID, MoCo, InfoMin, BYOL) on highly correlated, long-tail, or balanced datasets, as demonstrated by considerable experimentation. It not only sets new

benchmarks (CIFAR-10, CIFAR-100, and ImageNet) for self-supervision, semi-supervision, and transfer learning, but it also outperforms MoCo v2 and SimCLR on every reported performance achieved with a far larger compute in terms of accuracy. Unsupervised learning is effectively extended to natural data with CLD, bringing it closer to real-world applications.

4 Discussion

Based on this short and selective survey of deep clustering algorithms, we make the following observations:

- most deep clustering techniques have been tested in the area of image recognition;
- performances of these techniques are great in terms of recognition accuracy, as the study of [35], where obtained recognition accuracy achieves 98.9 %;
- most studies enhance the embedding of the data into a lower-dimensional space;
- several researchers use the MNIST database for experimentation and k-means algorithm for results comparison;
- we remark that the appearance of the hybrid version of Autoencoder gives satisfactory results too ;
- deep learning is a technology that continues to mature and has been applied to pattern recognition to great effect;
- we have identified the name of the proposed method, the category to which it belongs, a dataset of each approach with the methods of comparison, these are seen in table 1;
- Table 1 summarizes the sorted works in chronological order. We observed in Table 1, that the MNIST dataset provides good results compared to other databases like USPS; CIFAR-10; CIFAR-100;

Table 1: General comparison of various deep clustering algorithms for image recognition.

References	Method	Category	Dataset	Compared results with	Obtained results
[24]	SNNL Soft Nearest Neighbor Loss	AE	MNIST; Fashion-MNIST; and EMNIST Balanced.	SNNL-2; SNNL-4; Baseline AE; DEC; VaDE; N2D; and ClusterGAN;....	1. The best accuracy (acc)=96.2% with MNIST; 2. The best NMI=90.3% with MNIST; 3. The best ARI=91.8% with MNIST;
[25]	MPCC Matching Priors and Conditionals for Clustering	AE	MNIST; Onmiglot; FMNIST; CIFAR-10; and CIFAR-20.	DEC; VADE; InfoGAN; ClusterGAN; DAC; IMSAT (VAT); ADC; SCAE; and IIC.	The best accuracy (acc)= 98.76 ± 0.03% with MNIST;
[10]	DeepCluster is a new clustering strategy for large-scale end-to-end convent training.	AE	ImageNet; Places.	The methods have a standard AlexNet architecture.	The best is 73.7% on classification with deepCluster
[11]	Low-rank Constrained Deep Autoencoder for Subspace Clustering (LRAE)	AE	MNIST; COIL-100, and ORL	SSC; LRR; LRSC; LSR; AESC, and PARTY.	1. The best accuracy (acc)= 81.49 ± 2.19 with ORL; 2. The best NMI= 90.77 ± 2.01 with ORL; 3. The best ARI= 73.92 ± 2.11 with ORL;
[12]	Hybrid Autoencoder (BAE), the combination of three AE-based models—the convolutional autoencoder (CAE), adversarial autoencoder (AAE), and stacked autoencoder (SAE)	AE	MNIST and CIFAR-10.	Fuzzy objective function algorithm (FCM), Spectral clustering algorithm (SC), Low-rank representation algorithm (LRR), LSR1 and LSR2 are the variants of the least-squares regression (LSR), SLRR is the scalable LRR, LSC-R and LSC-K are the variants of the large-scale spectral clustering (LSC) algorithms, NMF is the non-negative matrix factorization algorithm, ZAC is the Zeta function based agglomerative clustering algorithm, and DEC is the deep embedding clustering algorithm.	1. The best accuracy (acc)= 83.67% with MNIST; 2. The best NMI= 80.85% with MNIST;

References	Method	Category	Dataset	Compared results with	Obtained results
[14]	Clustering with Optimal Clustering (OT) is a new approach where the embedding is performed by a differentiable model such as a deep neural network	GAN	MNIST and CIFAR10	k-means; AE + k-means; soft k-means and soft k-means (p)	The best NMI= 85.10% with MNIST;
[15]	ClusterGAN is a deep Generative Adversarial Clustering Network.	GAN	MNIST; USPS; FRGC; CIFAR-10 and STL-10.	Kmeans; N-Cuts; SC-LS; AC-PIC; SEC and LDMGI.	1. The best accuracy (acc)= 97% with USPS; 2. The best NMI= 93.10% with USPS; 3. The best accuracy (acc)= 96.4% with MNIST; 4. The best NMI= 92.10% with MNIST;
[27]	IFCA a new framework dubbed the Iterative Federated Clustering Algorithm	AE	Rotated MNIST; and Rotated CIFAR	The global model for IFCA; and local model	The best accuracy (acc)= 95.25 ± 0.40% with Rotated MNIST;
[9]	Deep Convolutional Embedded Clustering (DCEC)	AE	MNIST -full; MNIST-test; USPS	1. Deep Embedded Clustering (DEC); 2. K-means; 3. Stacked AutoEncoders (SAE).	1. The best accuracy (acc)=88.97% with MNIST-full; 2. The best NMI=88.49% with MNIST-full.
[3]	ASPC-DA is an Adaptive Self-Paced deep Clustering with Data Augmentation	AE	MNIST-full; MNIST-test; USPS and Fashion		1. The best accuracy (acc)=98.8% with MNIST-full; 2. The best NMI=96.6% with MNIST-full.
[16]	Kingdra is a framework that leverages semi-supervised models	AE	MNIST; STL; CIFAR10; Reuters and 20news.	k-means; AC; DEC; Deep RIM; and IMSAT...	The best accuracy (acc)=98.5% with MNIST.
[28]	A novel two-stage algorithm in which an embedding module for pretraining precedes a refining module that concurrently performs embedding and class assignment	AE	CIFAR-10; CIFAR-20; and STL-10	Random network; k-means; Autoencoder (AE); SWWAE; GAN; JULE; DEC; DAC; DeepCluster; ADC; and IIC	The best accuracy (acc)= 81% with CIFAR-10;
[29]	PICA a novel deep clustering method named PartItion Confidence mAximisation	AE	CIFAR-10; CIFAR-100; STL-10; ImageNet-10; ImageNet-Dogs and Tiny-ImageNet	K-means; SC; AC; NMF; AE; DAE; and IIC;...	1. The best accuracy (acc)= 87% with ImageNet-10; 2. The best NMI= 80.2% with ImageNet-10; 3. The best ARI= 76.1% with ImageNet-10;
[17]	MIADM is an approximate alternating direction method.	AE	USPS; MNIST-test; MNIST-full; YTF; CMU-PIE and FRGC	SR-K-means; DEPICT; DCN (K-means based) and DEC (KL based).	1. The best accuracy (acc)=97.9% with USPS; 2. The best NMI=94.8% with USPS.
[18]	IIC Invariant Information Clustering	AE	STL10; CIFAR10; CFR100-20 and MNIST.	Random network; Kmeans; Spectral clustering; Triplets; Variational Bayes AE and DeepCluster 2018,.....	The best accuracy (acc)=99.2% with MNIST;

[19]	LTVAE latent tree variational autoencoder.	VAE	MNIST; STL-10; Reuters and HHAR	AE+GMM; VAE+GMM; DEC and DCN.	The best accuracy (acc)=90% with STL-10;
[37]	CC Contrastive Clustering is an online clustering method	AE	CIFAR-10; CIFAR-100; STL-10; ImageNet-10; ImageNet-Dogs; and Tiny-ImageNet	k-means; SC; AC; NMF; DEC; JULE; VaE; DCGAN; DeCNN; DCCM; IIC; and PICA;...	1. The best accuracy (acc)= 89.3% with ImageNet-10; 2. The best NMI= 85.9% with ImageNet-10; 3. The best ARI= 82.2% with ImageNet-10;
[38]	N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding	AE	MNIST; MNIST-test; USPS; Fashion; pendigits; and HAR	k-means; SC; GMM; DEC; DCN; JULE; VaDE; DEPICT; DBC; and ASPC-DA;...	1. The best accuracy (acc)= 97.9% with MNIST; 2. The best NMI= 94.2% with MNIST;
[30]	DynAE Dynamic Autoencoder, a novel model for deep clustering that addresses a clustering–reconstruction trade-off.	AE	MNIST-full; MNIST-test; USPS; and Fashion-MNIST	K-Means; GMM; LSNMF; AC; SSC-OMP; EnSC; LMVSC; RBF K-Means −; DEC; JULE; and DEPICT;....	1. The best accuracy (acc)= 98.7% with MNIST-full; 2. The best NMI= 96.4% with MNIST-full;
[31]	ADEC (Adversarial Deep Embedded Clustering) is a novel autoencoder-based clustering model	AE	MNIST-full; MNIST-test; USPS; Fashion-MNIST; REUTERS-10K; and Mice Protein	DEC*; IDEC*; k-means; GMM; LSNMF; AC; RBF k-means;	1. The best accuracy (acc)= 98.6% with MNIST-full; 2. The best NMI= 96.1% with MNIST-full;
[13]	ClusterGAN method is a new mechanism for clustering using GANs (Generative Adversarial Networks)	GAN	Synthetic data; MNIST; Fashion-MNIST; 10x_73k and Pendigits.	WGAN (normal); WGAN (One-Hot) and Info GAN.	The best accuracy (acc)= 95% with MNIST;
[32]	SPICE, a Semantic Pseudo-labeling framework for Image ClustEring	AE	STL10; ImageNet-10; ImageNet-Dog-15; Cifar10; Cifar100-20; and Tiny-ImageNet-200	JULE; DEC; DAC; DeepCluster; DDC; IIC; DCCM; GATCluster; PIC; and CC	1. The best accuracy (acc)= 93.8% with STL10; 2. The best NMI= 87.2% with STL10; 3. The best ARI= 87% with STL10;
[40]	RUC is inspired by robust learning. RUC's novelty is at utilizing pseudo-labels of existing image clustering models as a noisy dataset.	AE	CIFAR-10; CIFAR-20; and STL-10	k-means; SC; Triplets; AE; GAN; JULE; DAC; DEC; DeepCluster; IIC; TSUC and SCAN;...	The best accuracy (acc)= 90.1% with CIFAR-10;
[33]	A method based on VAEs where we use a Gaussian Mixture before help cluster the images accurately	VAE	STL-10; CIFAR10; MNIST; and Fashion-MNIST	k-means; AE+k-means; and DEC	The best accuracy (acc)= 98.4% with MNIST;
[20]	DCC Deep Continuous Clustering	AE	MNIST; Coil100; YTF; YaleB; Reuters and RCV1	k-means++; AC-W; DBSCAN; SEC and LDMGI;.....	1. The best accuracy (acc)=91.3% with MNIST; 2. The best accuracy (acc)=98.5% with YaleB;

[41]	IDFD a clustering-friendly representation learning method using instance discrimination and feature decorrelation.	AE	CIFAR-10; CIFAR-100; STL-10; Imagenet-10; and Imagenet-Dog	AE; DEC; DAC; DCCM; ID; IIC; IDFO; and SCAN	1. The best accuracy (acc)= 95.4% with ImageNet-10; 2. The best NMI= 89.8% with ImageNet-10; 3. The best ARI= 90.1% with ImageNet-10;
[42]	MiCE Mixture of Contrastive Experts, a unified probabilistic clustering framework	AE	CIFAR-10; CIFAR-100; STL-10; and Imagenet-Dog	K-means; AE; DHOG; DAC; DCCM; MMDC; IIC; IDFO; and MoCo	1. The best accuracy (acc)= 83.5% with CIFAR-10; 2. The best NMI= 73.7% with CIFAR-10; 3. The best ARI= 69.8% with CIFAR-10;
[34]	SCAN Semantic Clustering by Adopting Nearest neighbors	AE	CIFAR10; CIFAR100- 20; STL10; and ImageNet	k-means; SC; Triplets; JULE; AEVB; SAE; DAE; GAN; DAC; and IIC	1. The best accuracy (acc)= 88.3% with CIFAR10; 2. The best NMI= 79.7% with CIFAR10; 3. The best ARI= 77.2% with CIFAR10;
[43]	CLD cross-level discrimination	AE	STL10; CIFAR10; CIFAR100; and ImageNet100	DeepCluster; MoCo; Exemplar; Inv. Spread; NPID; and BYOL;....	1. The best retrieval= 78.6% with CIFAR-10; 2. The best NMI= 69% with CIFAR-10; 3. The best kNN= 86.7% with CIFAR-10;
[23]	DCCM is a deep comprehensive correlation mining	AE	CIFAR-10; CIFAR-100; STL-10; ImageNet-10; Imagenet-dog-15; and Tiny-ImageNet.	K-means; SC; AC; NMF; AE; and DAE;.....	1. The best accuracy (acc)=60.8% with ImageNet-10; 2. The best NMI=71% with ImageNet-10; 3. The best ARI=55.5% with ImageNet-10;
[21]	DGG: Deep clustering via a Gaussian mixture variational autoencoder (VAE) with Graph embedding	VAE	MNIST; STL-10; Reuters and HHAR.	AE+GMM; DEC; IMSAT; VaDE; SpectralNet; and LTVAE.	The best accuracy (acc)=97.58±0.1% with MNIST;
[22]	A joint learning framework for discriminative embedding and spectral clustering	AE	MNIST-full; MNIST-test; USPS; Fashion-10; and YTF.	K-means; SC-Ncut; SC-LS; NMF; AC-GDL; and DASC;.....	1. The best accuracy (acc)=98% with MNIST-test; 2. The best NMI=94.6% with MNIST-test;
[35]	DCCS a novel deep image clustering framework to learn a category-style latent representation	AE	MNIST; and Fashion-MNIST	k-means; SC; AC; NMF; DEC; JULE; VaDE; DEPICT; IMSAT; ClusterGan; IIC; and DLS-clustering;...	1. The best accuracy (acc)= 98.9% with MNIST; 2. The best NMI= 97% with MNIST; 3. The best ARI= 97.6% with MNIST;
[36]	DRC Deep Robust Clustering	AE	CIFAR-10; CIFAR-100; STL-10; ImageNet-10; Imagenet-dog-15; and Tiny-ImageNet	k-means; SC; AC; NMF; DEC; JULE; VaDE; DEPICT; IMSAT; DCCM; IIC; and PICA;...	1. The best accuracy (acc)= 88.4% with ImageNet-10; 2. The best NMI= 83% with ImageNet-10; 3. The best ARI= 79.8% with ImageNet-10;

5 Proposed taxonomy of deep clustering

Figure 9 illustrates the taxonomy of Deep Clustering techniques that we describe, which in turn indicates the study's structure. The basic algorithmic structure, network architecture, loss functions, and training optimization methodologies for deep clustering systems vary (or learning the parameters).

We focus on deep learning for clustering approaches in this paper, where those approaches either use deep learning for grouping (or partitioning) the data and/or creating low-rank deep representations or embeddings of

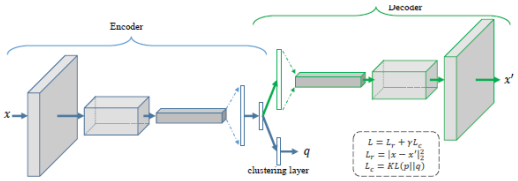


Figure 10: The structure of deep convolutional embedded clustering (DCEC). It is composed of convolutional autoencoders and a clustering layer connected to the embedded layer of autoencoders [9].

the data, which could play a significant supporting role as a building block of supervised learning, among other goals. There are numerous approaches to developing a taxonomy of deep clustering algorithms; in this study, we took the approach of seeing the methods as a process. As a result, we provide a simplified taxonomy based on deep clustering algorithms' overall procedural structure or architecture. Beginners and experienced readers will benefit from the simplified classification.

We have chosen to propose to divide deep learning into three categories:

AE-based deep clustering:

Artificial neural networks (ANNs) are a type of machine learning model made up of numerous nodes grouped in layers that compute an output depending on node activation mediated by weights in the connections between them. ANNs are capable of solving a variety of machine learning tasks, including classification, regression, and dimensionality reduction [44].

A neural network that has been trained to duplicate its input to its output is called an autoencoder. It has a hidden layer h on the inside that defines the code used to represent the input. The network is made up of two parts: an encoder function h=f(x) and a decoder function r=g(h) that provides a reconstruction. Figure 10 illustrates this architecture. If an autoencoder only succeeds in learning to set g(f(x)) =x everywhere, it isn't particularly useful. Autoencoders, on the other hand, are meant to be incapable of flawless copying. They are usually limited in some way, allowing them to copy only roughly and only input that closely mimics the training data. Because the model must prioritize which features of the input should be duplicated, it frequently discovers interesting data attributes. The following is an overview of representative methods of Autoencoder:

1. Deep Convolutional Embedded Clustering (DCEC): the DCEC system is composed of Convolutional Clustering (CAE) and a clustering layer that is connected to the embedded layer of CAE [9]. Each embedded point z_i of the input image xi is mapped into a soft label by the clustering layer. The Kullback-Leibler divergence (KL divergence) between the distribution of soft labels and the precisely defined distribution is then defined as the clustering loss Lc. The clustering loss leads the embedded features to be resistant to forming clusters, and CAE is used to learn embedded features.

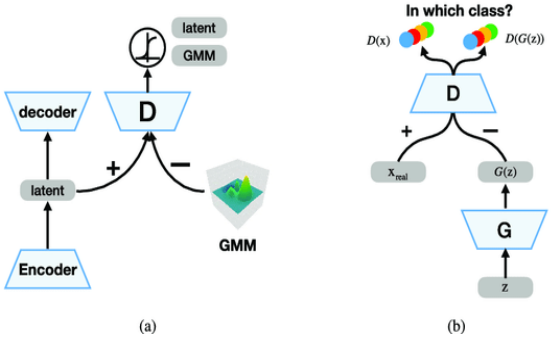


Figure 11: GAN-based deep clustering [47].

The objective of DCEC is:

$$L = L_r + \gamma L_c \quad (1)$$

where L_r and L_c are reconstruction loss and clustering loss respectively, and $\gamma > 0$ is a coefficient that controls the degree of distorting embedded space. When $\gamma = 1$ and $L_r \equiv 0$, (1) reduces to the objective of DEC.

2. Deep Clustering Network (DCN): this method [45] which combines the autoencoder and the k-means algorithm, is one of the most remarkable in the field. It pre-trains an autoencoder in the first stage. The reconstruction loss and the k-means loss are then optimized together. Because k-means relies on discrete cluster assignments, it necessitates the employment of a different optimization procedure. When compared to other methods, DCN's goal is simple, and the computing complexity is modest;
3. Hybrid Autoencoder (HAE): [46] CAE (convolutional autoencoder), VAE (adversarial autoencoder), and SAE (stacked autoencoder) combine the advantages of three autoencoders to learn low and high-level feature representation.

GAN-based deep clustering:

In recent years, the Generative Adversarial Network (GAN) has become a popular deep generative model. A min-max adversarial game is established between two neural networks in the (GAN) [47]: a generating network, G, and a discriminative network, D. The generative network attempts to map a sample z from a prior distribution p(z) to the data space, whereas the discriminative network attempts to compute the probability that an input is a real sample from the data distribution rather than one created by the generative network. GAN is an exciting idea since it offers an adversarial approach to matching the distribution of data or its representations to an arbitrary prior distribution.

VAE- based deep clustering:

[48] VAE is a generative variant of AE since it causes AE's latent code to follow a predetermined distribution. VAE blends variational Bayesian approaches with neural network flexibility and scalability. It applies neural networks to the conditional posterior and uses stochastic gradient descent and standard backpropagation to optimize the variational inference objective. It employs the reparameterization of the variational lower bound to produce a simple, differentiable, unbiased lower bound

estimator. In nearly every model with continuous latent variables, this estimator can be utilized for efficient approximate posterior inference:

1. Deep clustering via a Gaussian mixture VAE with Graph embedding (DGG): [21] a new VAE-based model that assumes the latent variables have a tree structure;
2. Variational Deep Embedding (VaDE): introduces a VAE-based generative model that assumes the latent variables are a mixture of Gaussians with trainable means and variances [49];
3. Latent Tree Variational Autoencoder (LTVAE): a VAE-based model that assumes the latent variables have a tree structure [19].

6 Conclusion and perspectives

Deep learning is made up of a number of well-known and effective models that are used to solve a variety of problems [50].

In the context of deep clustering, we have presented, in this article, an introductory study of the main deep unsupervised learning algorithms that have been found in the last 3-4 years in the literature.

We have presented an overview of clustering methods and algorithms for deep learning. We noticed the multitude of contributions developed in the area of image recognition and we studied and synthesized different recent works in this context.

We have proposed a taxonomy of clustering with deep learning algorithms based on previous studies and some treated representative methods in the survey.

This study is the first step of our research for which we can consider several future extensions, such as exploring the possibilities of hybridization between different deep clustering approaches and their application in evolving patterns. We will be able to make a comparative study of the performance of deep learning approaches based on the autoencoder, such as the work of [51]. We will be able to apply the deep clustering method in fields such as face recognition, etc [52].

Acknowledgment

The authors would like to thank the DGRSDT (General Directorate of Scientific Research and Technological Development) - MESRS (Ministry of Higher Education and Scientific Research), ALGERIA, for the financial support of LISCO Laboratory.

References

[1] A. Chefrour, and L. Souici-Meslati (2019). AMF-IDBSCAN: Incremental Density Based Clustering Algorithm Using Adaptive Median Filtering Technique. *Informatica*, vol. 43(4). https://doi.org/10.31449/inf.v43i4.2629

[2] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, vol. 6, pp. 39501-39514. https://doi.org/10.1109/access.2018.2855437

[3] X. Guo, X. Liu, E. Zhu, X. Zhu, M. Li, X. Xu, and J. Yin (2019). Adaptive self-paced deep clustering with data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, vol. 32(9), pp. 1680-1693. https://doi.org/10.1109/tkde.2019.2911833

[4] C. C. Wang, K. L. Tan, C. T. Chen, Y. H. Lin, S. S. Keerthi, D. Mahajan, and C. J. Lin (2018). Distributed newton methods for deep neural networks. *Neural computation*, vol. 30(6), pp. 1673-1724. https://doi.org/10.1162/neco_a_01088

[5] Z. Shen, H. Yang, and S. Zhang, S (2021). Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, vol. 33(4), pp. 1005-1036. https://doi.org/10.1162/neco_a_01364

[6] J. Gao, P. Li, Z. Chen, and J. Zhang (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, vol. 32(5), pp.829-864. https://doi.org/10.1162/neco_a_01273

[7] E. Aljalbout, V. Golkov, Y. Siddiqui, M., Strobel, and D. Cremers (2018). Clustering with deep learning: Taxonomy and new methods. *arXiv preprint arXiv*:1801.07648.

[8] G. C. Nutakki, B. Abdollahi, W. Sun, and O. Nasraoui (2019). An introduction to deep clustering. *Clustering Methods for Big Data Analytics* , pp. 73-89. Springer, Cham. https://doi.org/10.1007/978-3-319-97864-2_4

[9] X. Guo, X. Liu, E Zhu, and J. Yin (2017, November). Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, pp. 373-382. Springer, Cham. https://doi.org/10.1007/978-3-319-70096-0_39

[10] M. Caron, P. Bojanowski, A. Joulin, and M. Douze (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)* , pp. 132-149. https://doi.org/10.1007/978-3-030-01264-9_9

[11] Y. Chen, L. Zhang, and Z. Yi (2018). Subspace clustering using a low-rank constrained autoencoder. *Information Sciences*, vol. 424, pp. 27-38. https://doi.org/10.1016/j.ins.2017.09.047

[12] P. Y. Chen, and J. J. Huang (2019). A hybrid autoencoder network for unsupervised image clustering. *Algorithms*, vol. 12(6), pp.122. https://doi.org/10.3390/a12060122

[13] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan (2019, July). ClusterGAN: Latent space clustering in generative adversarial networks. In *Proceedings of*

the AAAI Conference on Artificial Intelligence , Vol. 33, No. 01, pp. 4610-4617.

[14] A. Genevay, G. Dulac-Arnold, and J. P. Vert (2019). Differentiable deep clustering with cluster size constraints.

[15] K. Ghasedi, X. Wang, C. Deng, and H. Huang (2019). Balanced self-paced learning for generative adversarial clustering network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4391-4400. https://doi.org/10.1109/CVPR.2019.00452

[16] Gupta, D., Ramjee, R., Kwatra, N., & Sivathanu, M. (2019, September). Unsupervised clustering using pseudo-semi-supervised learning. In *International Conference on Learning Representations*.

[17] M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed, (2019). Deep clustering: On the link between discriminative models and k-means. *IEEE transactions on pattern analysis and machine intelligence*. https://doi.org/10.1109/TPAMI.2019.2962683

[18] X. Ji, J. F. Henriques, and A. Vedaldi (2019). Invariant information clustering for unsupervised image classification and segmentation. https://doi.org/10.1109/ICCV.2019.00996

[19] X. Li, Z. Chen, L. K. Poon, and N. L. Zhang (2018). Learning latent superstructures in variational autoencoders for deep multidimensional clustering. *arXiv preprint arXiv*:1803.05206. iclr 2019.

[20] S. A. Shah, and V. Koltun (2018). Deep continuous clustering. *arXiv preprint arXiv*:1803.01449.

[21] L. Yang, N. M. Cheung, J. Li, and J. Fang (2019). Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6440-6449. https://doi.org/10.1109/ICCV.2019.00654

[22] X. Yang, C. Deng, F. Zheng, J. Yan, and W. Liu (2019). Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4066-4075

[23] J. Wu, K. Long, F. Wang, C. Qian, C. Li, Z. Lin, and H. Zha (2019). Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8150-8159. https://doi.org/10.1109/ICCV.2019.00824

[24] A. F. Agarap, and A. P. Azcarraga (2020, July). Improving k-Means Clustering Performance with Disentangled Internal Representations. In *2020 International Joint Conference on Neural Networks (IJCNN)* , pp. 1-8. IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207192

[25] N. Astorga, P. Huijse, P. Protopapas, and P. Estévez(2020, August). MPCC: Matching Priors and Conditionals for Clustering. In *European Conference on Computer Vision* , pp. 658-677. Springer, Cham. https://doi.org/10.1007/978-3-030-58592-1_39

[26] L. N. Darlow, and A. Storkey (2020). DHOG: Deep Hierarchical Object Grouping. *arXiv preprint arXiv*:2003.08821.

[27] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran (2020). An efficient framework for clustered federated learning. In *34th Conference on Neural Information Processing Systems* (NeurIPS 2020), Vancouver, Canada.

[28] S. Han, S. Park, S. Park, S. Kim, and M. Cha (2020, August). Mitigating Embedding and Class Assignment Mismatch in Unsupervised Image Classification. In *16th European Conference on Computer Vision, ECCV 2020*. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-58586-0_45

[29] J. Huang, S. Gong, and X. Zhu (2020). Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8849-8858. https://doi.org/10.1109/CVPR42600.2020.00887

[30] N. Mrabah, N. M. Khan, R. Ksantini, and Z. Lachiri (2020). Deep clustering with a Dynamic Autoencoder: From reconstruction towards centroids construction. *Neural Networks*, vol. 130, pp. 206-228. https://doi.org/10.1016/j.neunet.2020.07.005

[31] N. Mrabah, M. Bouguessa, and R. Ksantini (2020). Adversarial deep embedded clustering: on a better trade-off between feature randomness and feature drift. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2020.2997772

[32] C. Niu, J. Zhang, G. Wang, and J. Liang (2020, August). Gatcluster: Self-supervised gaussian-attention network for image clustering. In *European Conference on Computer Vision* , pp. 735-751. Springer, Cham. https://doi.org/10.1007/978-3-030-58595-2_44

[33] V. Prasad, D. Das, and B. Bhowmick (2020, July). Variational Clustering: Leveraging Variational Autoencoders for Image Clustering. In *2020 International Joint Conference on Neural Networks* (IJCNN) , pp. 1-10. IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207523

[34] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool (2020, August).

Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pp. 268-285. Springer, Cham. https://doi.org/10.1007/978-3-030-58607-2_16

[35] J. Zhao, D. Lu, K. Ma, Y. Zhang, and Y. Zheng (2020, August). Deep Image Clustering with Category-Style Representation. In *European Conference on Computer Vision* , pp. 54-70. Springer, Cham. https://doi.org/10.1007/978-3-030-58568-6_4

[36] H. Zhong, C. Chen, Z. Jin, and X. S. Hua (2020). Deep robust clustering by contrastive learning. *arXiv preprint arXiv*:2008.03030.

[37] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng (2021). Contrastive Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35(10), pp. 8547-8555.

[38] R. McConville, R. Santos-Rodriguez, R. J., Piechocki, and I. Craddock (2021, January). N2d:(not too) deep clustering via clustering the local manifold of an autoencoded embedding. In *25th International Conference on Pattern Recognition* (ICPR), pp. 5145-5152. IEEE. https://doi.org/10.1109/ICPR48806.2021.9413131

[39] C. Niu, and G. Wang (2021). SPICE: Semantic Pseudo-labeling for Image Clustering. *arXiv preprint arXiv*:2103.09382.

[40] S. Park, S. Han, S. Kim, S. Kim, D. Park, S. Hong, and M. Cha (2021). Improving Unsupervised Image Clustering With Robust Learning. *Accepted at Computer Vision and Pattern Recognition* (cs.CV); *Artificial Intelligence* (cs.AI); *Machine Learning* (cs.LG).

[41] Y. Tao, K. Takagi, and K. Nakata (2021). Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation. *arXi preprint arXiv:2106.00131.ICLR 2021 Workshop on Embodied Multimodal Learning* (EML).

[42] T. W. Tsai, C. Li, and J. Zhu (2021). MiCE: Mixture of Contrastive Experts for Unsupervised Image Clustering. *ICLR 2021*.

[43] X. Wang, Z. Liu, and S. X. Yu (2021). Unsupervised Feature Learning by Cross-Level Instance-Group Discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12586-12595.

[44] L. Amado, and F. Meneguzzi (2018). Q-Table compression for reinforcement learning. *The Knowledge Engineering Review*, vol. 33. https://doi.org/10.1017/S0269888918000280

[45] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong (2017, July). Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *International conference on machine learning*, pp. 3861-3870.

[46] K. Gupta, M. Y. Raghuprasad, and P. Kumar (2018). A hybrid variational autoencoder for collaborative filtering. *arXiv preprint arXiv*:1808.01006.

[47] F. Shoeleh, N. M. Yadollahi, and M. Asadpour. (2020). Domain adaptation-based transfer learning using adversarial networks. *The Knowledge Engineering Review*, vol. 35. https://doi.org/10.1017/S0269888920000107

[48] K. L. Lim, X. Jiang, and C. Yi (2020). Deep clustering with variational autoencoder. *IEEE Signal Processing Letters*, vol. 27, pp. 231-235. https://doi.org/10.1109/LSP.2020.2965328

[49] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou (2016). Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv*:1611.05148

[50] W. Etaiwi, D. Suleiman, and A. Awajan (2021). Deep Learning Based Techniques for Sentiment Analysis: A Survey. *Informatica*, vol. 45(7). https://doi.org/10.31449/inf.v45i7.3674

[51] A. S. Gaafar, J. M. Dahr, and A. K. Hamoud (2022). Comparative Analysis of Performance of Deep Learning Classification Approach based on LSTM-RNN for Textual and Image Datasets. *Informatica*, vol. 46(5). https://doi.org/10.31449/inf.v46i5.3872

[52] H. Ni. (2020). Face recognition based on deep learning under the background of big data. *Informatica*, vol. 44(4). https://doi.org/10.31449/inf.v44i4.3390

Remote Monitoring of Lab Experiments to Enhance Collaboration Between Universities

Arshad Ali
E-mail: dr.aali7pk@gmail.com, a.ali@iu.edu.sa
Faculty of Computer and Information Systems, Islamic University in Madinah, Almadinah Almunwarah, Saudi Arabia

Keywords: temperature control system (TCS), labview, user datagram, remote lab monitoring, web server

Received: November 25, 2021

LabView is used to design the system to monitor and control lab experiments remotely over the Internet. Web based embedded system is designed and presented in this paper along with required hardware to perform experiments remotely in allocated time slot. A unique system is designed to monitor and control room temperature remotely. To assess the performance of the designed system and the hardware, an experiment is designed to control room temperature remotely. The Temperature Control System (TCS) publish over the internet by using web server along with embedded live video of the equipment to control and change various required variables. LabView and the equipment from National Instruments is used to perform the experiment and the designed system can be controlled anywhere from the planet by using secure login detail only which exist in the database of server. As the experiments take long time to be completed and during this time the performer need to be presented in the Lab. By implementing this type of system, it will save the time of the performer and it can be utilized in other activities. Also, by designing and implementing this approach students from other Universities can also utilized the hardware and collaboration between institutions will be enhanced. The designed system presented in this article shows promising results which are presented in coming sections of this paper. The performed experiments are recorded and banked in the repository of the institution for future reference along with results.

Povzetek: Članek skuša izboljšati sodelovanje med univerzami preko oddaljenega nadzora laboratorijskih eksperimentov.

1 Introduction

The importance of the remote lab experimentation has been observed during the COVID-19 pandemic when the academic institutes are closed or running distance learning in almost whole world. To prevent the spreading of COVID-19, the governments have to closed and impose lock down in all or affected areas. Due to this the mobility has been affected and academic institutes went online and distance learning to continue education. With the increasing number of things on the internet the concept of providing integrated services as a result of communication amongst heterogeneous networks is gaining momentum [3]. In last half century, advancement in electronics and communication technology enable the scientist to connect the machineries with computer system which can be operated over the internet remotely. With increasing the use of internet in the globe, many people work from home instead of going to office premises. It saves time as well as cost for both employer and employee. These technologies are also beneficial for students to perform experiments remotely instead of going into laboratories. In this way students do not have to wait for the equipment time slot, but they need to register for an online experiment slot according to their convenience. By registering, system allocate time slot to each user for conducting experiment.

In this context National Instruments has great contribution by developing Lab View which is graphical programming and Lab View has built in servers and can be configured those servers according to the requirements. After configuring the servers one can perform experiments from a remote computer [1]. In most of the organization the National Instruments equipment is also used in laboratories with Lab View software, but in various organizations this software is not used in this context. With the traditional laboratories this is not possible for all students to perform the experiment at the same time because experiment equipment limitation. By using the IoT based resources it is possible that all the students perform experiment sitting remotely from home or from anywhere in the world. In this way distant learning students can also perform their experiments from their residences.

The major aim of this research work is to make it able to perform and monitor experiments remotely and collaboration between the organizations to share the resources by using Internet of Things (IoT).

For monitoring purposes live video is also published in the panel by using web cameras to facilitate the experiment performer to monitor the hardware set up and other things in the laboratory while performing the experiments.

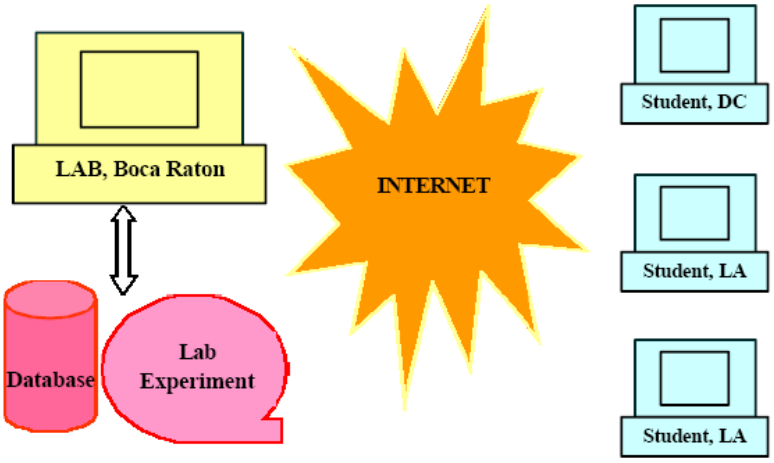


Figure 1: Framework of remote lab.

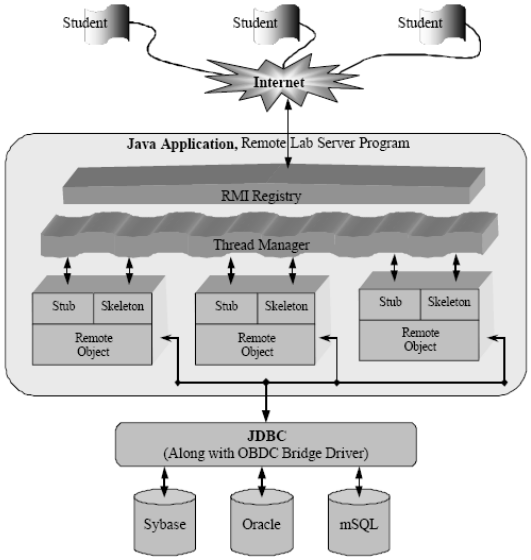


Figure 2: Java based client/ server communication model.

2 Related work

In last half century, the technology especially Internet related machines gained lot of attention from researcher who developed new models and architecture to build smart services, home and cities. Now a days traditional laboratories are not enough for all students to perform the experiment at the same time due to increased number of students in higher education sector especially in engineering and computer sciences. Due to lack of funds most of the laboratories don't have enough hardware for the labs to perform experiments. As a result, limited number of students can perform experiment in the daytime and tutors need to make groups to perform experiments. During the experiment time the laboratory supervisor needs to be present in the laboratory to supervise the students. However, there is another option that we can create virtual laboratories and at the same time many students can access that virtual experimentation but the drawback of this type of labs are that the experiment results of virtual lab are entirely different as compared with the real-world experiments results. Making the real setup of hardware in the laboratory and making it online for students to perform the experiment from anywhere in the globe. This reduces the costs and the time of the students. By making the Remote Labs it gives more opportunities to perform experiments more than once and gain the correct results of the experiments. The major benefit is for the distance learning students who learn from home, and they cannot come into the university every day or even once a week and making remote labs in universities, the student can get more practical knowledge than relying on theory [4].

By implementing the Remote Labs, it gives more opportunities to perform experiments more than once and gain the correct results of the experiments. The major benefit is for the distance learning students who learn from home, and they cannot come into the university every day or even once a week and making remote labs in universities, the student can get more practical knowledge than relying on theory [4,13]. Technologies are rapidly

growing and penetrate in every field of life in last few decades. In the education field it provides the new teaching techniques like Remote Monitoring or Tele Labs and etc. (e.g. a microscope is set up in NASA America and that is accessible from Russia and one benefit of this setup is that when Russian Scientists access that microscope at that time it is night time in USA). In this system both countries can utilize that microscope and other hardware. And there is no time clash between two countries because when it is daytime in Russia it is America's nighttime. Due to this time difference, scientists of both countries utilize that microscope with no time problem [6,14]. The operation modal of remote lab is show in figure 1. The figure shows that students present at different geographical location can access remote lab via internet to perform desired experiments. The hardware is setup for an experiment in remote lab. Multiple students can access the experiment at same time but only one student can control the experiment at same time. Every student can perform experiment in his allocated time slot and the results of experiments are stored in database and can be retrieved and analyzed later. [2]

The software design for remote lab for client/server Java based software model. The figure 2 shows major component of java based remote lab. The complete software is composed of two parts. Client interface which is designed for the end user and other is remote lab that server part. Using JDBC the server objects are allowed to talk with the database. Any client who wants to talk with the database can do so by talking with the stub, which is in turn interacting with the skeleton. The following figure shows the actual experiment running in remote lab at Atlantic University. [2]

3 Remote lab architecture and hardware

3.1 Architecture

For connecting with the system through web serve by using username and password, the system verifies the credentials of the user then allow the user to start performing experiment. Then the client clicks on the remote panel for performing the experiment through web browser. Web browser sends a request to the experiment server and remote panel is downloaded for remote client and the whole system as shown in the following figure 3. The user can change the input variables and other required values to complete experiment. The data is automatically

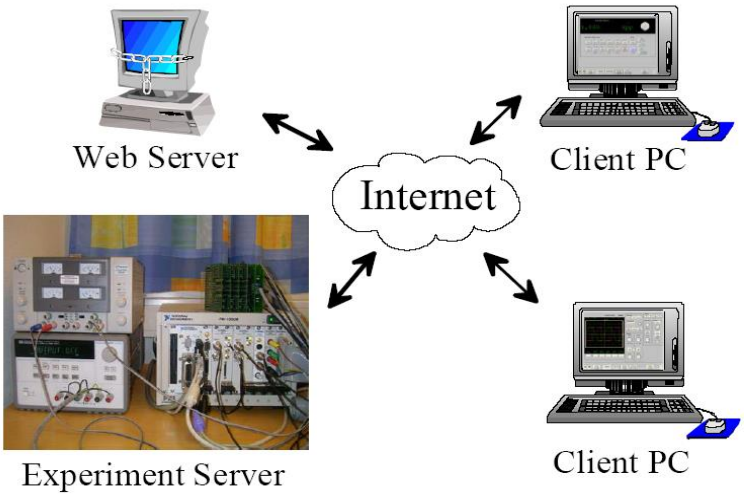


Figure 3: Client/server communication with experiment server.

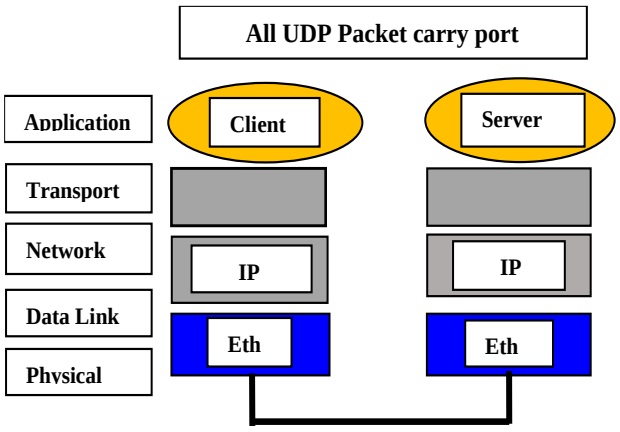


Figure 4: Client/server communication model.

Figure 5: Data acquisition card (DAQ).

saved on sever for future user and reference along with username and id.

3.2 User datagram protocol

User datagram protocol is the basically used as communication protocol to establish low latency and packet loss toleration between client/server communication over internet through web server. UDP transfer data very fast as it starts transferring data before agreement between sender and receiver. User datagram protocol is used for the for the purpose of transferring the input data to the hardware connected with experimental setup and UDP is connectionless protocol. It contains eight-bit header information, and it does not grantee to the destination for reliable data. The working of UDP is as follow.

3.3 Hardware

To demonstrate the proposed idea of remote monitoring of lab experiment, a sample experiment is designed by using following hardware components to access it remotely. To setup the experimentation following hardware is needed.

Transistor: The BDW93C is a silicon epitaxial-Base NPN power transistor in monolithic Darlington configuration mounted and is shown in the figure below.

Resistor: Light Bulb: 5-volt light bulb is used for increasing or decreasing the intensity of light and by varying the light of the bulb through DAQ input voltage, it varies the resistance of the Thermistor and by varying the resistance of the Thermistor temperature varies.

Power Supply: Power supply is needed to provide voltage to drive the circuit.

DAQ NI-USB 6008: DAQ is used for attaching all the hardware with the computer. The DAQ which is used in this project is shown below: [1]

3.4 Hardware Design

For controlling and monitoring room temperature remotely, the system is designed by using various hardware components and block diagram is shown in the figure 6.

A 5-volt supply is connected with designed system to switch on/off and reduce the intensity of heat in the environment to increase or decrease room temperature. 5-volt power is supplied through the DAQ and it can be varied between 1 and 5 volt to vary the brightness of the bulb.

3.5 Thermistor theory

Mathematical calculation of the VI is given below:

V = Voltage from supply, Vin = Voltage cross Thermistor, RA = Resistor Resistance, RB = Thermistor Resistance

$r = R0 * e^{-B/T0}$

Where R0 resistance at T0 (room temperature 289.15K) [25], T = Temperature, B =Parameter at room Temperature,

By using the following equations

$RB = (Vin * RA)/ (V-Vin)$

Here RB is Thermistor resistance and it now equal to R

$T = B/\ln(R/r)$

Temperature is in Kelvin, to calculate resistance across Thermistor

$RB = (Vin * RA)/ (V-Vin)$
$= (2.4 * 10000)/ (5-2.4)$
$=9231$

By using the second equation, Here RB is a Thermistor resistance, and it is now equal to R

$T = B/\ln(R/r)$
$= 3450/\ln (9231/0.0943296)$
$= 3450/\ln (97859)$
$= 3450/11.4913$
$= 300.228$ Kelvin

4 Remote lab experiment setup

To verify the proposed idea an experiment is designed which consist of hardware components and software. The figure 8 below shows the block diagram of the system designed by using LabView programming tool. The purpose of selecting the LabView to design the system is basically in hardware component data acquisition card (DAQ) is used from national instrument which is

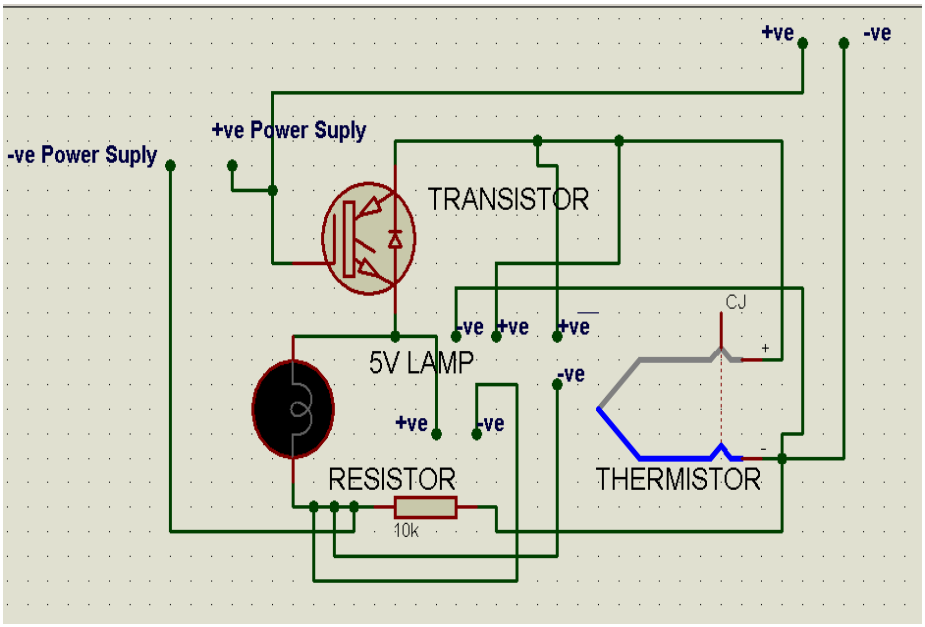


Figure 6: Block diagram of experimentation hardware.

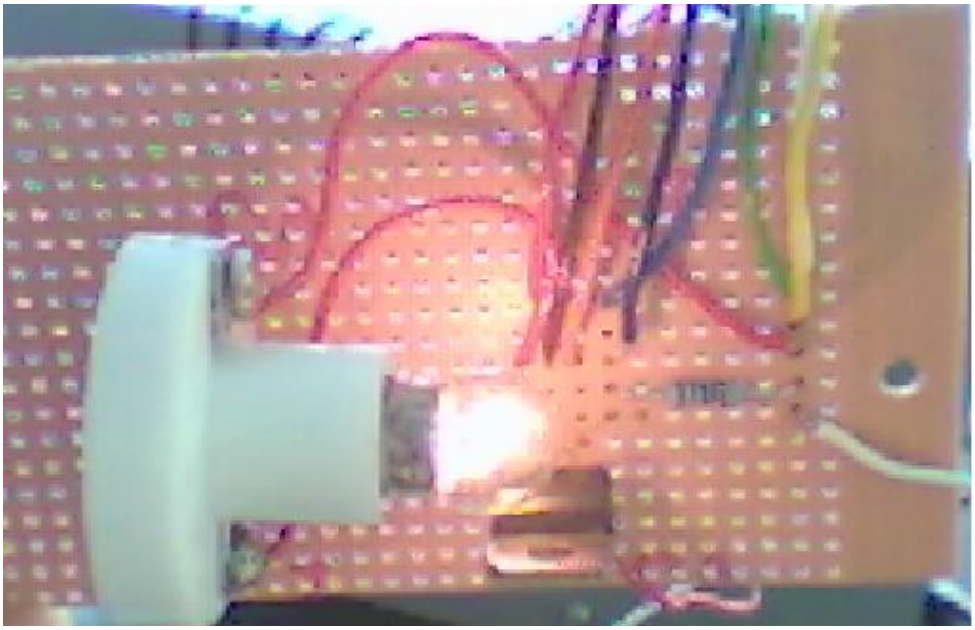


Figure 7: Experiment setup for temperature control.

compatible with LabView as both are developed by same vendor. The performer of the experiment can change the voltage which is going to bulb through resistor and due to change of voltage the temperature changes in the surrounding environment.

The DAQ Assistant is used to read the voltage across thermistor, supply voltage and voltage cross resistor.

The formula node is used to calculate temperature due to change of voltage across the Thermistor and supply voltage. For purpose of measurement, we need to configure DAQ with the number of samples, sample rate and timeout or delay between reading the samples from the circuit. If more readings are provided to the system, then it became slow, due to this reason less than 100 samples and timeout is 0.005 second is used and it gives best results. If the number of samples are increased more than 100 and the timeout more than 0.005 second, then the system response becomes slow.

Second DAQ Assistant is used for voltage variation, varying the voltage to the circuit the brightness of the bulb is changed. By changing voltage to the bulb, the brightness of bulb changes, due to change in brightness the temperature of the surrounding is also changed which is read through the Thermistor. The voltage timeout can be changed and by changing the timeout, the speed of the system varies, and this is measured in seconds. By increasing the timeout, the update becomes slow and decreasing the timeout update becomes fast. It is tested by using different values and by analyzing the results.

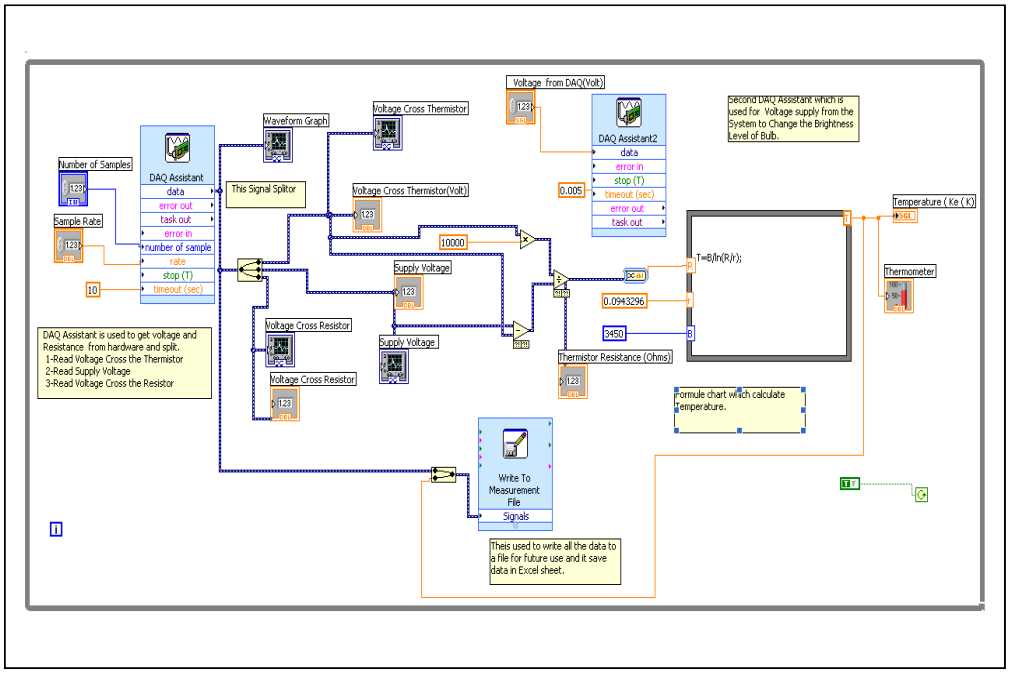


Figure 8: Experiment Block diagram designed in Labview.

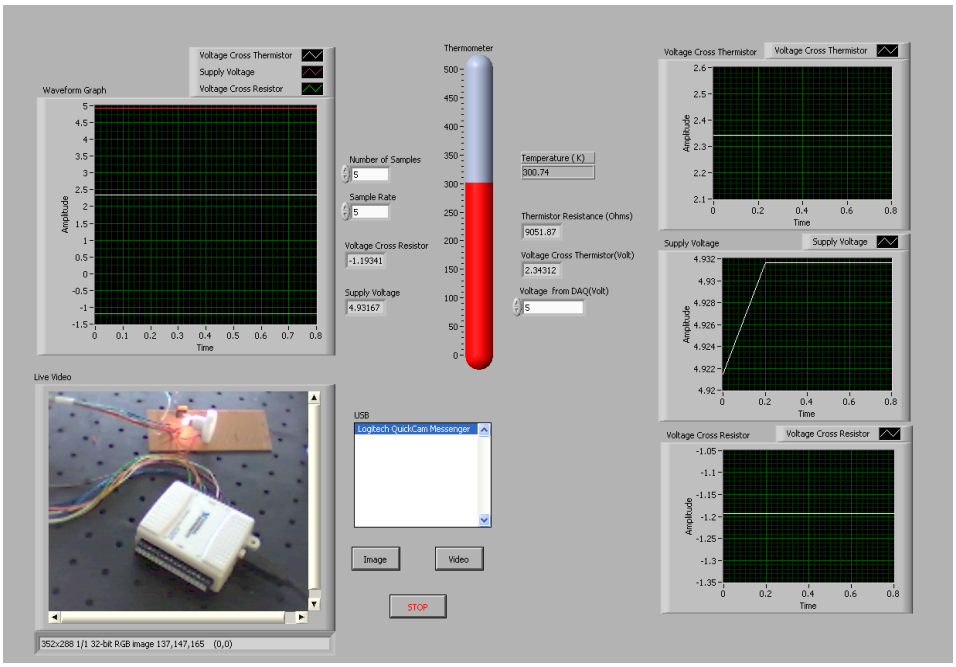


Figure 9: LabView based Front Panel for experiment.

A series of experiments performed it is found that 0.005 second is an ideal value for update to be sent to change the voltage. The experiment data is stored on the device by the using function “Write to Measurement File” provided by the LabView and it generate excel file for future retrieval. In the excel file, it records Voltage across the Thermistor, Supply Voltage and the Voltage across the resistor, and Temperature in Kelvin’s. Live video of experiment setup is also embedded in the panel to provide live view of the experimental environment. The control for publishing video live is given to the experimental performer to stop or start video on the panel which is embedded in webpage. The panel includes the voltage supply and change of voltage to the DAQ. Due to change of voltage the temperature change in the environment and it reflect the change in embedded thermometer.

5 Results and discussion

After designing the required software for temperature measurement remotely and client can vary the supplied voltage to the bulb. Due to change in voltage the brightness of light bulb is also change, by change in brightness the environment temperature change which is measured with respect to voltage change. The following figure shows variation in temperature with respect DAQ

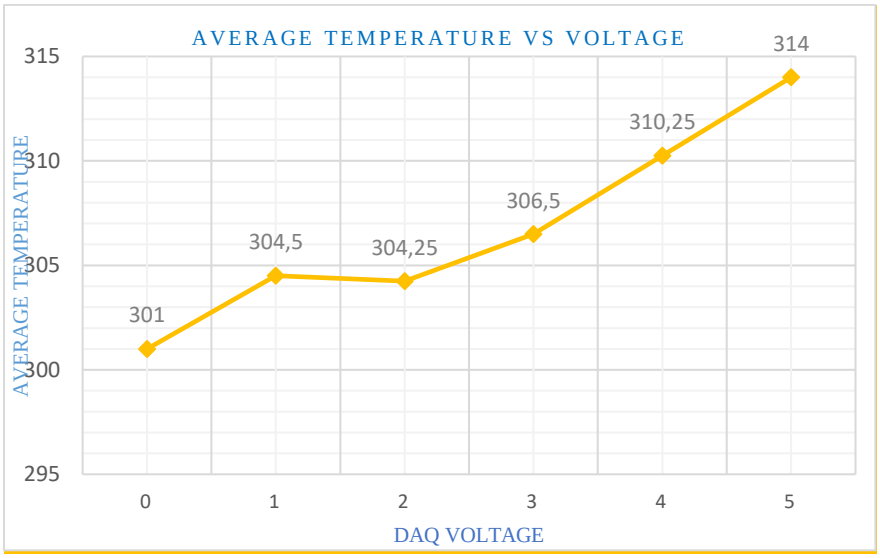


Figure 10: Average temperature vs voltage variation.

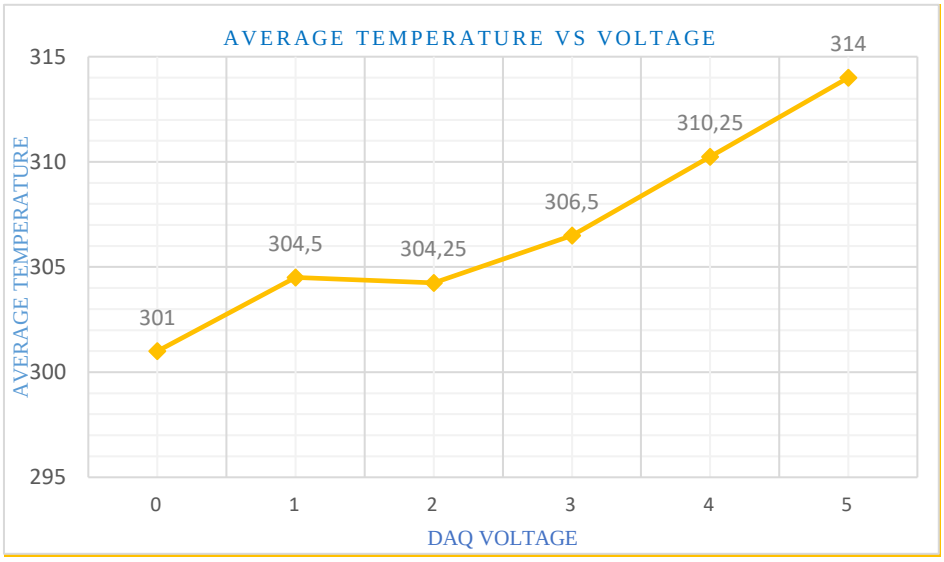


Figure 11: DAQ voltage vs average thermistor resistance.

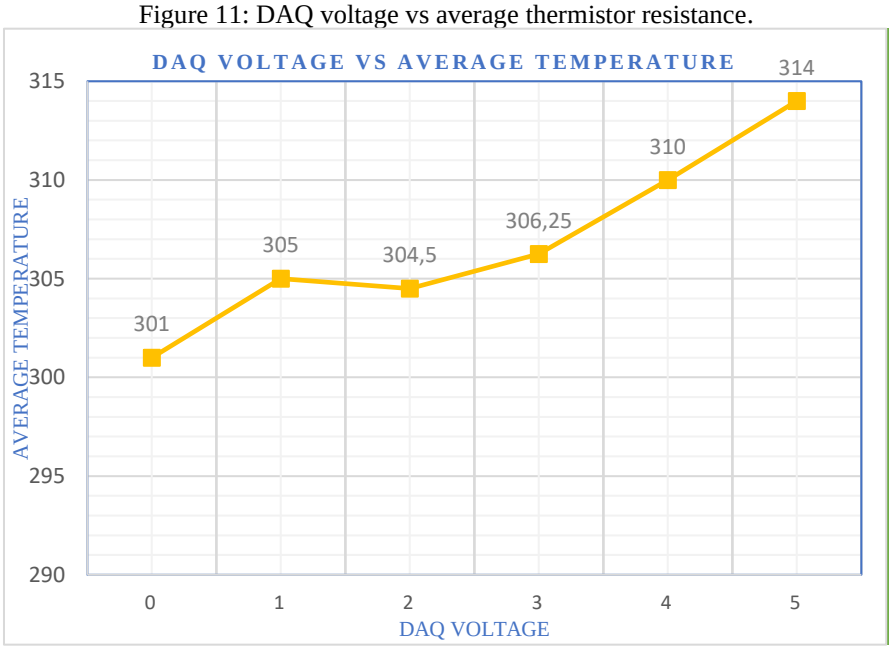


Figure 12: DAQ voltage vs average temperature.

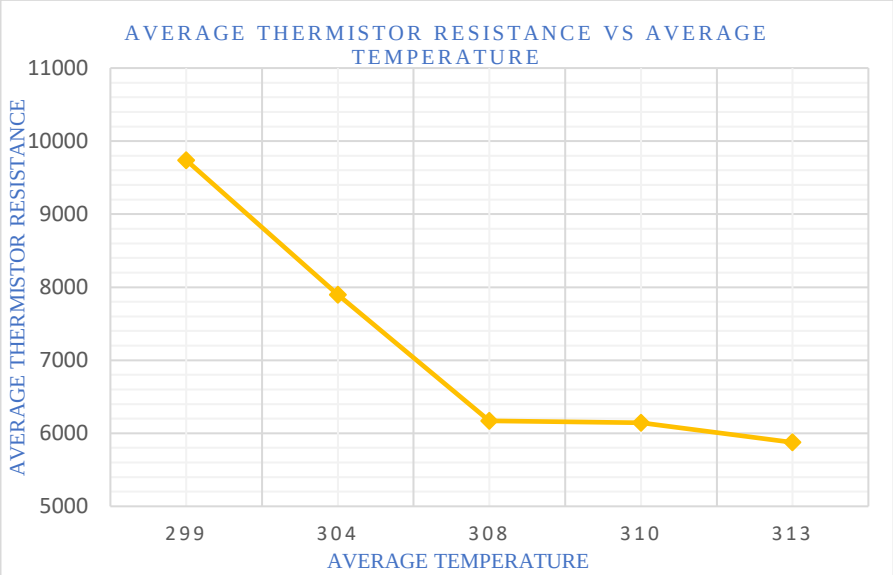


Figure 13: Average thermistor resistance vs average temperature.

voltage. By increasing the DAQ voltage the temperature increases because of the brightness of the light bulb. The presented graph below shows that by increasing the temperature the resistance of the Thermistor decreases and vice versa.

This graph shows the relationship between Thermistor resistance and DAQ voltage. DAQ voltage is taken along the x-axis and Thermistor resistance is along the y-axis. As the DAQ voltage increases the Thermistor resistance decreases and by decreasing DAQ voltage Thermistor resistance increase.

This figure 12 explains the relation between DAQ voltage and temperature, as the voltage increases the surrounding temperature increases. In this graph DAQ voltage is taken along the x-axis and temperature is along the y-axis. By analyzing the graph, one can concludes that by increasing the DAQ voltage to the circuit temperature increases and by decreasing the DAQ voltage to circuit temperature decreases.

Thermistor Resistance and Temperature relationship is shown in this graph. According to graph resistance decreases when temperature increases, and it increases when temperature decreases.

6 Conclusion

With the advancement of technologies and IoT devices in 21st century, it is very important that academia should enhance the collaboration between to utilize the resources of each other to benefit the learners. Due to COVID-19, it became more important for educational and technical institution that implement online laboratory systems in case to support online learning and experimentation. This project put the effort in regards of this area of research and development utilize online resources for remote experimentation along with enhancing collaboration between institutions. In this paper a remote-controlled experiment is designed and presented by using LabView integrated with web services for remote handling of lab experiments. The results presented in this paper is very much satisfactory and proposed model can be use to enhance collaboration between academia. By using proposed model, the universities, which are lack of the latest experimental bed can be benefitted from other universities labs. By using the presented collaboration model, the lab experiments can be done any time even out of hours from home or anywhere from the globe.

Acknowledgment

The researchers wish to extend their sincere gratitude to the Deanship of Scientific Research at the Islamic University of Madinah for the support provided to the Post-Publishing Program 1.

References

[1] Caragliu, A.; Del Bo, C.; Nijkamp, P. Smart cities in Europe. J. Urban Technol. 2011,18, 65–82. https://doi.org/10.1080/10630732.2011.601117

[2] Ishida, T.; Isbister, K. Digital Cities: Technologies, Experiences, and Future Perspectives; Number 1765; Springer Science & Business Media: Berlin, Germany, 2000.

[3] Alshmrany, Sami & Ali, Arshad "Health Monitoring and Management System by Using Wireless Sensor Network and Internet of Things (IoT)", VOL.19 No.12, December 2019, 179-184

[4] A. Bouroumi and R. Fajr, "Collaborative and Cooperative E-learning in Higher Education in Morocco: A Case Study," Int. J. Emerg. Technol. Learn. IJET, vol. 9, no. 1, p. 66, Feb.2014. https://doi.org/10.3991/ijet.v9i1.3065

[5] Wang Ping, Wang Zheng. Design and Implementation of Open Computer Lab Monitoring and Management System [J]. Computer and Modernization, 2007,11: 125-128.

[6] Jing Li, Yong Xu. Remote Monitoring Systems Based on Embedded Database [C]. Third International Conference on Genetic and Evolutionary Computing.

2009 :381-384.
https://doi.org/10.1109/wgec.2009.54

[7] M. A. Bochicchio and A. Longo, "Hands-On Remote Labs: Collaborative Web Laboratories as a Case Study for IT Engineering Classes," IEEE Trans. Learn. Technol., vol. 2, no.4, pp. 320–330, Oct. 2009. https://doi.org/10.1109/tlt.2009.30 https://doi.org/10.1109/tlt.2009.30

[8] Asraf, H.M. & Dalila, Nur & Zakiah, M.Y. & Faiz, Z.A. & Tahir, Nooritawati. (2018). Computer assisted e-laboratory using LabVIEW and internet-of-things platform as teaching aids in the industrial instrumentation course. International Journal of Online Engineering. 14. 26-42. https://doi.org/10.3991/ijoe.v14i12.8992

[9] Donose, Costel & Schreiner, Cristina-Mihaela & Macovei, Stefan. (2018). Using the Virtual Online Laboratories for the pre-University. 0609-0613. https://doi.org/10.1109/icepe.2018.8559617

[10] Chew A.S 1997, Development of a Feedback System for Position Control, Western Australia, Perth

[11] Davies T., Chan D. 2002 Network Cameras versus Streaming Video, Vacation work project, The University of Western Australia, Perth

[12] Herrera O.A., Alves G.R., Fuller D., Aldunate R.G. (2006) Remote Lab Experiments: Opening Possibilities for Distance Learning in Engineering Fields. In: Kumar D., Turner J. (eds) Education for the 21st Century — Impact of ICT and Digital Resources. IFIP WCC TC3 2006. IFIP International Federation for Information Processing, vol 210. Springer, Boston,https://doi.org/10.1007/978-0-387-34731-8_38

[13] Gennady Veselov, Alexey Tselykh, Ashutosh Sharma, Ruihang Huang, Applications of Artificial Intelligence in Evolution of Smart Cities and Societies, Informatica 45 (2021) https://doi.org/10.31449/inf.v45i5.3600

[14] Jianwei Liu, Research on Campus Network Equipment Environment Monitoring Based on Internet of Things, Informatica, 45 (2021) 303–307,https://doi.org/10.31449/ inf.v45i2.3550

A Study of Stressed Facial Recognition Based on Histogram Information

Barlian Henryranu Prasetio*, Edita Rosana Widasari and Fitra Abdurrachman Bachtiar
E-mail: barlian@ub.ac.id, editarosanaw@ub.ac.id, fitra.bactiar@ub.ac.id
Faculty of Computer Science – University of Brawijaya, Veteran No. 8 Malang, Indonesia

Keywords: emotion recognition, face, machine learning, Support Vector Machine, Gabor filter, HOG feature, histogram

Received: December 11, 2021

Stress represents our subconscious emotions. The majority of the unconscious content is unacceptable or unpleasant such as pain, anxiety, or conflict. Most individuals do not realize that they are experiencing stress. Prolonged stressful experiences are likely to lead to health problems and affect one's facial appearance, specifically wrinkles shown in the face. This paper discussed the introduction of facial stress with histogram information. There are three stages in recognizing the stress pattern on the face of the registered image, feature extraction and classification. The registered image process takes three important parts of the face, i.e. the eyes, nose, and mouth. The feature extraction process was performed using the histogram method, i.e. Gabor filter and HOG feature. Each extracted feature was used as the model input to determine whether or not an individual is suffering from stress. Two classification methods were applied to learn stress patterns from the extracted feature. The classification process was performed using SVM with six kernel functions and a Tree algorithm with three numbers of split. Each model is trained using ten cross-fold validation strategies. The test results showed that the Gabor filter and HOG feature accuracy were 55% and 65%, respectively. The effectiveness of the proposed method is evaluated by comparing it with the existing methods in term of accuracy.

Povzetek: Predstavljena je študija ugotavljanja stresa iz obraznih mimik.

1 Introduction

Stress is a non-specific response of the body in every claim [1]. It defines an individual's suppressed psychological state due to limitations and barriers arising in their effort to seize an opportunity [2]. Stress arises due to failure in achieving demands due to an imbalance in demands (physical and psychological) and ability [3]. According to [4] in his book, stress denotes one's physical or psychological response to a change in his environment perceived as disturbing and threatening [4].

Stress is our subconscious emotion. The subconscious mind is the shelter of feelings, thoughts, drives, and memories that are in the subconscious of our consciousness. Much of the unconscious content is unacceptable or unpleasant, such as pain, anxiety, or conflict. The subconscious will be influencing our behavior and experience, even though we are not aware of this fundamental influence [5].

Stress may affect one's physical condition. A prolonged stressful experience may result in heart and circulatory problems. As reported by [6], it can increase heart rate. In 2017, one-three of people reported feeling stressed, meaning that approximately 322 million people worldwide have an anxiety disorder [7]. If this condition continues, one's may suffer from increased blood pressure, affecting their facial appearance. Wrinkles may appear in an individual undergoing stressful experiences [8]. One's facial expression may indirectly show his/her stressed condition. However, it should be noted that stress is known to affect only several face areas such as the eye, nose, and lips [9].

Several methods are known to be helpful for recognizing stress conditions. The medical test is known to be the most accurate method in identifying stress level. However, it is very difficult to identify under-stress person. Most things have used the intrusive method, that is, through facial recognition. Over the past few decades, research on face recognition has been instrumental in computer vision technology [10].

Computer vision is a combination of image processing and pattern recognition. Image processing is a process of image transformation that aims to get good image quality [11]. Pattern recognition is the process of identifying an image object to extract information from the image. In neuroscience, researchers have been mostly concerned with models of the perception and classification of expressions [12], [13]. However, facial expression is a complex element that is difficult to understand. Even in a static pose, a face may present information related to emotion and mood [14]. It investigated how human observers use information from different face areas to successfully recognize the emotion expressed. More than a half of face part presents their conditions. Not only focusing on the perception of facial emotions and human behaviors [15], but the researchers also focused on the methods developed [16].

* Corresponding author

The current study focused on comparing two texture feature extractions. The feature extraction process is the process of encoding images in visual form. The encoding stage used the HOG descriptor (Histogram of Oriented Gradients) and Gabor filter. These two feature extractions were selected since the stress experience is identifiable through the wrinkles in the face area. Furthermore, a classification algorithm was applied to evaluate the effectiveness of the selected feature extractions. Previous studies showed that SVM and Tree-Based algorithm have better recognition rates than other machine learning methods [17]. Considering those findings, this study applied the SVM and Tree algorithm. The vector array of the extracted image features is modeled using SVM (Support Vector Machine) and Tree Algorithm.

2 Face and stress

Stress is likely to cause sleep difficulty [18]. It may slow down the circulatory system, expanding blood vessels below the eyes and thus creating dark circles around the eyes [19].

Although aging is a natural, common skin experience, stress may accelerate this process. Stress can affect our brain, causing wrinkles to emerge earlier due to anxiety, depression, exhaustion, or lack of rest. These lines or wrinkles may emerge around the forehead, mouth, and eyes [20].

Stress indirectly affects the production of healthy collagen in the skin. Thus, the skin will look dull and create a tired-looking face. Stress also causes the production of melanin, a pigment that gives color to the skin, to decrease [21]. One's signs of stress can be easily detected from his/her face. Figure 1 above displays the lines or wrinkles around the mouth and eyes.

3 Methods

The stressed face recognition system has two main parts: the feature extraction process and the classification [22]. Figure 2 shows the stress recognition system block diagram.

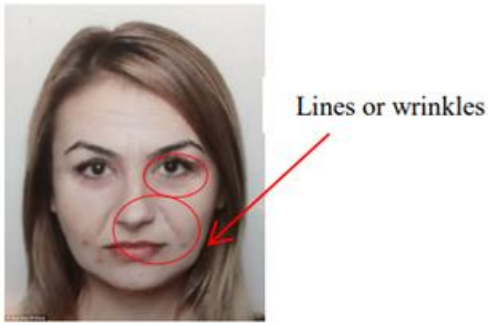


Figure 1: The lines or wrinkles around the mouth and eyes.

As shown in Figure 2, Specific face areas were segmented first before the feature extraction process. This study focused on three areas. The segmented images were extracted to obtain the numerical value for further process.

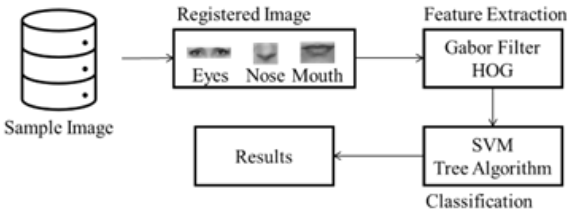


Figure 2: The stressed recognition system block diagram.

Lastly, the extracted features were fed to classification algorithm for facial pattern learning.

3.1 Image dataset

The training phase is the process of providing the images face that is known to the class (stress or neutral). The collection of facial images is called the image dataset. This study employed JAFFE consisting of 213 images of different facial expressions from ten different Japanese female subjects categorized as stress and neutral expression.

3.2 Registered image

The registered image phase has 3 steps to get the needed facial parts to be processed further. The first step is detecting the face area. The next step was to greyscale the image. Finally, the selected area of the image was cropped. In particular, the three areas to be cropped are the eye, nose, and mouth [23]. These areas are chosen as the features since wrinkles may appear in these three regions. The image registration steps are displayed in Figure 3.

Face detection and area cropping is detected and performed using Viola-Jones object detection framework in MATLAB Toolbox.

3.3 Feature extraction

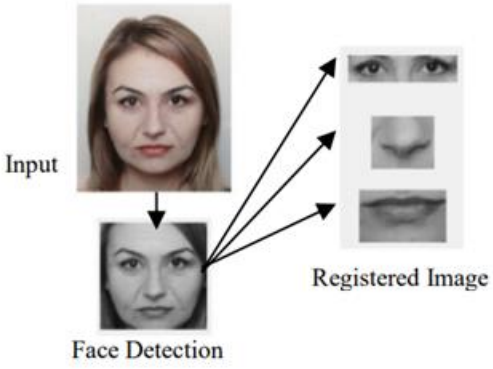


Figure 3: The pre-processing steps of registered image.

3.3.1 Gabor feature

Gabor Filter is a linear filter used to detect edges and frequency decomposition [24]. There are 2 main processes in Gabor feature extraction: the process of making Gabor array and the extraction process, i.e., taking vector feature from the Gabor array [25].

Gabor filter consists of several steps. The first step is the initialization variable. Four variables are initialized, including wavelength (u), number of orientations (v), and two dimensions of Gabor filter bank (m, n) [26], [27]. Each Gabor filter variable used in this study was identified to obtain the maximum extraction results. These four parameters were used to calculate the Gabor filter bank. The visual illustration of the Gabor matrix is displayed in Figure 4.

Fig. 4 shows that the 5x5 filter matrix. Each filter array represents a different wavelength and orientation [28].

3.3.2 Histogram of Gradients (HOG)

Histogram of Oriented Gradients (HOG) is a rotationally invariant descriptor [29]. HOG feature consists of several stages, including gradient calculation, spatial weighting and orientation of cells, normalization of spatial block overlapping and HOG detection windows.

The gradient calculation process is to apply the vertical and horizontal Sobel method with the kernel filter on the greyscale image (Rekha and Kurian, 2014). The magnitude and gradient values can be calculated after obtaining the values of x, y derivatives Equation 1.

$$|G|=\sqrt{(I\_x\text{^}2+I\_y\text{^}2)} \quad (1)$$

Where, I_x and I_y are input image after convolution operation x and y.

The spatial weighting and orientation of the cells process divide the image into small spaces called cells. Each pixel in the cell is grouped in bins based on the orientation value obtained from the gradient calculation [30]. Each cell creates a histogram with 9 channel histograms distributed into an angular orientation of 0°-180°.

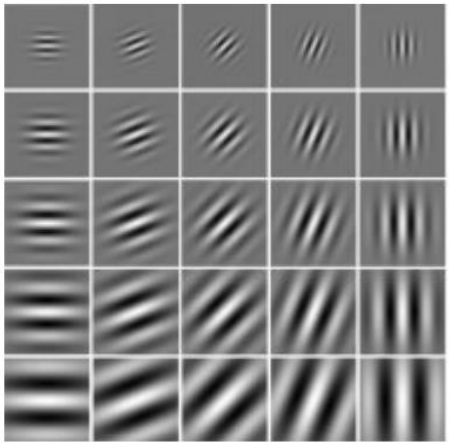


Figure 4: The visual illustration of Gabor matrix (u=5, v=5).

Since the gradient calculation process obtained different values, it is necessary to group each cell into larger groups called blocks. After grouping into a block, any overlapping block should be normalized. All histograms for each block were combined to produce a feature vector. This process is called a HOG detection window. The illustration of the HOG feature stages is presented in Figure 5. All of the HOG parameters in this study were explored to get the most extraction results.

3.4 Classification

SVM is the traditional classification method used in this study because it can find the hyperplane optimally [31]. SVM maps each input and output data to represent the similarity vector [32]. Vector mapping can be either linear or non-linear [33]. For non-linear functions, a kernel is required in the mapping process. Three types of kernels are known, namely Linear (Equation 2), Polynomial (Equation 3) and Gaussian (Equation 4).

$$k(x,y)=x\text{^}T\ y \quad (2)$$

$$k(x,y)=〖(1+x\text{^}T\ y)〗\text{^}d \quad (3)$$

$$k(x,y)=e\text{^}(-(〖||x-y||〗\text{^}2/〖2\sigma〗\text{^}2)) \quad (4)$$

Where x and y represent classes, d represent polynomial degree; if d = 2 then quadratic kernel function; if d = 3 then cubic kernel function; $1/2\sigma = \gamma$; if $\gamma = 0.43$ then fine Gaussian; if $\gamma = 1.7$ then medium Gaussian if $\gamma = 6.9$ then coarse Gaussian.

The second classification method was Tree algorithm. Tree algorithm has influenced machine learning as classification and regression. A tree has presented a node and split of the branch. The splitting number σ (T) of tree T has an effect of successive [34]. In this study, feature extraction data from Gabor filter and HOG feature were classified into three models of Tree algorithm, namely, Fine Tree, Medium Tree, and Coarse Tree. The number of fine tree's split is 100, for medium tree is 20, and for the coarse tree is 4.

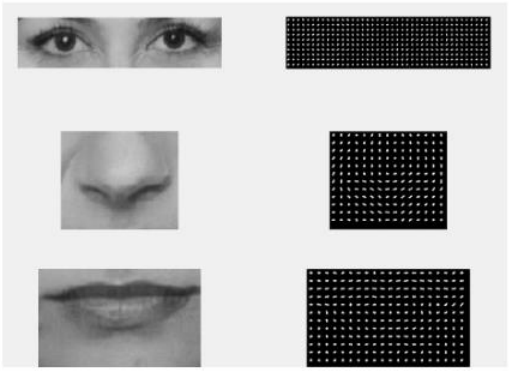


Figure 7: The gradient histogram of eyes, nose, and mouth.

3.5 Validation method

In these experiments, the stress classification is validated using a k-fold cross-validation method. This is a technique to validate the accuracy of a model built on a particular dataset. The development of a model usually aims to predict and classify new data [35]. The data used in the model development process is called training data. The data used to validate the model is called the test data. In

this study, we evaluated the effectiveness of the proposed system by applying some "k" value.

4 Result and discussions

4.1 Gabor filter

Initially, we created a Gabor filter matrix (m=7, n=5). This matrix was filled by the Gabor array with the number of wavelengths of 5 and the number of orientations of 7. The wavelengths were 3, 6, 13, 28, and 58. The orientations were 0, 30, 45, 90, 120, 135, and 150. Every face part of a registered image (eyes, nose, and mouth) is convoluted its magnitude based on Gabor array. So, each face part has 7x5 arrays of images. The Gabor magnitude image array for the eye can be seen in Figure 6.

The next step was converting those magnitudes matrices to feature vectors. The feature vector consists of Local Energy and Mean Amplitude [36]. Local energy is the mean of the sum of squares of the magnitude matrix. Mean Amplitude is the mean of the absolute values of the magnetic matrix.

Furthermore, both local energy and mean amplitude are input of SVM and Tree Algorithm.

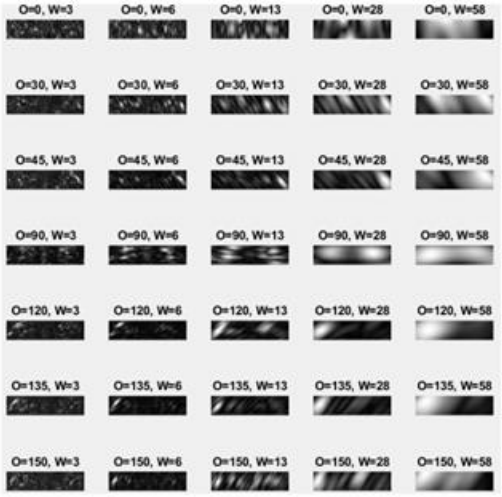


Figure 6: The Gabor magnitude image array for the eye.

4.2 HOG feature

The first step of the HOG feature is calculating a HOG descriptor. We need the horizontal and vertical gradients. The next step is to divide the image into a 16x16 cell. Then the gradient histogram is counting on each cell. Each gradient histogram consists of 2 matrix values, i.e. magnitude and direction. A gradient histogram of three face areas (eyes, nose, and mouth) can be seen in Figure 7.

As shown in Figure 7, the calculation of the feature vector values was performed after the gradient histogram matrix value was normalized into dimensional vectors. The dimensional vectors in each face area will take the average value of each cell and will be as the input of SVM and Tree Algorithm.

4.3 Support Vector Machine

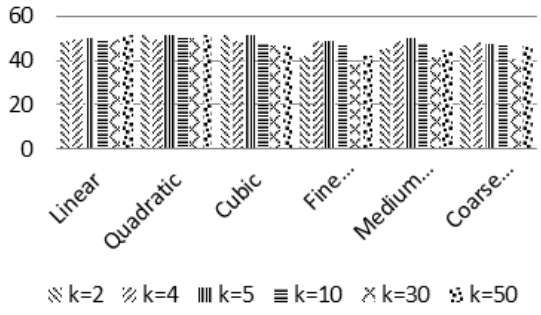
The result data from feature extraction was classified using SVM algorithm. The classification was performed with 6 SVM kernel functions, i.e. linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian. Each classification was validated using a cross-validation method with some "k" values. The accuracy of SVM algorithm can be seen in Figure 8.

Figure 8 (a) displays the SVM accuracy for Gabor filter feature extraction. The highest accuracy was around 50%. Fig. 8 (b) shows the SVM accuracy for HOG feature extraction. The highest accuracy was around 60%. Based on the classification results, the combination of SVM and HOG feature shows higher recognition result compared to SVM and Gabor Filter by showing accuracy of 62.8%.

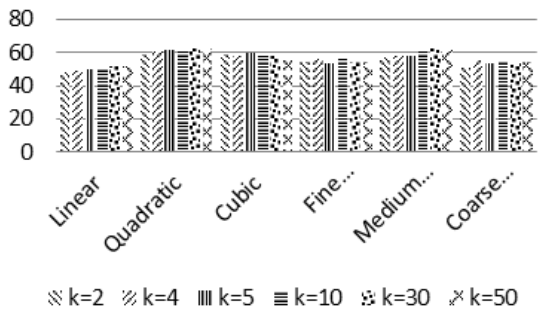
4.4 Tree algorithm

Each feature extraction (Gabor Filter and HOG Feature) was classified using the Tree algorithm. The accuracy value was cross-validated. The accuracy of the Tree algorithm can be seen in Figure 9.

Figure 9 shows that the average accuracy using Tree's algorithm is above 65%. Figure 9 (a) shows that the best accuracy of the Tree's algorithm classification for the Gabor feature is the 30-folds Fine Tree. Figure 9 (b) shows that the most accurate of the tree algorithm classification for the HOG feature is the 50-folds Medium Tree. The combination of Medium Tree and HOG Feature was found to be more accurate than other combinations with a 70.4% accuracy.



(a) Gabor Filter



(b) HOG Feature

Figure 8: The accuracy of SVM algorithm.

4.5 Comparison performance

The effectiveness of the proposed method is evaluated in term of accuracy and compares it to baseline methods, as showed in Table 1.

The evaluation result shows that Bayesian Network (BN) able to incorporate the past information about a event parameters. However, BN sometimes produce a posterior distribution that is strongly influenced by priors, it might difficult to assure the validity of the selected priors. On the other hand, the DNN based technique (ResNet and SOTA) outperforms the BN. DNN better to represent complicated patterns of emotions. However in a simple case, such as binary classification of stress, our work more accurate than DNN due to a low set of categorical values in training data. DNN and Tree has a same mechanism in finding non-linear solutions by interacting between independent variables. Thus, decision trees are better when the scenario demands an explanation over the decision. It is because its deterministically one, whereas DNN take a probabilistic view towards the piece-by-piece model fitting.

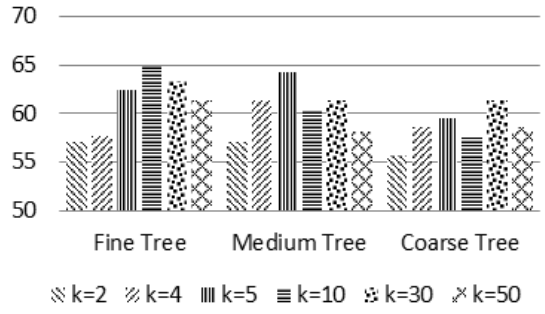
4.6 Summary

The summary of classification performance can be seen in Figure 10.

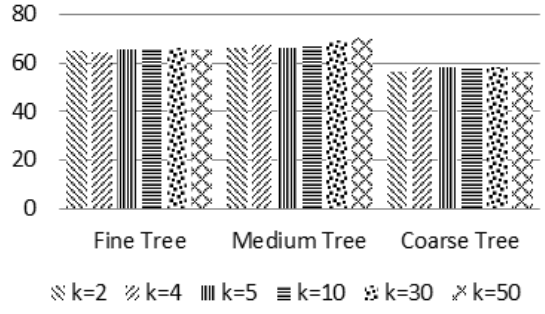
As shown in Figure 10, the accuracy of SVM-Gabor filter feature was 51.9%, while that of SVM-HOG feature was 62.8%. The accuracy of Tree algorithm and Gabor filter feature was 64.8% and that of Tree algorithm and HOG feature was 70.4%. The results indicated that the HOG feature is more efficient for classifying stress extracted from facial elements. HOG feature extracted the orientation of the histogram which can extract nonlinear pattern in the image. The registered images used in this study were not linear. Thus, nonlinear feature extraction would be suitable in this situation.

We also evaluated the proposed method effectiveness by comparing it with the existing methods such as Bayesian Network (BN) and Self Organizing Tree Algorithm (SOTA). The evaluation result shows that our work outperforms the existing methods because decision trees explicitly fit parameters to direct the information flow.

Stress, anxiety, and depression were the highest mental health problems with 74% for stress, 28% for anxiety disorders, and 48% for depression [38]. Currently, mental health problems are increasing and worrying, especially because of the COVID-19 pandemic, affecting individuals and society in an increasingly diverse nature. The pandemic has also given rise to a wide-ranging mental health crisis. This is forcing decision makers to turn to technology to open up research opportunities into a coherent framework that serves as an initial effort to develop interdisciplinary research between technology and mental health [39][40].



(a) Gabor Filter



(b) HOG Feature

Figure 9: The accuracy of Tree algorithm.

5 Conclusion

In this paper, we have recognized the facial signs caused by stress. Three face areas (i.e., eyes, nose, and mouth) were selected as stress features. We applied a Gabor filter and HOG feature extraction methods. The classification

Table 1: The comparison performance evaluation.

Reference	Dataset	Method and Classifier	Accuracy (%)
L. Surace [10]	GAF (positive/negative event)	Bayesian Network	64.68
S. D. Viet [15]	FERC-2013 (7 expressions)	Multi-task ResNet	69.33
B. H. Prasetio [37]	SUSAS (5 classes of stress)	SOTA	68.72
Our work	Jaffe (stress/neutral)	Tree	70.4

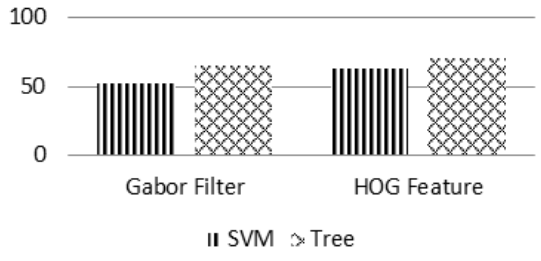


Figure 10: The summary of classification performance.

process was performed by SVM and Tree algorithm. The effectiveness of the proposed system was evaluated using k-fold cross-validation method with some "k" value. The experimental result showed that SVM and Tree algorithm with Gabor filter feature exhibit an average accuracy of 55% while the combination of SVM and Tree Algorithm with HOG exhibited a 67% accuracy. The current study achieved satisfactory results by recognizing stress from facial images by involving three face areas, namely eyes, nose, and mouth. Future studies might consider adding more face areas as additional features. It is also recommended to explore other algorithm methods as the model could affect the recognition accuracy. In a simple classification, decision tree better than DNN based method due to its explicitly fit parameters to direct the information flow and deterministically one.

References

[1] H. Selye, Stress in health and disease. Butterworth-Heinemann, 2013. https://doi.org/10.1016/C2013-0-06263-9

[2] S. P. Robin, "Organizational Behavior," New York Prentice Hall, 2001.

[3] R. S. Weinberg and D. Gould, Foundations of sport and exercise psychology, 7E. Human Kinetics, 2019.

[4] H. Yaribeygi, Y. Panahi, H. Sahraei, T. P. Johnston, and A. Sahebkar, "The impact of stress on body function: A review," EXCLI J., vol. 16, p. 1057, 2017. https://doi.org/ 10.17179/excli2017-480

[5] B. J. Carducci, The psychology of personality: Viewpoints, research, and applications. John Wiley \& Sons, 2009.

[6] B. . Tovian, S.; Thorn, B.; Coons, H.; Susan, L.; Matthew, B.; Richard, S.; Daniel, "Stress effects on the body," Am. Psychol. Assoc.

[7] W. H. Organization and others, "Depression and other common mental disorders: global health estimates," 2017.

[8] J. H. Dunn and J. Koo, "Psychological Stress and skin aging: a review of possible mechanisms and potential therapies," Dermatol. Online J., vol. 19, no. 6, 2013. https://doi.org/10.5070/D3196018561

[9] C. Daudelin-Peltier, H. Forget, C. Blais, A. Deschênes, and D. Fiset, "The effect of acute social stress on the recognition of facial expression of emotions," Sci. Rep., vol. 7, no. 1, pp. 1–13, 2017. https://doi.org/10.1038/s41598-017-01053-3

[10] L. Surace, M. Patacchiola, E. Battini Sönmez, W. Spataro, and A. Cangelosi, "Emotion recognition in the wild using deep neural networks and Bayesian classifiers," in Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 593–597. https://doi.org/10.1145/3136755.3143015

[11] A. Adriyendi, "A Rapid Review of Image Captioning," J. Inf. Technol. Comput. Sci., vol. 6, no. 2, pp. 158–169, 2021. https://doi.org/10.25126/jitecs.202162316

[12] A. Martinez and S. Du, "A model of the perception of facial expressions of emotion by humans: research overview and perspectives.," J. Mach. Learn. Res., vol. 13, no. 5, 2012.

[13] I. Azam and S. A. Khan, "Feature extraction trends for intelligent facial expression recognition: A survey," Informatica, vol. 42, no. 4, 2018. https://doi.org/10.31449/inf.v42i4.2037

[14] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. How individual face parts contribute to successful emotion recognition," PLoS One, vol. 12, no. 5, p. e0177239, 2017. https://doi.org/10.1371/journal.pone.0177239

[15] S. D. Viet and C. L. T. Bao, "Effective Deep Multi-source Multi-task Learning Frameworks for Smile Detection, Emotion Recognition and Gender Classification," Informatica, vol. 42, no. 3, 2018. https://doi.org/10.31449/inf.v42i3.2301

[16] M. K. Benkaddour, "CNN based features extraction for age estimation and gender classification," Informatica, vol. 45, no. 5, 2021. https://doi.org/10.31449/inf.v45i5.3262

[17] D. S. Trigueros, L. Meng, and M. Hartnett, "Face recognition: From traditional to deep learning methods," arXiv Prepr. arXiv1811.00116, 2018. https://doi.org/10.48550/arXiv.1811.00116

[18] D. A. Kalmbach, J. R. Anderson, and C. L. Drake, "The impact of stress on sleep: pathogenic sleep reactivity as a vulnerability to insomnia and circadian disorders," J. Sleep Res., vol. 27, no. 6, p. e12710, 2018. https://doi.org/10.1111/jsr.12710

[19] R. Sarkar, R. Ranjan, S. Garg, V. K. Garg, S. Sonthalia, and S. Bansal, "Periorbital hyperpigmentation: a comprehensive review," J. Clin. Aesthet. Dermatol., vol. 9, no. 1, p. 49, 2016.

[20] C. Parrado, S. Mercado-Saenz, A. Perez-Davo, Y. Gilaberte, S. Gonzalez, and A. Juarranz, "Environmental stressors on skin aging. Mechanistic insights," Front. Pharmacol., vol. 10, p. 759, 2019. https://doi.org/10.3389/fphar.2019.00759

[21] A. Sitek, E. Zadzinska, and I. Rosset, "Effects of psychological stress on skin and hair pigmentation in Polish adolescents," 2012. https://doi.org/10.2478/v10044-012-0001-7

[22] M. Usman, S. Latif, and J. Qadir, "Using deep autoencoders for facial expression recognition," in 2017 13th International Conference on Emerging Technologies (ICET), 2017, pp. 1–6. https://doi.org/10.1109/ICET.2017.8281753

[23] K. Tsuchida and M. Kobayashi, “Oxidative stress in human facial skin observed by ultraweak photon emission imaging and its correlation with biophysical properties of skin,” Sci. Rep., vol. 10, no. 1, pp. 1–7, 2020. https://doi.org/10.1038/s41598-020-66723-1

[24] L. El Shafey, R. Wallace, and S. Marcel, “Face verification using gabor filtering and adapted gaussian mixture models,” in 2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), 2012, pp. 1–6.

[25] K. Yan, Y. Chen, and D. Zhang, “Gabor surface feature for face recognition,” in The first asian conference on pattern recognition, 2011, pp. 288–292. https://doi.org/10.1109/ACPR.2011.6166553

[26] T. Barbu, “Gabor filter-based face recognition technique,” Proc. Rom. Acad., vol. 11, no. 3, pp. 277–283, 2010.

[27] A.-A. Bhuiyan and C. H. Liu, “On face recognition using gabor filters,” World Acad. Sci. Eng. Technol., vol. 28, 2007. https://doi.org/10.1109/ICCITECHN.2008.4803030

[28] M. Sharif, A. Khalid, M. Raza, and S. Mohsin, “Face Recognition using Gabor Filters.,” J. Appl. Comput. Sci. \& Math., no. 11, 2011.

[29] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, “On effectiveness of histogram of oriented gradient features for visible to near infrared face matching,” in 2014 22nd International Conference on Pattern Recognition, 2014, pp. 1788–1793. https://doi.org/10.1109/ICPR.2014.314

[30] S. Paisitkriangkrai, C. Shen, and J. Zhang, “Face detection with effective feature extraction,” in Asian Conference on Computer Vision, 2010, pp. 460–470. https://doi.org/10.1007/978-3-642-19318-7_36

[31] M. G. L. Putra, W. Ariyanti, and I. Cholissodin, “Selection and Recommendation Scholarships Using AHP-SVM-TOPSIS,” JITeCS (Journal Inf. Technol. Comput. Sci., vol. 1, no. 1, pp. 1–13, 2016. https://doi.org/10.25126/jitecs.2016111

[32] L. Muflikhah and D. J. Haryanto, “High performance of polynomial kernel at SVM Algorithm for sentiment analysis,” JITeCS (Journal Inf. Technol. Comput. Sci., vol. 3, no. 2, pp. 194–201, 2018. https://doi.org/10.25126/jitecs.20183260

[33] I. Oktanisa, W. F. Mahmudy, and G. Maski, “Inflation Rate Prediction in Indonesia using Optimized Support Vector Regression Model,” JITeCS (Journal Inf. Technol. Comput. Sci., vol. 5, no. 1, pp. 104–114, 2020. https://doi.org/10.25126/jitecs.202051173

[34] P. Hansen, A. Hertz, and N. Quinodoz, “Splitting trees,” Discrete Math., vol. 165, pp. 403–419, 1997. https://doi.org/10.1016/S0012-365X(96)00187-2

[35] Y. Bengio, Y. Grandvalet, and others, No unbiased estimator of the variance of K-fold cross-validation. Citeseer, 2003.

[36] N. Joshi, “Combinational neural network using Gabor filters for the classification of handwritten digits,” arXiv Prepr. arXiv1709.05867, 2017. https://doi.org/10.48550/arXiv.1709.05867

[37] B.H. Prasetio, H. Tamura, K. Tanno, K. "Semi-Supervised Deep Time-Delay Embedded Clustering for Stress Speech Analysis". Electronics. vol. 2019, no. 8, 1263. https://doi.org/10.3390/electronics8111263

[38] T. Kolenik and M. Gams, "Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality?," in IEEE Technology and Society Magazine, vol. 40, no. 1, pp. 80-86, March 2021. https://doi.org/10.1109/MTS.2021.3056288.

[39] Kolenik, T. (2022). Methods in Digital Mental Health: Smartphone-Based Assessment and Intervention for Stress, Anxiety, and Depression. In: Comito, C., Forestiero, A., Zumpano, E. (eds) Integrating Artificial Intelligence and IoT for Advanced Health Informatics. Internet of Things. Springer, Cham. https://doi.org/10.1007/978-3-030-91181-2_7

[40] X. Peng. "Research on Emotion Recognition Based on Deep Learning for Mental Health," Informatica, vol. 45, no. 1, 2021. https://doi.org/10.31449/inf.v45i1.3424

A Prestudy of Machine Learning in Industrial Quality Control Pipelines

Jože Ravničan[1], Anže Marinko[2], Gjorgji Noveski[2], Stefan Kalabakov[2], Marko Jovanovič[3], Samo Gazvoda[4] and Matjaž Gams[1]
E-mail: joze.ravnican@unior.com, anze.marinko@ijs.si, gjorgji.noveski@ijs.si, stefan.kalabakov@ijs.si, marko.jovanovic@smm.si, samo.gazvoda@gorenje.com, matjaz.gams@ijs.si
[1]UNIOR Kovaška industrija d.d., Zreče, Slovenia
[2]Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia
[3]SMM proizvodni sistemi d.o.o., Maribor, Slovenija
[4]Cooking Appliances Division Gorenje Group, Velenje, Slovenia

Keywords: machine learning, manufacturing, quality control, home appliances, car industry

Received: January 24, 2022

Today's fast paced industrial production requires automation at multiple steps during its process. Involving humans during the quality control inspection provides high degree of confidence that the end products are with the best quality. Workers involved in the control process may have an impact on production capacity by lowering the throughput, depending on the complexity of the control process at the time the control is carried out, during the process which is a time-critical operation, or after the process is completed. Companies are striving to fully automate their quality control stages of production and it comes naturally to focus on using various machine learning methods to help build a quality control pipeline which will offer high throughput and high degree of quality. In this paper we give an overview of applying several machine learning approaches in order to achieve an autonomous quality control pipeline. The applications for these approaches were used to help improve the quality control pipeline of two of the biggest manufacturing companies in Slovenia. One of the most challenging part of the study was that the tests had to be performed only on a small number of defective products, as is in reality. The motivation was to test several methods to find the most promising one for later actual application.

Povzetek: Z nekaj prototipi je bila narejena analiza možnosti uporabe umetne inteligence v nekaj velikih slovenskih podjetjih.

1 Introduction

In the world of machinery, products are expected to be available in high quantities as well with high quality. To make this possible, many companies around the world introduce an efficient production process which is streamlined as much as possible. To achieve this, many go to such lengths by having a zero defects policy. This policy suggests that product defects, while unavoidable, should be kept to the lowest percentage possible. The reason for this is to increase profits by lowering the cost of failure. Defects can occur at any stage during the production pipeline. It is important to catch these product defects as early as possible so they don't have a chance to propagate to later stages causing even more damage. Some approaches, such as [4], reduce defect propagation by shifting quality control at earlier stages of the production process.

Quality control is a process that is present throughout all aspects of production, ensuring standards and quality criteria are met before products leave the assembly line. In order for this process to be as efficient as possible, it needs to be integrated into the manufacturing process. Depending on the products manufactured, quality control might need a separate physical area devoted to it in order to perform multiple analyses, although this segregation of tasks in a production environment is often cumbersome and error prone. Traditionally, the quality control process requires specialized labor who are skilled in correctly identifying mistakes or inconsistencies in the developed product. The downside of using such a specialized labor force is that humans are prone to error [9]. Additionally, repetitive tasks are not suited for one person to perform them for a long period of time with consistent quality. One solution is to use computer aid in order to achieve a higher quality of the overall processes.

Implementing a machine learning quality control pipeline is a challenging task. Difficulties arise from the specifics of the industry, requiring expert knowledge to be used to develop a system which is capable of satisfying the quality needs while also being fast. In [20] a comprehensive overview regarding the scope of those challenges is given and one could see the plethora of categories that machine learning has been applied to.

Looking at machine learning as an enabler to industrial processes, [15] provide with an overview of the effort put into machine learning in manufacturing. Most noticeably are the uses of neural networks and support vector machines (SVMs) since they are generically applicable and

provide competitive results with a low investment time. Regarding industrial maintenance, [18] provides an overview on current and future trends which are shaping companies towards Industry 4.0. Prevalent in these efforts is the inter-connectivity of different sensors and systems, namely internet-of-things (IoT).

In this paper we provide a technical overview of various applications of machine learning into a manufacturing environment. The solutions presented here are provided for industrial products each having their own specific purpose. Due to the different types of products, we will test which methods are more suited than others. We also show how making use of transfer learning in the image recognition domain can provide encouraging results on domains that the model was not trained on.

The products that were used for our experiments included **metallic forks**, **metallic surfaces**, **oven fans** and **broaches**. These are included in the following sections and the data is structured as follows. Section 2 presents similar papers which focus of quality control in manufacturing environments. In Section 3 we present our solutions for quality control on metallic forks from car parts. Regarding home appliances, applicable solutions are presented in Section 4 and Section 5 where we examine oven parts. Section 6 examines broaches in metalworking to determine the amount of wear. Discussion is provided in section 7. Lastly a conclusion is presented in Section 8.

2 Related work

Many applications of machine learning proved successful in a wide set of industries, ranging from automotive to pharmaceutical. Each industry specializes on its problem domain to achieve a great end product or a service.

Looking at the field of pharmaceuticals, high-throughput screening (HTS) is a method to quickly test a large number of chemicals and/or compounds. Using robotics, data processing, sensitive detectors and more, this allows a researcher to conduct millions of chemical, genetic and pharmacological tests. Because of the necessity to conduct a large amount of tests, machine learning can be used here to speed up the process. In [10], a large problem of the process represents data mining and data storage of the images acquired during testing. There exists a trade-off between quality and high throughput. More time is needed to get results with higher quality though that lowers the overall throughput. Often, image analysis is a bottleneck in this process, meaning good and fast machine learning methods for visual tasks are paramount.

An interesting implementation of HTS is presented in [2], where a whole quality control pipeline is created using open source software. Their goal is eliminating images which deviate from the standard quality baseline. Their tool of choice can implement multiple image processing modules all of which are configurable. The modules implement machine learning models which achieve good results.

In manufacturing environments, where products need to conform to high standards, [5] use a knowledge-based intelligent supervisory system to find events of rare quality. These rare quality events can be considered as finding defects in around one million opportunities. By using several methods, such as logistic regression and feature elimination, they have achieved great results, showcasing what kind of exceptional quality machine learning can bring.

In the automotive industry, timely detection of car body defects is important since a bad end product can cost tens of thousands of dollars. To ensure good car body dimensions, [12] use XGBoost [3], random forest, and other methods to detect dimensional defects. From their production line we can see that they carry out quality control in two stages while the stages that are in-between are human operated.

Integrating machine learning solutions into industrial processes needs careful selection of best performing models and parameters. The deployed models have to provide a prediction in less time than a unit needs to be manufactured. Because of this, not all production processes can utilize machine learning easily. In [14], an integrated quality inspection solution is presented which uses the power of machine learning and edge computing. Edge computing was used to provide faster model training and evaluation in order to find the one that gave the best results. They work on improving the quality inspection in electronics manufacturing which uses surface mount technology (SMT) assembly. Printed circuit boards (PCBs) are assembled and go through a solder paste inspection (SPI) phase, which is the phase that the predictive models are applied. They used several models such as Naïve Bayes, Decision Tree, Logistic Regression, SVM and Gradient Boosted Tree. Since in their use case the manufacturing plant was outputting a device every single second, the model needs to be trained and predict in less time than that. All of the models were capable of achieving this except Gradient Boosted Tree where the training time took 2 second. It was shown that this approach could improve the product throughput by lessening the need for units to go to additional quality checks.

Related to food, [1], explore the usefulness of using machine vision in food packaging. Specifically, they analyze the seals of thermoforming packaging using three convolutional neural networks (CNNs): ResNet [6], VGG [16], DenseNet [8]. These tasks are still done by human operators which presents a possibility to greatly automate and improve the production line. Line workers were observed to see which packages were accepted and which were rejected. With that knowledge, a system was built to classify packages with improper seals that could cause food to spoil. Five different datasets were created depending on the type of product that was produced using an inline image acquisition system. Their solution detects 99.93% of defects in production. It is worth mentioning that this was tested in both laboratory conditions and also in a real scenario.

Efforts are made into automating quality control processes in order to be closer to Industry 4.0. In [11], they mention several manufacturing paradigms with the main

focus being on Holonic Shop Floor Control Systems. To aid in bridging the flow of information between a management level and a manufacturing process level, several CNNs were trained in order to find flaws in cast products, namely impellers of submersible pumps. In order to obtain more samples for training, the authors used several image augmentation techniques such as: different axis rotations, zooming in or out and shear mapping. The best performing CNN had a total of 7 layers and achieved precision of 99.82%.

In order to get a better representation of the efforts which are put into building quality control system, a summary of related work, their description and results is shown in Table 1.

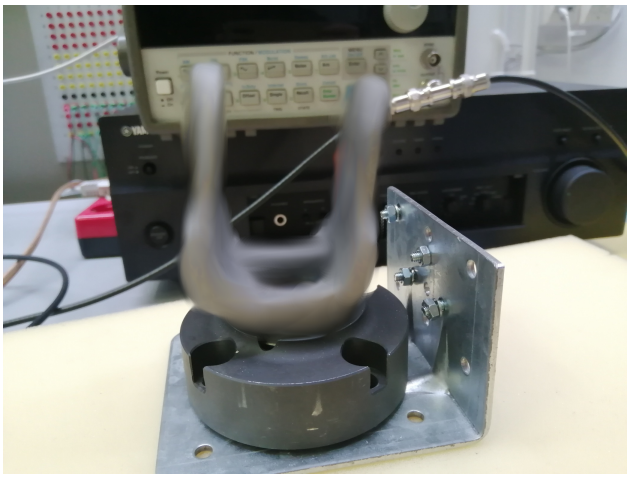


Figure 1: Vibration analysis.

3 Metallic forks

In this section, we describe our quality control solutions which were tested in different manufacturing environments. Our goal is to utilize machine learning accompanied by preprocessing techniques in order to classify if a given industrial product contains a defect or not. The solutions created must be easy to implement and have little interference during the manufacturing of products. To detect defects on metal forks we used three approaches: vibration analysis, 2D and 3D machine vision analysis.

3.1 Vibrations

Looking at utilizing vibrations, we measured the signal of displacement of the rack which holds and shakes the product and measured the signal of displacement for the product. To measure the displacement we used a laser distance meter with high accuracy. In the end we observed the relationship between the two displacement signals. Before going deeper with this technique, we did a basic validity test in which we glued a small piece of insulating tape on the fork. Results showed that the displacement signals had very high deviation from the baseline which gives hope that real defects will also be noticeable on the displacement signals. Then we tested the real products by comparing ones with and without an imperfection. Each output signal was 10 seconds long which was obtained from using a total of 24 products. Four different vibration settings were used which were:

1. Amplitude: 0,389 Vpp; frequency: 50Hz
2. Amplitude: 0,389 Vpp; frequency: 60Hz
3. Amplitude: 0,2026 Vpp; frequency: 60Hz
4. Amplitude: 0,2026 Vpp; frequency 50Hz

We split each 10 second measurement into 10 pieces of 1 second long data points. On each of these pieces, 22 features were expertly selected based on the characteristics from the time and frequency spectrum. On the dataset 5-fold cross validation was done. For the vibration analysis a convolutional neural network was also used, having 2 hidden layers and an input of 256 neurons. The input was the whole 10 second measurement of the signal. The results obtained from the CNN were: Accuracy 48%, Recall 42%, Precision 91%. There was no point going further with this approach, so other methods had to be tested.

3.2 2D machine vision analysis

The major problem with machine learning in our case was lack of faulty products. For the machine vision approach we had 9 defective products in total from which 46 images were captured. Each of the images was manually annotated with a region of interest which encompasses the area with an imperfection. This was done so we could have an annotated dataset we can use to train our deep neural network. This dataset was further divided into a training and testing dataset.

In order to perform image analysis we used a pre-trained convolutional neural network, Faster RCNN [13]. This networks consists of several convolutional layers and at the end several fully connected layers. The convolutional layers are good at extracting features from the image such as shapes and edges, whereas the fully connected layers learn to classify these features for the task at hand. The whole network consists of two parts, a segmentation part and a classification part, both of which were used to detect defects on the surfaces. The first part, namely segmentation, was used in order to extract specific regions of interest in the picture which in our case was regions with defects. These regions are then fed into the second part of the neural network whose tasks is to classify if a certain region contains an imperfection or not. The pre-trained weights were used from the Faster RCNN network and only the fully connected layers were trained on our training set, effectively freezing the convolutional layers so we can utilize the features learned in previous object detection tasks.

Paper	Description	Results
[5]	Using feature elimination and classification threshold search to find best models for classifying good or bad welds	Logistic regression model - maximum probability of correct decision (MPCD): 1
[12]	Testing several machine learning models to predict dimensional defects in cars	Both XGBoost and Random Forest - ROC AUC: 0.97
[14]	Training machine learning models to detect defective units during PCB manufacturing	Gradient boosted tree - precision: 93.1%, recall: 89.9%, accuracy: 92.6%
[1]	Proper seal detection in thermoforming food packaging	ResNet18 - accuracy: 95%
[11]	Convolutional neural networks in detecting faults in cast metal products	Custom CNN - accuracy: 99.82%

Table 1: Overview of related work.

During training we used 2900 epochs on the segmentation network and 200 epochs on the classification network. In Figure 2 we can see a schematic view of the metallic fork that was used in our experiments.

The test set used the segmentation network from the Faster RCNN which contained 21 images with an imperfection, 7 without, are presented in Table 2, wheres the results on the classification network on the testing dataset which had 18 images with an imperfection, 1 without are presented in the Table 3.

TP	FP	TN	FN	Accuracy	Recall	Precision
17	4	3	4	71.4%	81%	81%

Table 2: Results from the testing dataset on the segmentation network.

TP	FP	TN	FN	Accuracy	Recall	Precision
1	0	1	17	10%	5%	100%

Table 3: Results from the testing dataset on the classification network.

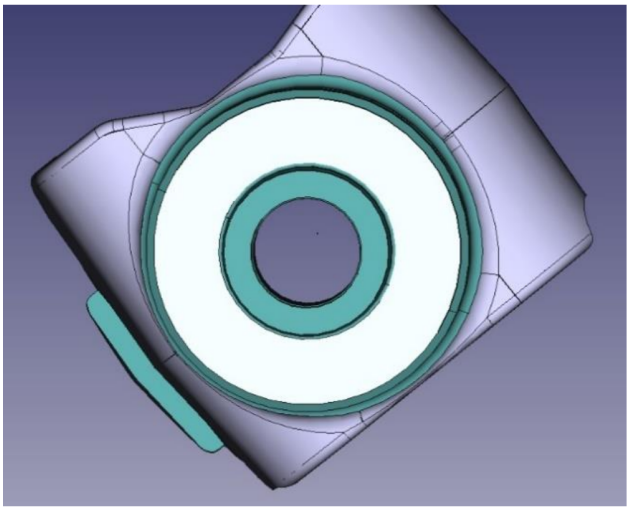


Figure 2: Schematic view of the metallic fork.

The small number of learning examples was insufficient for a hard real-life problem, as expected.

3.3 3D machine vision analysis

The last technique for detecting imperfections on metallic surfaces that was tested was through using a 3D hardware system. This is the most expensive option from all the others but is the one that was expected to deliver the most promising results. We scanned a total of 8 items, 4 of which contained a scratch on their surface and 4 without. The hardware system in question is 3D camera "Ranger 3" manufactured by SICK with laser light source of 660 nm (red laser). It is a line scanner which means it scans objects not at once, but line by line. During scanning, either the object or the scanner might be moved in order to get a complete scan. In our case, the scanner was mounted on a fixed point and only the object was moved along one axis. Figure 3 showcases the setup used to scan objects, in this example a saw tooth blade. The point clouds obtained from the system are 3D representations of the scanned object. Each point in the point cloud has a specific color value associated with it, which represent the reflectance value of that point. The scratches in the point cloud exhibited lower number of 3D points at their location, and additionally a change in reflectance values. The observed scratches varied by length and width, some of them even forming curved lines. Concerning width, a single scratch might also have varying width along its whole length. In reality, the width of the scratches varies between 0.15mm to 2mm.

In order to have a bigger dataset to work with, new samples were generated by adding artificial imperfections on samples or adding some noise. Random noise was introduced by moving each point 0.001 mm in any direction, ending up with a new point cloud. In order to generate a point cloud with a scratch, the following steps were taken:

1. Two points on the surface are chosen which will represent the beginning and the end of a scratch.
2. A random point on that line will represent an area where the scratch is the widest.
3. All points that are contained within a circle which its center is defined by the random chosen point are affected points.
4. Randomly sample affected points. The probability of a point being sampled and removed is higher when

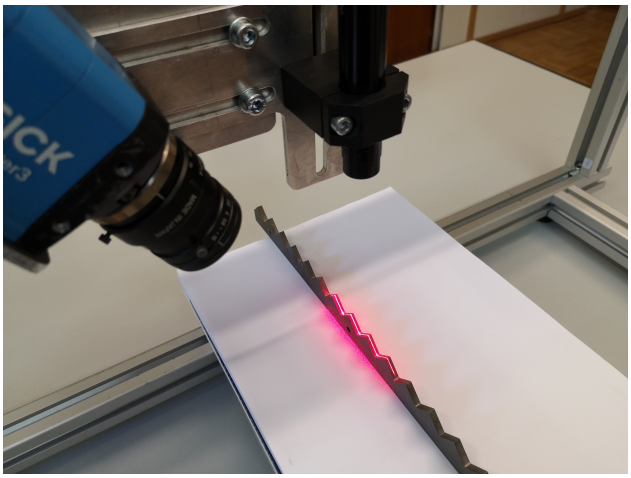


Figure 3: 3D vision system setup.

it is closer to the circle center and center line of the scratch. The radius of the circle is random and varies between 0.15mm to 2mm.

5. The points which were not removed in the previous step are sampled again using the same logic but this time instead of being removed their reflectance is lowered.

Lastly we reduce the problem dimensionality by removing the Z-axis. In order not to lose the Z-axis information it is encoded as brightness of the pixels. Next we select an area of interest on which we will do our analysis. After selecting an area of interest, a median 3x3 filter is applied to remove parabolic lines from the scan since those are a by-product of the whole scanning process. After blurring, a binary threshold is applied using a constant T=60. Deciding whether the region of interest includes a scratch, the image is compared to an error-free surface using a metric called "structural similarity index". The values returned by this metric are between 1 and 0.

Using this metric as the only feature, a random forest classifier is trained and the results are show in Table 4. The classifier achieves a perfect score.

True\predicted	No error	With error
No error	113	0
With error	1	106

Table 4: Confusion matrix of the random forest classifier.

In summary we tested three approaches to solve quality control of metallic car objects: by vibrations, by using a 2D hardware system and by a 3D hardware system. The first two approaches soon turned out unsuccessful. Only the last, hardware costing around 10.000€ enabled good results with machine learning.

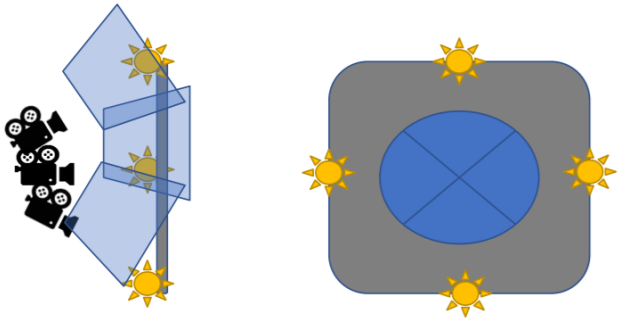


Figure 4: Camera and lighting arrangement.

4 Oven surfaces

This section is focused on detecting imperfections on oven faceplate's metallic surface and the interior surface of the oven. During the final stages of the production of a commercial oven, a key step is visual inspection on its surfaces. For the decision of image acquisition we opted to use a simple RGB camera that will have almost no interference on the production line. The amount of oven faceplates we obtained was again very small, in total 5 faceplates from which we took our images. Out of the 5 faceplates, 3 had scratches on the metallic surface and 2 did not have any scratches.

Before using machine learning models, we also tested some feature extraction methods, such as Speeded Up Robust Features (SURF) and Scale Invariant Feature Transform (SIFT). These were used in order to compare images with and without defects from 13 different key areas on the oven. These key areas are places where the most errors occur, such as panel buttons, door alignment, etc. These methods proved to be effective, achieving an almost perfect F1 score, or at least above 0.9 on all the 13 key areas.

4.1 Machine vision

Alongside detecting scratches on the front metallic surface of the oven, pictures were taken from inside the oven at varying angles and light conditions. In Figure 4 we can see the triple camera arrangement with 4 different light sources. Multiple light sources were used and these caused a certain amount of noise and glare in the final picture, remedying it by performing post-processing of the images. This resulted in improved stability in the classification algorithms. To increase the number of faults on the metallic surface, adhesive tape was used to mimic imperfections alongside with using a sharp metal object to create new scratches on the interior oven surface.

In order to be able to use deep learning models, segmentation and augmentation was used to increase the number of images. Before segmentation, we manually annotated the images that contained an imperfection using a binary mask, 1 indicating a pixel which is part of an imperfection, 0 otherwise. We used various methods to increase the number of images we could get from the faceplates, from

Architecture	Simple-CNN	VGG16	InceptionV3	ResNet101V2
(Macro) F1-score	9.8%	13.0%	58.75%	60.75%

Table 5: Average model F1 score after leave-one-out cross-validation.

covering individual scratches with adhesive tape, to coloring with a felt-tip pen and finally using image augmentation. To segment an image we used a "sliding window" approach where a window slides over and extracts smaller parts of the image. The segments which overlap with the binary mask from before are labeled as images that contain an imperfection. Figure 5 shows an example of how an oven faceplate is segmented, avoiding the display area.

After segmentation, class-invariant transformations are performed in order to increase the dataset size. Traditional image augmentation techniques such as flipping and rotating perform well on the results of a classifier as can also be seen in [19]. With all these methods combined, we managed to obtain more than 20,000 images. We have chosen 4 convolutional neural network models for this task, one simple CNN and the others which are VGG16 [16], InceptionV3 [17], ResNet101V2 [7]. In order to train all of these with transfer learning, we disregarded the last fully connected layers, substituting them with our own layers with randomly assigned weights. During training the weights of all except the final fully connected layers were updated. By making use of transfer learning we shorten the time needed to train a neural network and simultaneously use knowledge which was previously obtained by the pre-trained network on a wide domain of images. The performance of the models was evaluated using leave-one-out cross-validation (LOOCV),with 4 folds. For an evaluation metric F1 macro average score was used. The reason for using the F1 metric with averaging is because of our imbalanced dataset where the majority of samples belong to the "No imperfection" class.

The results of all the models are shown in Table 5. As we can see our simple CNN and VGG16 weren't able to capture the difference between images with or without imperfections. On the other hand, InceptionV3 and ResNet101V2 showed better and similar results to each other.

5 Oven fans

Also regarding commercial ovens, the next quality control solution focuses on oven fans. In order to inspect if a fan which is located inside an oven is in working condition, videos were recorded with a frame rate of 30 frames per second and from those videos images were obtained totalling 7200 images. Out of all those frames 4000 were of a working fan and 3200 were with one which is not working. The visual data was obtained through a closed oven door, since opening the door in a manufacturing environment would take too much time.

Preprocessing of the images consisted of three steps:

1. Object detection
2. Glare reduction
3. Thresholding

5.1 Extracting image features

In object detection the Hough Gradient Method [21] was used to extract certain shapes from images. We are only interested in the area where the fan is located which happens to be a circular opening. Before running the method to detect circles, a median filter is applied to blur the image, since using the original image results in many false positive detection of circles. The output of this method was a single circle although the radius of each circle from all the images varied. Because of this the mean value of the circles' radius was calculated and used.

For glare reduction, each image was decomposed into a color, saturation and brightness component (HSV). After this, particularly bright areas of the image were identified and their pixel values inpainted in respect to their surrounding pixel. We followed few rules while reducing glare. Because reflections coming from the light source are white, any pixel that contains glare will not have saturation, since white has no color or saturation. Then we filtered out the areas that have low saturation. Following that, the non-saturated areas were reduced by an erosion morphological operator and the brightness values of the saturated pixels were set to 0. Lastly, the final glare mask was obtained by choosing pixels that have high brightness (threshold 130). The pixels from the original image that overlap with the glare mask were then interpolated with an inpainting operation. The final image without glare is almost identical to the original image.

The last step is thresholding. In an image frame where a fan is not working, the fan stands out more, and in a frame where a fan is working the fan is slightly blurred. Using this information we decided to use binary thresholding and count the number of white pixels in an image. The idea is that if the thresholded image represents a non working fan, then the number of white pixels would be greater than in a thresholded image representing a working fan. For the binary thresholding we used a constant of T=90. Next, to calculate the white pixel count threshold by which we decide the class of the image, that is working/non-working we calculate the number of pixels in all binary-thresholded images of a non-working fans and take the 5^{th} percentile of that. In the end, there were some abrupt changes in classification during consecutive frames and to remedy that the class of each frame was taken as the majority of the last 20 frames.

The confusion matrix for the results is shown in Table 6.

	Non-working	Working
Non-working	3117	82
Working	280	3720

Table 6: Confusion matrix from the results.

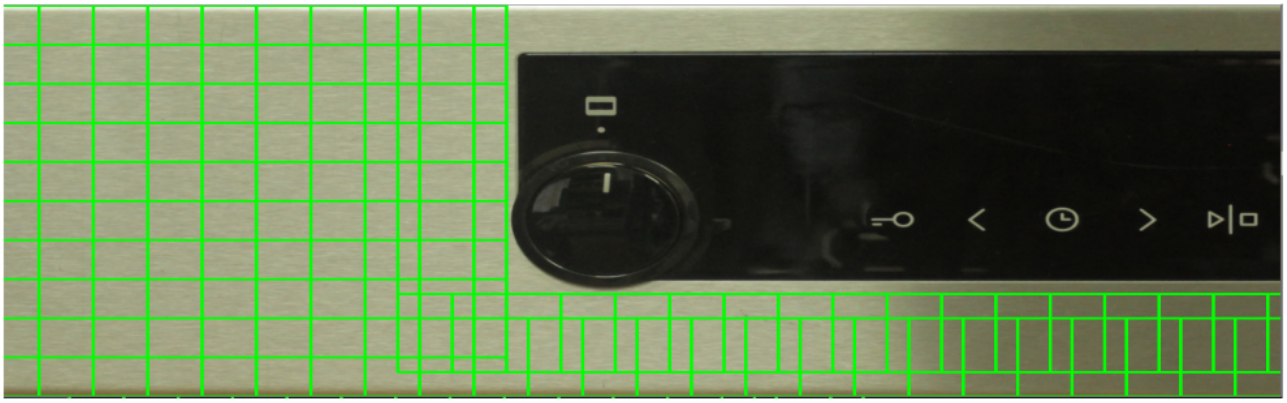


Figure 5: Example of image segmentation.

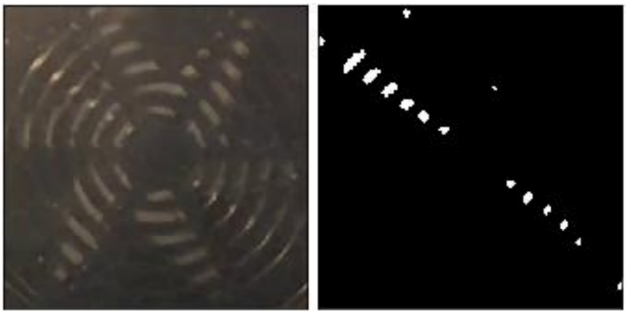


Figure 6: Oven fan and corresponding glare mask.

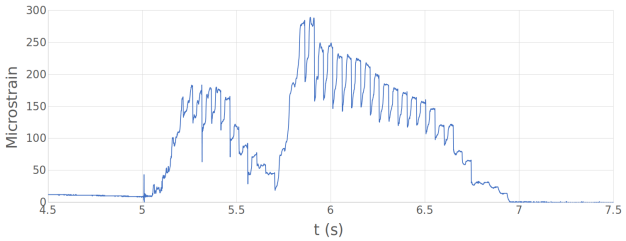


Figure 7: Microstrain parameter over time.

5.2 Audio analysis

A technique that is worth mentioning is the use of signal analysis in order to detect the workings of the oven fans. Using an ordinary microphone from a telephone, samples were gathered which then were analysed. The mobile phone was placed on top of the oven housing in order to record the audio signal. It was observed that the operation of the working oven fan was present in the frequency spectrum at the peak of 100Hz. These results were obtained in a controlled laboratory setting. Recording samples were also obtained in a production environment where more noise is present. This added noise introduced some disturbances in the final signal spectrogram which made the detection of a working oven fan more difficult, nevertheless, correct detection of the oven fan was still possible. The ovens are equipped with two fans, a cooling fan and a heating one. During the period when they are operating, they both exhibit a signal of 100Hz on the spectrogram. This makes their separation and identification very difficult since the signals overlap. With this technique, at any given time when the fans are working, it can be concluded that at least one is operational, but not both because of the previously mentioned problem.

6 Broaches

During metalworking it is necessary to form a metal into a certain shape. Broaching comes into play because it cuts and shapes metal using so called broaches. Broaches come in different shapes and sizes with multiple blades on them, all with varying size. In the broaching process, internal or external broaching can be used and in our approach we experiment with internal broaching. These broaches need to be replaced after prolonged use and our goal is to identify how many operational cycles can be done using the broach before replacement is needed. If the blades of the broach are too worn but not damaged, then they can also be sharpened instead of replacing the whole broach, which is cheaper. To find out when a broach reaches it's maximum operational period before irreversible damage, we measure the microstrain parameter by the cut time, which tells us the strain change in parts per million.

6.1 Signal features

The normal amount of operational cycles for a single broach measures around 1500. Using a broach beyond this number has a very high probability that the cuts done by it will not be accurate or that a tooth from it might be damaged and in the end, broken. External factors such as heat and material lubrication affect the rate at which broaches are getting worn out. In defect recognition the most important factors are the number and shape of peaks in the measured stretch signal. In Figure 7 we can see a signal

of the amount of stretch over time which was taken during a cut of the broach and we can see how it changes over time when different sized blades pass through the object. Automatic peak recognition was done by measuring when the signal rises from its standard deviation. Doing this we were able to isolate a time window in which the signal was generated when the broach was used for cutting.

Various features were extracted from the signal such as minimum and maximum signal value, frequencies, patterns etc. The features were filtered using a statistical significance test and after using this test alongside with the Benjamini–Yekutieli procedure, it turned out that there are only three most important features which are: area, number of peaks and maximum value of the signal. These features were used to train several machine learning models such as:

1. Linear regression
2. Gradient boosting
3. AdaBoost
4. K-nearest neighbour

The results show that for broaches that are worn, the area of the signal is larger since more force is needed to make a cut resulting in more stretch. This is analogous to a dull knife requiring more force to make a cut. What is also noticed is that the number of peaks is inversely proportional to the number of cuts done by the broach. Since the blades are getting worn out, they don't scrape enough metal which is why we see a lower amount of peaks. For the regression metrics we used Mean Absolute Error (MAE). Our model achieved a MAE of 27.58 when predicting the current cut cycle. That corresponds to 98.18% accuracy of the model that predicts in which cycle the needle needs to be replaced. In Table 7 we can see the MAE results on all the tested regressors.

Regressor	MAE
Linear regression	101.25
Gradient boost	27.58
AdaBoost	165.44
KNN	74.16

Table 7: Regressors and their corresponding MAE metrics.

7 Discussion

Manufacturing is a large field which encompasses many different production areas, from automotive to pharmaceutical, to also food and packaging. Our proposed quality control models are from a range of applications and each is tailored to the specific use case. Several papers from the Related Work section are comparable to our solutions. For example, we use CNNs in detecting imperfections in oven faceplates, which similar application is used in [11]. Although the accuracy of their proposed network is higher, we showcase that there is underlying knowledge which can be used in pre-trained networks that brings forth faster training. In the future, custom networks can be built to facilitate a more specific problem learning. Moving on to oven fans, we see that good results are achieved by using simpler techniques such as extracting image features, glare reduction, thresholding and so on. While the work in Table 1 analyzes only static products, our scenario posed a different problem because of the movement of the oven fans. Taking a look at random forest classifiers, comparable results are achieved in our 3.3 section in regards to the efforts of [12]. The results from our random forest classifier can be seen in Table 4.

8 Conclusion

In this paper we presented tests of many different methods in improving the quality control pipeline in manufacturing, using typically only a couple of imperfect products. Our contribution includes novel methods used in conjunction with machine learning such as processing 3D point clouds and vibration analysis of industrial products. An addition to that is the application of various machine learning models on data obtained from an industrial environment. We presented augmentation techniques to increase our dataset size since obtaining defective industrial products in large numbers is difficult. The results from each applied method were presented in the previous sections and they are encouraging. With more data it is possible to train better classifiers which will have better performance in classifying defective samples from the production line. This research tests different kinds of industrial products for some of which only a sparse number of samples could be obtained. It is worth mentioning that some of the niche industrial products have such specific defects that so far using machine learning methods proved to be unsatisfactory. The current classifiers have difficulties learning meaningful information from those samples, nevertheless results on other industrial products give hope that new models with better quality can be learned if sufficient data is acquired.

The purpose of this study was to test which hardware and which methods should be applied in the second stage of a project, that is, actual implementation. The study showed that several approaches demand special care for our specific real-life quality control, namely:

1. Some methods like DNNs demand lot of learning examples. In some cases, it was possible to artificially generate these learning examples and the machine learning approach was successful, otherwise the approach failed.
2. Depending on the difficulty of the quality control, cheaper or more expensive hardware was needed. For several visual problems, quality cameras were sufficient, but for demanding visual problems the only hardware that enabled good results was the 10.000 euro 3D hardware system.

3. Machine learning proved successful when sufficient number of learning examples were given, either real or artificially generated. However, each approach needed special machine learning methods. For example, DNs sometimes performed best and sometimes worst. In summary: having quality input data enables creation of proper machine learning models.

References

[1] Núria Banús et al. "Deep learning for the quality control of thermoforming food packages". In: *Scientific Reports* 11.1 (2021), pp. 1–15. DOI: 10.1038/s41598-021-01254-x.

[2] Mark-Anthony Bray and Anne E Carpenter. "Quality control for high-throughput imaging experiments using machine learning in cellprofiler". In: *High Content Screening*. Springer, 2018, pp. 89–112. DOI: 10.1007/978-1-4939-7357-6_7.

[3] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[4] D Coupek et al. "Proactive quality control system for defect reduction in the production of electric drives". In: *2013 3rd International Electric Drives Production Conference (EDPC)*. IEEE. 2013, pp. 1–6. DOI: 10.1109/edpc.2013.6689762.

[5] Carlos A Escobar and Ruben Morales-Menendez. "Machine learning techniques for quality control in high conformance manufacturing environment". In: *Advances in Mechanical Engineering* 10.2 (2018), p. 1687814018755519. DOI: 10.1177/1687814018755519.

[6] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778. DOI: 10.48550/arXiv.1512.03385.

[7] Kaiming He et al. "Identity mappings in deep residual networks". In: *European conference on computer vision*. Springer. 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38.

[8] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708. DOI: 10.48550/arXiv.1608.06993.

[9] Anil Mital, M Govindaraju, and B Subramani. "A comparison between manual and hybrid methods in parts inspection". In: *Integrated Manufacturing Systems* (1998). DOI: 10.1108/09576069810238709.

[10] Antje Niederlein et al. "Image analysis in high content screening". In: *Combinatorial chemistry & high throughput screening* 12.9 (2009), pp. 899–907. DOI: 10.2174/138620709789383213.

[11] Przemysław Oborski and Przemysław Wysocki. "Intelligent Visual Quality Control System Based on Convolutional Neural Networks for Holonic Shop Floor Control of Industry 4.0 Manufacturing Systems". In: *Advances in Science and Technology. Research Journal* 16.2 (2022), pp. 89–98. DOI: 10.12913/22998624/145503.

[12] Ricardo Silva Peres et al. "Multistage quality control using machine learning in the automotive industry". In: *IEEE Access* 7 (2019), pp. 79908–79916. DOI: 10.1109/access.2019.2923405.

[13] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015), pp. 91–99. DOI: 10.1109/tpami.2016.2577031.

[14] Jacqueline Schmitt et al. "Predictive model-based quality inspection using Machine Learning and Edge Cloud Computing". In: *Advanced engineering informatics* 45 (2020), p. 101101. DOI: 10.1016/j.aei.2020.101101.

[15] Michael Sharp, Ronay Ak, and Thomas Hedberg Jr. "A survey of the advancing use and development of machine learning in smart manufacturing". In: *Journal of manufacturing systems* 48 (2018), pp. 170–179. DOI: 10.1016/j.jmsy.2018.02.004.

[16] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014). DOI: 10.48550/arXiv.1409.1556.

[17] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826. DOI: 10.1109/cvpr.2016.308.

[18] Chris J Turner et al. "Intelligent decision support for maintenance: an overview and future trends". In: *International Journal of Computer Integrated Manufacturing* 32.10 (2019), pp. 936–959. DOI: 10.1080/0951192x.2019.1667033.

[19] Jason Wang, Luis Perez, et al. "The effectiveness of data augmentation in image classification using deep learning". In: *Convolutional Neural Networks Vis. Recognit* 11 (2017), pp. 1–8. DOI: 10.48550/arXiv.1712.04621.

[20] Jing Yang et al. "Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges". In: *Materials* 13.24 (2020), p. 5755. DOI: 10.3390/ma13245755.

[21] HK Yuen et al. “Comparative study of Hough transform methods for circle finding”. In: *Image and vision computing* 8.1 (1990), pp. 71–77. DOI: 10.1016/0262-8856(90)90059-e.

An Efficient Procedure for Removing Salt and Pepper Noise in Images

Guangyu Xu and Muhammad Jibril Aminu
E-mail: xgy761220@163.com, gyxu@aust.edu.cn
School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China

Keywords: noise removal, salt and pepper noise, edge preservation, extremum-compressing median filtering

Received: January 10, 2022

In this paper, we propose an efficient algorithm for removing salt and pepper noise in images. The process of denoising is implemented in two stages: noise detection followed by noise removal. For noise detection, two extreme intensity values in an image are used to detect possible "noise pixels". For noise removal, the switching mechanism only selects "noise pixels" for processing to avoid altering any fine image details, and only the identified noise-free pixels are used to achieve better denoising performance. Two filtering techniques, the edge-preserving filtering (EPF) and the extremum-compressing median filtering (ECMF), are employed for edge-preserving and noise removal. The EPF provides higher correlation between the corrupted pixel and neighborhood pixel, which gives rise to better edge preservation. The ECMF can yield an appropriate estimation by selecting the median pixel from the noise-free pixels of current filtering window. The proposed algorithm is tested on different images and provides a better restoration performance over some of the salt and pepper noise filters.

Povzetek: V prispevku je predstavljena učinkovita metoda za odstranjevanje šuma iz slik.

1 Introduction

Digital images are often corrupted by salt and pepper noise in the process of image acquisition and transmission. Salt and pepper noise is a special case of impulse noise, where a certain percentage of pixels in the image are randomly digitized into two extreme intensities. Normally, these intensities have the maximum and minimum intensities within the dynamic range. It is very important to remove such noise before subsequent image processing tasks such as edge detection, segmentation or object recognition is carried out, because the occurrence of salt and pepper noise can severely damage the information or data embedded in the original image. To this end, a variety of techniques have been proposed for removal of salt and pepper noise.

The standard median filter (MF) [1] was once the most popular filter for removing impulse noise because of its good denoising power and computational efficiency. However, since each pixel in the image is replaced by median value in its neighborhood, the median filter often removes desirable details and blurs it, too. The weighted median filter [2] and the center-weighted median filter [3] were proposed to improve the MF by giving more weight to some selected pixels in the filtering window. Unfortunately, these two filters are still carried out uniformly across the image without considering whether the current pixel is noise free or not. These filters are effective only for low noise densities.

Over the years, better noise removal methods with different kinds of noise detectors have been proposed, for example, switching median filter (SMF) [4], adaptive median filter (AMF) [5], tri-state median filter (TSMF) [6], adaptive center-weighted median filter (ACWMF) [7], decision-based algorithm (DBA) [8], [9], a two-stage filter [10], etc. The basic idea of the methods is that the noisy pixels are detected first and eliminated afterward, whereas the uncorrupted pixels are left unchanged to prevent the blurred or removal of fine details. The main drawback of these filters is that they only use median values or their variations to restore the noisy pixels without considering local features such as the possible presence of edges, and hence they cannot usually preserve image details even when the images are corrupted by low noise densities. To overcome the above drawbacks, an efficient edge-preserving algorithm (EEPA) in [11] was introduced for removal of salt and pepper noise without degrading image fine details. It only performs well when an image is corrupted by 50% salt and pepper noise or lower. Kenny and Nor in [12] proposed a noise adaptive fuzzy switching median filter (NAFSMF) for removal of salt-and-pepper noise. This method can suppress high density of noise, at the same time preserving fine image details. Similarly, an efficient adaptive fuzzy switching weighted mean filter was presented in [13] for salt and pepper noise removal. In [14], an adaptive type-2 fuzzy method was introduced for eliminating salt and pepper noise.

Apart from the median-based filters, some other types of methods were also presented. The edge-preserving regularization method [15] and total variation L1 regularization [16] were applied to the noisy pixels to preserve the edges and noise suppression. These two methods are shown to be most efficient in dealing with

high-density of salt-and-pepper noise, meanwhile preserving image edges. In [17], the cardinal B-splines was used to restore the noisy pixels by selecting four nearly noise-free pixels. A novel filter based on the continued fractions interpolation was introduced in [18] for removal of salt-and-pepper noise. Then, Ramadan [19] used the simple average filtering for the noise suppression and edge-preserving; however, it is unsatisfactory in restoring details and edges, especially when the noise density is high. Moreover, the weighted average based filtering strategy has successfully been applied in removing salt and pepper noise [13], [20]. In [21], a universal impulse noise filter was proposed to remove any type of impulse noise by combining the noise detection with the nonlocal means filter.

In this paper, an efficient algorithm is proposed for removing salt-and-pepper noise from noisy images. The process of denoising is carried out in two stages: noise detection followed by noise removal. For noise detection, the noisy pixels are detected, based on the fact that their values must be the extreme intensities of the image. For noise removal, the switching mechanism only selects "noise pixels" for processing to avoid altering any fine image details, and only the identified noise-free pixels are used for noise removal to achieve better denoising performance. The proposed algorithm employs two different filtering techniques for edge-preserving and noise removal, namely edge-preserving filtering (EPF) and extremum-compressing median filtering (ECMF). The EPF adopts a directional correlation filtering technique based on observing the sample correlations of four different directions. Higher correlation gives rise to better edge preservation. The ECMF can yield an appropriate estimation of the processed pixel by selecting the median pixel from the noise-free pixels in the current filtering window. Extensive simulations show that the proposed algorithm not only can provide better performance of suppressing salt and pepper noise, but also can preserve more details. Furthermore, our method produces good restoration results in filtering color images.

The rest of the paper is organized as follows. In Section 2, noise model is introduced. The proposed algorithm is presented in Section 3. Experimental results are given in Section 4. Section 5 concludes the work in this paper.

2 Noise model

When an image is corrupted by salt and pepper noise, a portion of the pixel is changed with random values. To be precise, let $x_{i,j}$ and $u_{i,j}$ denote the intensity values of a noise-free image and the noisy image at the pixel location (i,j). Let the dynamic range of the image be $[L_{min}, L_{max}]$. If the noise ratio is p, then

$$u_{i,j} = \begin{cases} n_{i,j}, & with\ probability\ p \\ x_{i,j}, & with\ probability\ 1-p \end{cases} \quad (1)$$

where $n_{i,j}$ is the intensity value of the noisy pixel in noisy image u. There are two models of impulse noise: the salt and pepper noise where $n_{i,j}$ is equal to L_{min} or L_{max}, and the random-valued impulse noise where $n_{i,j}$ takes random values from the interval $[L_{min}, L_{max}]$ with a uniform distribution. In this paper, we only focus on the removal of salt and pepper noise.

3 Proposed algorithm

The proposed algorithm is a two-stage filter for salt and pepper noise detection and removal. Initially, the detection stage utilizes two extreme intensity values to identify possible "noise pixels". These detected "noise pixels" will be subjected to the following filtering action, while noise-free pixels are left unchanged. Then, two different filtering techniques, namely the EPF and the ECMF, are employed for edge-preserving and noise removal.

3.1 Noise detection

Noise detection is based on the fact that a digital image corrupted with salt and pepper noise would produce two extreme intensity values, the minimal intensity value L_{min} and the maximal intensity value L_{max}. One can identify these two intensities by analyzing the noisy image histogram, where the isolated peaks clearly indicate the two salt and pepper noise intensities. Normally, for 8-bit grayscale images corrupted by salt and pepper noise, noisy pixel only takes either L_{min} or L_{max} as its intensity value. Therefore, these two noise intensities will be used to identify possible "noise pixels" in the images. A binary noise map $F_{i,j}$ will be created to mark the location of "noise pixels" by using

$$F_{i,j} = \begin{cases} 0, & u_{i,j} = L_{\min}\ or\ L_{\max} \\ 1, & otherwise \end{cases} \quad (2)$$

where $F_{i,j} = 1$ represents noise-free pixels to be retained from the noisy image, whereas $F_{i,j} = 0$ represents "noise pixels" subjected to the next filtering process.

Since the detection of "noise pixels" is based on L_{min} and L_{max}, noise-free pixels may be falsely identified as "noise pixels" at image uniform regions having same values as L_{min} or L_{max}. In this case, it is difficult for the filter to determine the size of the filtering window. On the other hand, when the filtering action is applied these pixels, their values may be changed. In [19], the intensity values 0 and 255 are employed to identify noise pixels, but the author does not provide the solution on how to prevent the above problems from occurrence. In the section 3.3, we describe in detail how to deal with this problem.

3.2 Edge-preserving filtering (EPF)

After the binary noise map $F_{i,j}$ is created, noise-free pixels marked with $F_{i,j} = 1$ will be retained and the filtering action is omitted to avoid altering any details in the original images, whereas "noise pixels" marked with $F_{i,j} = 0$ will be replaced by an estimated correction term.

Before introducing the EPF, we review a basic fact, that is, a noise-free image consists of locally smoothly varying areas separated by edges. Considering this characteristic, the proposed EPF adopts a directional correlation-dependent filtering technique based on observing the sample correlations of four main directions. Here, we only focus on the edges aligned with four main directions. The proposed algorithm uses a $(2N+1)\times(2N+1)$ filtering window $W_{i,j}^{N}$, given as

$$W_{i,j}^{N}=\{u_{i+s,j+t}\} where s,t\in(-N,\cdots,0,\cdots,N). \quad (3)$$

Then, the number of noise-free pixels, $G_{i,j}^{N}$ in the window $W_{i,j}^{N}$, is counted using

$$G_{i,j}^{N}=\sum_{s,t\in(-N,\cdots,0,\cdots,N)}F_{i+s,j+t}. \quad (4)$$

For each "noise pixel" marked with $F_{i,j}=0$, the EPF will detect edges in its four directions in the 3×3 filtering window, i.e., $W_{i,j}^{1}$. For simpler representation, let $a,\ b,\ c,\ d,\ e,\ f,\ g,$ and h represent those pixel values, $u_{i-1,j-1}$, $u_{i-1,j}$, $u_{i-1,j+1}$, $u_{i,j-1}$, $u_{i,j+1}$, $u_{i+1,j-1}$, $u_{i+1,j}$, and $u_{i+1,j+1}$, respectively, around the current pixel $u_{i,j}$ as shown in Figure 1. The detailed steps of the proposed EPF are described as follows.

If the current filtering window $W_{i,j}^{1}$ has at least two noise-free pixels, i.e., $G_{i,j}^{1}>1$, do the following.

Step 1: Calculate the four directional differences around the pixel $u_{i,j}$ in the $W_{i,j}^{1}$.

$$\begin{cases}D_1=|a-h|, & D_2=|b-g|\\ D_3=|c-f|, & D_4=|d-e|\end{cases} \quad (5)$$

Step 2: Check whether the pixels ($a,\ b,\ c,\ d,\ e,\ f,\ g,$ and h) are possible "noise pixels" marked as $F_{i,j}=0$, respectively. If yes, the pixel might be corrupted, and thus we do not consider the directional differences containing it by setting those differences to 512.

Step 3: Find the minimum value among four directional differences and denote it as D_{min}. The minimum directional difference has the strongest correlation and probably has an edge in its direction. Hence, the restored value $y_{i,j}$ of the corrupted pixel $u_{i,j}$ is estimated as follows:

$$y_{i,j}=\begin{cases}\frac{(a+h)}{2,} & if D_{min}=D_1\\ \frac{(b+g)}{2,} & if D_{min}=D_2\\ \frac{(c+f)}{2,} & if D_{min}=D_3\\ \frac{(d+e)}{2,} & if D_{min}=D_4\end{cases} \quad (6)$$

Note that the EPF is used only when $G_{i,j}^{1}>1$, here the filtering window size is 3×3. The reason is that there is lower directional correlation between the central pixel and its neighbors which are spatially far away from the central pixel.

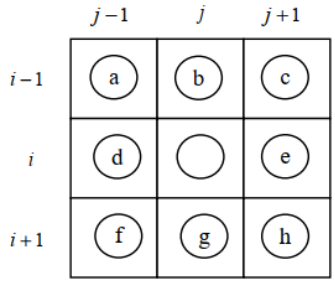


Figure 1: The pixels around the current pixel.

3.3 Extremum-compressing median filtering (ECMF)

Obviously, the EPF can prevent the noise-free pixels from being changed and give better estimates of the noisy pixels than median alone. However, there is an exception in the EPF. If D_{min} is equal to 512, it means that there exist the corrupted pixels on four directions in the window $W_{i,j}^{1}$. In this condition, no edge is considered. Here, noisy pixels marked with $F_{i,j}=0$ are excluded and not involved in the estimate of the currently processed pixel. That is, only those noise-free pixels marked with $F_{i,j}=1$ are used as candidates for selecting the median pixel, $y_{i,j}$, given by

$$y_{i,j}=median\{u_{i+s,j+t}\}\ \ with\ \ F_{i+s,j+t}=1. \quad (7)$$

On the other hand, if the current filtering window $W_{i,j}^{1}$ does not have a minimum number of one noise-free pixel (i.e., $G_{i,j}^{1}=0$), then the filtering window will be expanded by one pixel at each of its four sides. This procedure is repeated until the criterion of $G_{i,j}^{N}\geq1$ is met, and then the filtering action is applied to the current pixel. For example, when $G_{i,j}^{2}\geq1$ (i.e., the window $W_{i,j}^{2}$ contains noise-free pixels), the median pixel in the $W_{i,j}^{2}$ will be selected as the estimate of the current pixel $u_{i,j}$ by using (7). Accordingly, if $G_{i,j}^{3}\geq1$, the median filtering based on (7) is applied to the window $W_{i,j}^{3}$. This strategy of choosing only noise-free pixels is imposed to avoid selecting a "noise pixel" as the estimated median.

As mentioned above, the noise detection based on L_{min} and L_{max} may falsely identify noise-free pixels as "noise pixels" at image uniform regions having same intensities as L_{min} or L_{max}. Consequently, the filtering window will be expanded continuously and the selected median pixel may be improper to be used as an estimate term. Considering this possibility, the search for "noise-free pixels" is halted when the size of the window has reached 7×7 (or $N=3$) although no "noise-free pixel" is marked, i.e., $G_{i,j}^{3}=0$. Here, the window $W_{i,j}^{1}$ centered at (i,j) will be used to estimate the median pixel $y_{i,j}$, i.e.,

$$y_{i,j}=median\{u_{i+s,j+t}\}\ \ where\ \ s,t\in(-1,0,1). \quad (8)$$

By using (8), the proposed algorithm can avoid modifying uncorrupted pixels at image uniform regions having same intensities as L_{min} or L_{max}, meanwhile removing corrupted pixels. We call the above approach

as the ECMF. However, it would remove the isolated noise-free pixels with intensity L_{min} or L_{max}. To the best of our knowledge, this is still an unsolved problem.

4 Experimental results and discussions

4.1 Experimental environment

In this section, we compare our method with a number of existing filters. To verify the characteristics and performances of various filters, a variety of simulations are carried out on six 8-bit gray-scale test images: Lena, Boats, Peppers, Goldhill, Barbara, and Rice. In the simulations, images are corrupted with salt and pepper noise, where 0 represents the "pepper" noise and 255 represents the "salt" noise with equal probability. A wide range of noise ratios varied from 10% to 90% with increments of 10% is tested. Totally, five recent filters are compared with our method in terms of objective testing (quantitative evaluation) and subjective testing (visual quality): AMF [5], ACWMF [7], DBA [8], EEPA [11], NAFSMF [12]. In each experiment, the parameters or thresholds of the compared methods are set as suggested by its authors. To be fair, for high-density of noise, AMF, ACWMF, and EEPA filters are iteratively executed to obtain the best results.

The experiments are implemented in Matlab R2018b on PC equipped with an Intel Core i7-8700 3.20GHz, 4.0GB RAM, and Windows 10. We employ the peak signal-to-noise ratio (PSNR) to assess the quantitative quality of the restored images for various methods. Higher PSNR values indicate better image restoration.

4.2 Experimental results

The comparison of restoration results in PSNR for six test images (Top to bottom and left to right: Lena, Boats, Peppers, Goldhill, Barbara, and Rice) corrupted with various salt and pepper noise ratios are shown in Figure 2. It is easy to see that our method is very good in removing salt-and-pepper noise and has the highest PSNR values at all levels of noise. In Table 1, we list the restoration results in PSNR of different filters for images "Lena" and "Boats" corrupted with salt and pepper noise. Apparently, our method provides the best results than the others.

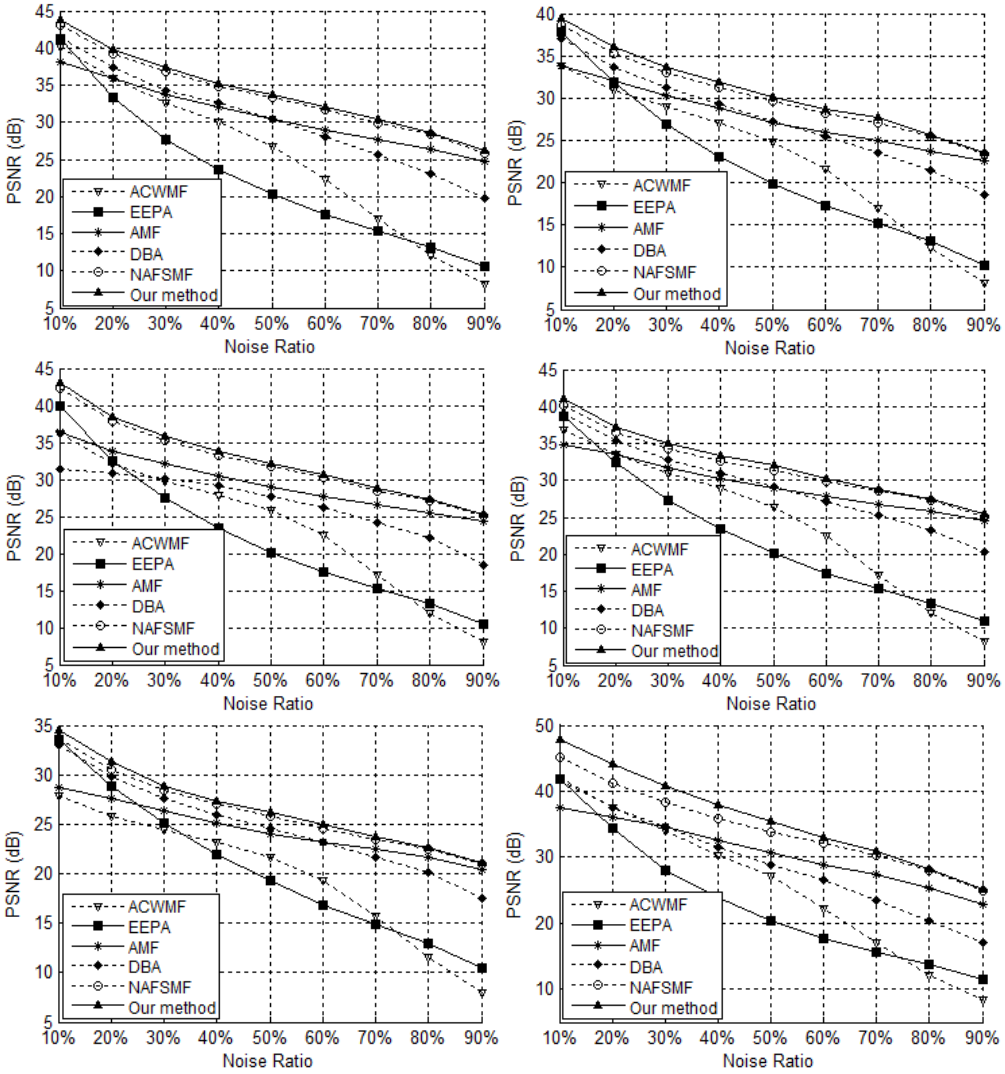


Figure 2: Comparison of restoration results in PSNR.

Filters	Lena			Boats		
	30%	70%	90%	30%	70%	90%
ACWMF	32.83	17.09	8.15	28.96	16.93	8.05
EEPA	27.74	15.31	10.53	26.97	15.18	10.27
AMF	33.66	27.75	24.80	30.36	24.92	22.65
DBA	34.31	25.72	19.86	31.31	23.50	18.63
NAFSMF	36.85	29.95	25.84	32.99	27.04	23.34
Our method	**37.43**	**30.37**	**26.14**	**33.64**	**27.67**	**23.54**

Table 1: Comparisons of restoration results in PSNR.

Furthermore, a subjective visual result of the noise removal is presented in Figure 3. Figure 3(b) is the noisy "Boats" image with 70% salt and pepper noise. Among the restorations, the ACWMF and EEPA have plenty of visible noise patches and produce disappointing results, whereas the AMF, DBA, NAFSMF, and our method work well to remove noise. One can observe that our method gives the best performance in terms of noise removal and edge preservation such as ropes between masts.

In Figure 4, we show the restoration results of different filters in restoring image "Lena" corrupted with 90% salt and pepper noise. Obviously, only the NAFSMF and our method can produce perceptible reconstructed images. However, our method has a better noise suppression ability as compared to the NAFSMF since the restored images does not have any spot of unremoved noise. Moreover, our method can preserve edges better. The blurring caused by the NAFSMF is mainly due to its median-based restoration mechanism. The DBA can suppress noise satisfactorily, but at the expense of image details, causing edges to be jagged and distorted. The AMF is capable of removing noise, but it blurs image details significantly. On the other hand, the ACWMF and EEPA have completely failed to remove noise.

The average processing time takes by the filters at various noise densities is shown in Figure 5, where a total of six test images are processed. Due to the computational complexity of the edge-preserving filtering is high, our method has a higher processing time than the AMF and DBA, but in general it is lower than the ACWMF, EEPA, and NAFSMF. However, the higher processing time is compensated by the better restoration results as shown above.

4.3 Application on color images

In this paper, the RGB color space is selected to represent the color images. The noisy color images are obtained by applying the noise model in (1) to each of the R-, G-, and B- channels independently. This means when a color image is corrupted by salt and pepper noise with noise ratio p, then each color component image is being corrupted with p. Accordingly, our proposed filter can be extended for removing noise from corrupted color images by applying the proposed algorithm to the R-, G-, and B- channels independently.

Simulation results for color image "Lena" corrupted with 80% salt and pepper noise are depicted in Figure 6. It is observed that our proposed mehtod exhibits excellent noise suppression performances, meanwhile preserving image fine details well.

5 Conclusion

An efficient algorithm for the removal of salt-and-pepper noise is presented. To remove the salt-and-pepper noise with edge and fine detail preservation, the switching mechanism is used to avoid altering any fine details in the images, and then the EPF provides more edge details, leading to better edge preservation.

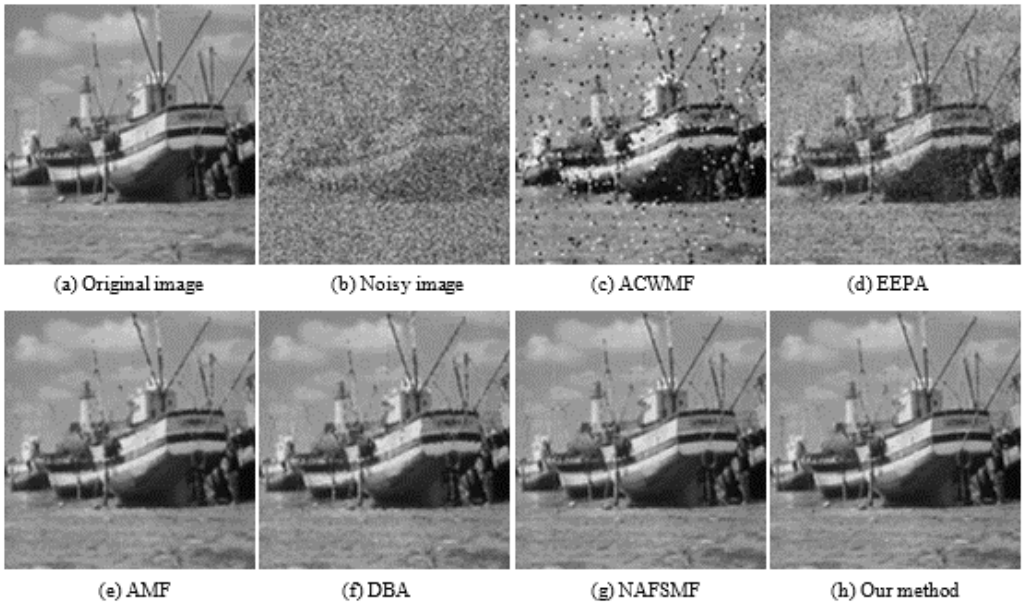


Figure 3: Comparison of visual results for image "Boats".

Meanwhile, the ECMF is employed for the high-densities of noise. The proposed algorithm shows consistent and stable performance across a wide range of noise densities varying from 10%-90%. In contrast to the other existing filters, it produces the best restoration results both visually and quantitatively.

Acknowledgements

The author would like to thank the editors and the anonymous reviewers for their valuable suggestions. This work is supported by the National Natural Science

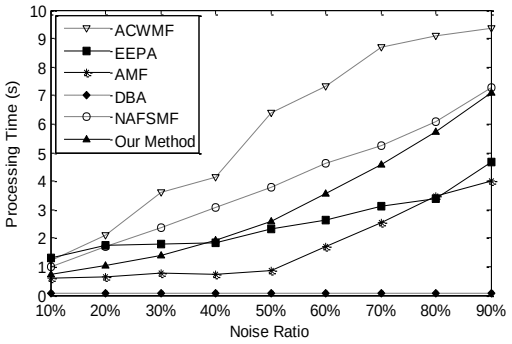


Figure 4: Average processing time (s) versus noise ratios.

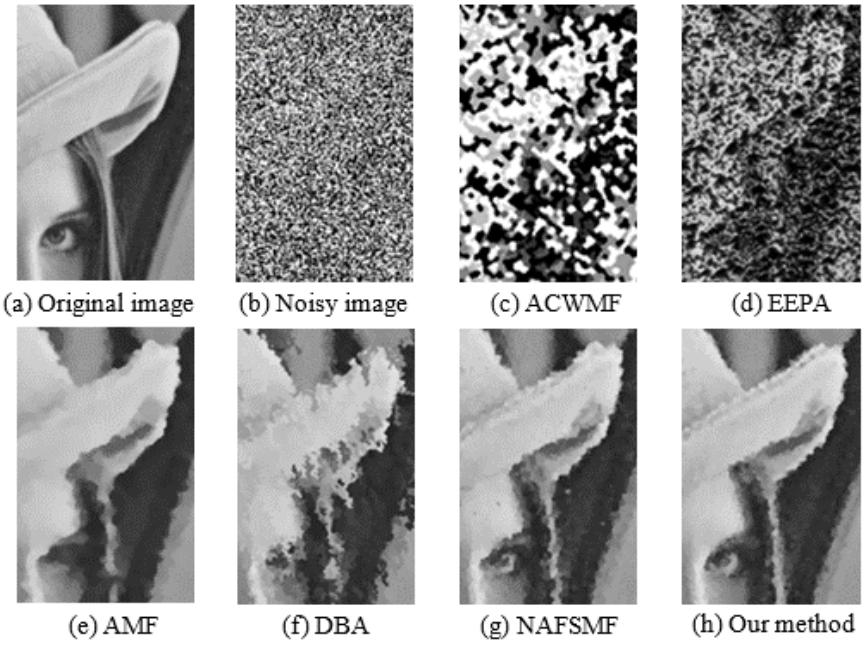


Figure 6: Comparison of visual results for image “Lena”.

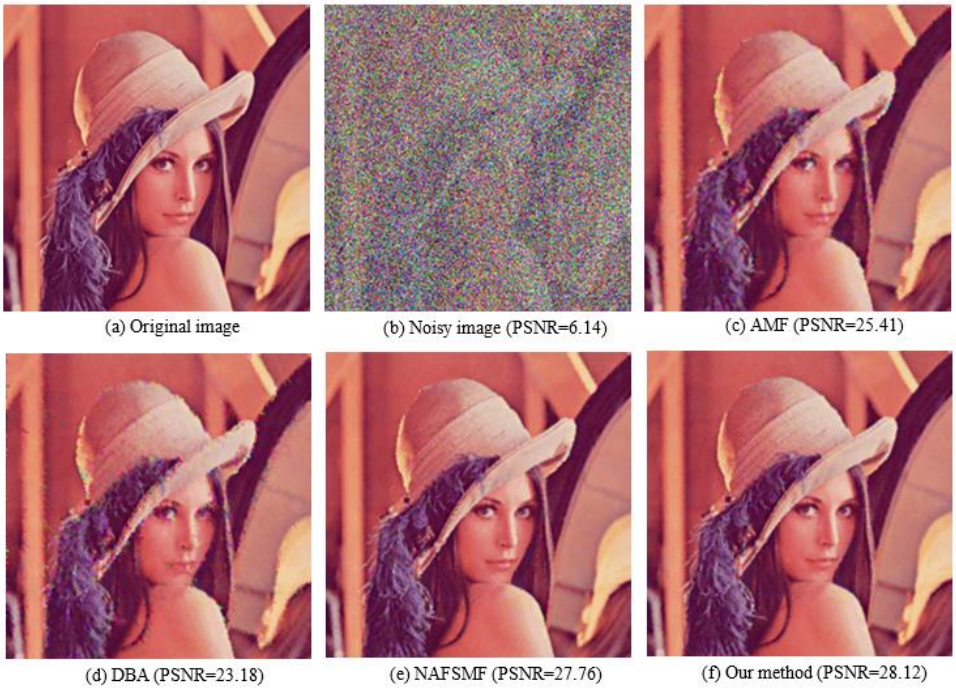


Figure 5: Comparison of visual results for color image “Lena”.

Foundation of China (Grant No. 61471004) and by Doctoral Foundation of Anhui University of Science and Technology (Grant No. ZX942).

References

[1] Huang T. S., Yang G. J., and Tang G. Y. (1979). A fast two-dimensional median filtering algorithm. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(1), pp. 13-18. https://doi.org/10.1109/tassp.1979.1163188.

[2] Brownrigg D. R. K. (1984). Weighted median filter. Communications of the ACM, 27(8), pp. 807-818. https://doi.org/10.1145/358198.358222.

[3] Ko S. J., and Lee Y. H. (1991). Center weighted median filters and their applications to image enhancement. IEEE Transactions on Circuits and Systems, 38(9), pp. 984-993. https://doi.org/10.1109/31.83870.

[4] Sun T., and Neuvo Y. (1994). Detail-preserving median based filters in image processing. Pattern Recognition Letters, 15(4), pp. 341-347. https://doi.org/10.1016/0167-8655(94)90082-5.

[5] Hwang H., and Haddad R. A. (1995). Adaptive median filters: New algorithms and results. IEEE Transactions on Image Processing, 4(4), pp. 499-502. https://doi.org/10.1109/83.370679.

[6] Chen T., Ma K. K., and Chen L. H. (1999). Tri-state median filters for image denoising. IEEE Transactions on Image Processing, 8(12), pp. 1834-1838. https://doi.org/10.1109/83.806630.

[7] Chen T., and Wu H. R. (2001). Adaptive impulse detection using center-weighted median filter. IEEE Signal Processing Letters, 8(1), pp. 1-3. https://doi.org/10.1109/97.889633.

[8] Srinivasan K. S., and Ebenezer D. (2007). A new fast and efficient decision-based algorithm for removal of high-density impulse noises. IEEE Signal Processing Letters, 14(3), pp. 189-192. https://doi.org/10.1109/ lsp.2006.884018.

[9] Kkishorebabu V., and Varatharajan R. (2020). A decision based unsymmetrical trimmed modified winsorized variants for the removal of high density salt and pepper noise in images and videos. Computer communications, 154, pp. 433-441. https://doi.org/10.1016/j.comcom.2020.02.048.

[10] Thanh D. N. H., Hai N. H., Prasath V. B. S., Hieu L. M., and Tavares J. M. R. S. (2020). A two-stage filter for high density salt and pepper denoising. Multimedia tools and applications, 79(29), pp.21013-21035. https://doi.org/10.1007/s11042-020-08887-6.

[11] Chen P. Y., and Lien C. Y. (2008). An efficient edge-preserving algorithm for removal of salt-and-pepper noise. IEEE Signal Processing Letters, 15, pp. 833-836. https://doi.org/10.1109/lsp.2008.2005047.

[12] Toh K. K. V., and Mat Isa N. A. (2010). Noise adaptive fuzzy switching median filter for salt-and-pepper noise reduction. IEEE Signal Processing Letters, 17(3), pp. 281-284. https://doi.org/10.1109/ lsp.2009.2038769.

[13] Wang Y., Wang J., Song X., and Han L. (2016). An efficient adaptive fuzzy switching weighted mean filter for salt-and-pepper noise removal. IEEE Signal Processing Letters, 23(11), pp. 1582-1586. https://doi.org/10.1109/ lsp.2016.2607785.

[14] Singh V., Dev R., Dhar N. K., Agrawal P., and Verma N. K. (2018). Adaptive type-2 fuzzy approach for filtering salt and pepper noise in grayscale images, IEEE Transactions Fuzzy system, 26(5), pp.3170-3176. https://doi.org/10.1109/tfuzz.2018.2805289.

[15] Chan R. H., Ho C. W., and Nikolova M. (2005). Salt-and-pepper noise removal by median-type noise detectors and detail preserving regularization. IEEE Transactions on Image Processing, 14(10), pp. 1479-1485. https://doi.org/10.1109/tip.2005.852196.

[16] Thanh D. N. H., Thanh L. T., Hien N. N., and Prasath V. B. S. (2020), Adaptive total variation L1 regularization for salt and pepper image denoising. Optik, 208, pp. 1-10. https://doi.org/10.1016/j.ijleo.2019.163677.

[17] Jayasree P. S., Raj P., Kumar P., Siddavatam R., and Ghrera S. P. (2012). A fast novel algorithm for salt and pepper image noise cancellation using cardinal B-splines. Signal Image Video Process, 5(2), pp. 1-8. https://doi.org/10.1007/s11760-012-0368-3.

[18] Bai T., and Tan J., Hu M., and Wang Y. (2014). A novel algorithm for removal of salt and pepper noise using continued fractions interpolation. Signal Processing, 102, pp. 247-255. https://doi.org/10.1016/j.sigpro.2014.03.023.

[19] Ramadan Z. M. (2012). Efficient restoration method for images corrupted with impulse noise. Circuits, Systems, and Signal Processing, 31, pp. 1397-1406. https://doi.org/10.1007/s00034-011-9380-z.

[20] Enginoglu S., Erkan U., and Memis S. (2019), Pixel similarity-based adaptive riesz mean filter for salt-and-pepper noise removal, Multimedia Tools Applications, 78, 35401-35418. https://doi.org/10.1007/s11042-019-08110-1.

[21] Xu G., and Tan J. (2014). A universal impulse noise filter with an impulse detector and nonlocal means. Circuits, Systems, and Signal Processing, 33, pp. 421-435. https://doi.org/10.1007/s00034-013-9640-1

Exploring the Parametric Impact on a Deep Learning Model and Proposal of a 2-Branch CNN for Diabetic Retinopathy Classification with Case Study in IoT-Blockchain based Smart Healthcare System

Manaswini Jena[1], Smita Prava Mishra[1], Debahuti Mishra [1], Pradeep Kumar Mallick[2] and Sachin Kumar[3*]
E-mail: manaswini.jena88@gmail.com, smitamishra@soa.ac.in, debahutimishra@soa.ac.in, pradeepmallick84@gmail.com, sachinagnihotri16@gmail.com
[1]Department of Computer Science and Engineering, Siksha 'O' Anusandhana Deemed to be University, Bhuvaneswar Odisha, India
[2]School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhuvaneswar, Odisha, India
[3]Department of Computer Science, South Ural State University, Chelyabinsk, Russia

Keywords: healthcare system, 2-branch CNN, diabetic retinopathy, fundus images, medical diagnosis, internet of things

Received: January 9, 2022

Smart healthcare has changed the way how the patient interacts with the specialists for treatment. However, security and support for various diseases are still the concern for such smart automated systems. One of the critical diseases namely Diabetic Retinopathy (DR), is a major concern for the person with prolonged diabetes and may lead to complete blindness irrespective of age groups. Moreover, in recent years blockchain has gained popularity in providing secure communication between sender and receiver. Hence, this work focus on designing a blockchain-based smart healthcare system for the early detection of diabetic retinopathy. However, early detection of DR impose complexities and requires expert diagnosis, which is not available everywhere. Hence, the proposed smart healthcare model contains a Computer-Aided Diagnosis (CAD) assistance for early detection of symptoms of the disease. The CAD model may assist the ophthalmologists in the early detection of DR, which requires intensive research in developing an efficient and accurate model that can operate without human interaction. This study provides an empirical analysis of these factors to design the best model for early detection of DR. The best model can be used to develop IoT based smart devices to detect DR in diabetic patients. The study also explains the importance of IoT and blockchain-based technology for the development of smart healthcare systems. The values of the parameters and type of hyperparameters choosen from the study is used in a proposed 2-branch CNN model, and the model is validated using the Kaggle fundus image set. Analysis of various parameters and using their best values gives an outstanding performance in the proposed 2-branch CNN model.

Povzetek: Predstavljena je bločna metoda za varno komunikacijo med zdravnikom in pacienti, pri tem je uporabljena metoda globokih nevronskih mrež za problem diabetične slepote.

1 Introduction

Healthcare systems are an essential part of modern society to save people's lives and provide services remotely. However, the concern of the healthcare system is the need for the protection of information of the patient and correctly recognizing the symptoms for early detection. Diabetic Retinopathy(DR) is one of the fatal diseases where early diagnosis is very crucial. Computer-Aided Diagnosis(CAD) assists experts in the early diagnosis of the disease. In recent years, researchers have developed many CAD models to detect DR effectively. However, a complete healthcare model is not yet developed. This research attempts to propose a CAD model with security to protect patient data and a deep learning model to detect DR accurately.

Blockchain, IoT, and AI have shown their potential in almost every domain of our life. Smart healthcare is one of the major sectors that has been influenced by IoT infrastructure and solutions [1-2]. IoT-based healthcare systems have immensely added value to our lifestyle and health monitoring with the use of portable wearable devices. However, there are few complex healthcare areas where Blockchain and IoT have to do some evolutionary steps to protect the healthcare data to ensure patient confidentiality.

Similarly, Convolutional Neural Network (CNN) is widely used for accurate CAD in the healthcare domain. These automatic diagnostic systems can improve the quality and productivity of healthcare services. Moreover, CAD can help in the early detection of diabetic retinopathy, one of the primary causes of blindness in the world [3]. The limited resources of healthcare facilities lead to difficult ac-

*Corresponding author

cess to expert doctors for diabetic retinopathy analysis [4]. Currently, CAD includes convolutional networks to identify the early signs of diabetic retinopathy disease. The CNN is a type of deep feed-forward artificial neural network. This architecture is mainly designed and adopted to classify images using a multilayer architecture [5-7]. Various types of CNNs are proposed and their efficiencies are proved in the complicated image classification techniques [8-11]. Prolonged diabetics result in diabetic retinopathy leading to the significant cause of vision loss worldwide [10-11]. Several technical implementations have done using CNN for diabetic retinopathy identification [12-17]. The performance of the CNN model depends on the number of convolution layers, the number of nodes in convolution layers with several parameters, and hyper-parameters such as learning rate, activation functions, and pooling [18-19]. To our best knowledge, no study has been carried out that can justify the choices made in designing a CNN for diabetic retinopathy classification.

The fundamental units for this convolution network are shared weight, bias, local perceptive, and pooling. The activation function and pooling are critical for model performance. The activation function is an elementary part of every neural network architecture. Moreover, it is responsible for transforming the summed weighted input into its output [20-21]. The activation function and threshold are the basic terminology for every neural network for execution and flow of network as the activation of neurons depend upon it [22-23]. Non-linear transformations and back-propagation are made using activation functions. Several different activation functions are utilized for NN [24-27]. Pooling is applied as a downsampling approach to reduce the input signal resolution without affecting the representable, essential, and large feature elements [24-25]. In addition to activation and pooling, the learning rate and optimizer used for back-propagation in the network has also a great impact on the performance of the neural network. So, all four features are going to be analyzed for the classification of diabetic retinopathy using the CNN model.

The learning rate is a hyper-parameter used to control the learning speed of the neural network model. The key role of the learning rate is to scale the magnitude of weight to reduce the network's loss function. The training or learning is generally performed through updating weights to map the input with the output. These updated weights are calculated by analytical methods via empirical optimization techniques called stochastic gradient descent (SGD). The stepping amount needed to change the weights during the training process is called the learning rate. This is the value used for adjusting the weights according to loss gradient through back-propagation [26].

The value of the learning rate is a small positive value in the range of 0.0 to 1.0. It may vary differently for different models and we can find a good learning rate value via the trial and error method. Though the optimal learning rate calculation is analytically not possible yet performance improvement can be done by achieving a good learning rate. During training, the learning rate can be adjusted to improve the performance and this is called adaptive learning [27-28]. Between the epochs or iterations, the value can be increased or decreased, instead of having a fixed value. The change in value is made is called a learning rate schedule. Here, an attempt is made to visualize the change in accuracy and performance of the model according to the change in learning rate values.

Another feature considered here is the types of optimizers used for back-propagation in the neural network. Back-propagation is a weight updater in neural networks used to update the weights by optimizing the values to get the target result. For this optimization, SGD is generally [29-30] used. SGD is an iterative optimization method that uses shuffled samples for the evaluation of the gradient values. Sometimes the random probabilistic nature of SGD for choosing samples results in a noisy path to reach the destination from typical gradient descent algorithms as it may take a higher number of iterations due to randomness. But, the computationally inexpensive nature of this algorithm makes it suitable for usage. Another type of optimizer available to be used instead of this SGD is Adam. The memory requirement is very less for this optimizer as well as it can be used for large problems in terms of data/parameters and appropriate for non-stationary objectives along with a problem with noisy gradients [31]. AdaDelta is another optimizer based on a moving window of gradient updates. It does not accumulate the past gradients rather continues learning from updates [32]. Here these three optimizers are chosen for the experimentation.

The objective of this study is to analyze the impacts of these four features i.e. activation function, pooling, learning rate, and optimizer parameters on a CNN model for diabetic retinopathy classification. The study is conducted in phases to analyze the impact of each parameter. Finally, the best specifications are used in a proposed CNN model for diabetic retinopathy classification.

The rest of the article is organized as follows: Section II represents the state-of-the-art literature review. Section III and IV provides the methodolgy with model description and an overview of the dataset respectively. Section V provides the experiment details. The results and discussion are in Section VI . Finally, Section VII concludes the study.

2 Literature review

IoT, Blockchain and Cloud technologies are integrated in the medical environment for offering healthcare and telemedical laboratory services along with AI. For the classification of diabetic retinopathy through fundus image CNN is widely used and proved as a great classifiers, but sometimes the data taken directly from the patient may not give the expected accurate report. More features and parameters need to be tuned very minutely and this is the main objective here. This section represents some studies where CNN is used for the classification of DR. In [33] transfer learn-

ing is used on pre-trained CNN models and an accuracy of 74.5% is achieved for binary classification model. Tertiary and quaternary classification models are also tested for the classification the study. A set of pre-processing technique and data augmentation is followed by transfer learning for the experiment. In [34] the CNN used with augmentation for the classification of diabetic retinopathy gives better result compared to the CNN used without augmentation *i.e.* 94.5% and 91.5% respectively. Here augmentation improves the performance. DIARETDB1 data is used in [35] for classification of DR by N. Gharaibeh *et al.* Here, in the first phase feature extraction is done followed by classification of features acquiring 98.4% accuracy using CNN. In [39] retinal images are transformed into entropy images to improve the classification. The complexity of original fundus images are represented by quantifying the information of the image. Then classification is done using deep learning which improves the accuracy to 86.0% from 81.8% that was achieved using normal fundus images. Image enhancement techniques like histogram equalization is used in [57] along with deep leaning to improve the classification of DR and an accuracy of 97% is achieved. In [37] the segmented image classification using CNN for DR achieves an accuracy of 99.18% accuracy results. Another work in [36] uses CNN along with Support vector machine (SVM) as classifier and results with 86.17% accuracy. Here max-pooling is replaced by fractional max-pooling in the CNN for extraction of features before using it in SVM. Following Table 1 highlights the findings of the above discussion:

However, a study [56] states that although high accuracy is achieved for detection of DR using various models of machine learning, but the requirements of real time deployment and clinical validation is not yet fulfilled completely. It also indicates to solve the problem arising due to the diversity of datasets in terms of ethnicities and camera used for capturing the images in real practical clinical field. So, fine-tuning of every parameters and features of the network is required to get a perfect model. However, this study of various features and parameters and their effects will help in development of prediction models. Furthermore, IoT/Blockchain based devices provides portable and secure technology for healthcare monitoring, however it is a challenging task to achieve such facility with high security. In [41], authors proposed an approach to deal with these challenges for Internet of Medical Things [IoMT]. Several research studies focused on the effectiveness of IoT and bloackchain for the healthcare ecosystem [42-43]. In addition, cloud-server can be utilized to provide a global secure network to monitor a large number of petients in several regions with cloud based IoT network in health care domain [44-47].

It is observed that, image features are also a very important factor [48], to obtain a high classification performance. Various techniques have been used for accurate classification of the image by extracting the discriminative features. The amalgamation of both global and local features can be fed to a classifier for classification of the image. For this a 2 branch CNN can be designed having the ability to extract features in an optimized way for the classification in a unified optimized framework [49]. A multi-domain CNN performs better than a single frequency domain CNN and spatial domain CNN for detection of image compression [50]. Hyperspectral image classifications show the dominance of a multi-branch CNN in many studies [51-53]. These kinds of architecture extracts feature from each source of the image at their native pixel of resolution and it is done by exploiting both spatial and spectral information from the image. Keeping this in mind, the workflow is designed to create a better model for diabetic retinopathy detection as compared to the existing models using the analysis.

3 Methodology

At first a two class classification model using several CNN configurations were developed to understand the importance of several parameters for the classification of DR in diabetic patients. These configurations are mentioned in Table 2 as follows:

The result of the empirical analysis is used in a 5 class classification model. A two-branch CNN is proposed for the classification of DR using fundus image. The proposed flow diagram of a two-branch CNN model is given in **Figure 1**. Two separate feature sets collected from two separate branches of CNN are assembled and used in the fully connected neural network for classification of DR.

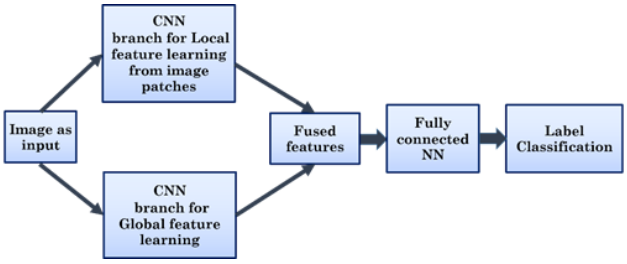


Figure 1: A two branch CNN model.

Some basic pre-processing steps are applied here for the classification as follows:

(i) **Cropping:** Extracting the part of the region of interest (ROI) from the image is done through cropping. Most of the fundus images have a black border which needs to be reduced as possible to concentrate on the region of interest.

(ii) **Resizing of image:** For any machine learning technique or deep learning technique, resizing of the image is required. All the images of data need to be standardized. Here, all the images are converted into one size *i.e.* 256×256×3.

Table 1: Comparison of existing studies for DR classification.

Model used	Dataset used	Accuracy
CNN with pre-processing, feature extraction and augmentation (used for binary classification) [33]	Messidor MildDR	74.50%
CNN with augmentation [34]	Kaggle	94.50%
CNN without augmentation [34]	Kaggle	91.50%
Feature extraction and classification using CNN [35]	DIARETDB1	98.40%
Transformed entropy images and classification using CNN [36]	Kaggle	86.00%
Histogram Equalization and classification using CNN [57]	Messidor	97%
Segmentation and pre- processing with CNN [37]	DiaretDB0,	99.17%
	DiaretDB1	98.53%
	DrimDB	99.18
CNN with SVM as classifier [38]	Kaggle	86.17%

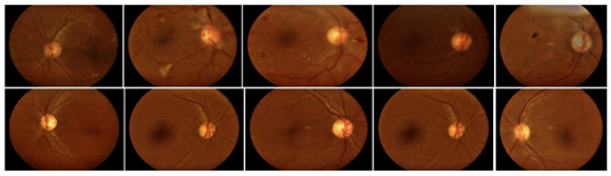


Figure 2: First row represents diseased fundus images and second row represents normal fundus images.

(iii) **Contrast Enhancement:** It is a process of making the image features more prominent by changing the range of pixel values in the image. Contrast Limited Adaptive Histogram Equalization (CLAHE) is the method which improves contrast in digital images. The results of medical image enhancement using this technique are better than other histogram equalization techniques [54]. Here, this CLAHE is used for the contrast enhancement to make the hidden features visible.

(iv) **Normalization:** Max-min normalization is applied here for normalizing the pixel values.

4 Dataset description

The dataset under study have been obtained with the support of the administration of Ruby Eye hospital in Orisha, India. The fundus images have been captured with a high resoulion cameras. The database shows a male and female ratio of 50:26 with diabetic retinopathy patients' in 30-65 age group. For diabetic retinopathy, the separated fundus images of diseased and non-diseased patients are collected. The samples of the infected and non-infected fundus images are shown in Fig 2.

The dataset consists of the fundus images of 102 diabetic patients with the presence of both healthy and diseased images with equal distribution. The two classes of these fundus images are either diabetic retinopathy (DR) *i.e.* the diseased or normal *i.e.* non-diseased. Normalized of the images are performed with max-min normalization in the range [0-1] in order to achieve zero mean and unit variance. This dataset is available online publically for research purpose [55].

Kaggle retina fundus image dataset is also used here for experiment. It contains the images based on five severity levels such as 0=No DR, 1=Mild NPDR, 2=Moderate NPDR, 3=Severe NPDR, 4=PDR. For both the data, the training and testing ratio is taken as 80:20 here.

5 Experimental analysis

The detailed evaluation of hyper-parameters on a large model is a computationally intensive task. Hence, the study uses simple model to evaluate various hyper-parameter settings like pooling, activation, learning rate, and optimization. Finally, the proposed 2-branch CNN model uses the best combination of hyper-parameters learnt from the simple model.

To analyse the significance of pooling, activation, learning rate, and optimization speciation of the CNN model, multiple experiments have been conducted, and these are divided into four phases.

In the first phase (Phase 1), the performance of the parameters or factors are analyzed to be selected for further testing. Following this in the second phase (Phase 2), a selective combination of factors are evaluated to establish the relationship of these to the model performance. In next two phases (Phase 3 and Phase 4), more factors are evaluated by incremental tests to obtain maximum performance for fundus image classification using this dataset.

Here, data augmentation is used for virtually increasing the number of samples in the dataset during training. It is a technique used to manufacture or generate own data

Table 2: Used CNN model and its layer description.

Type	Layer Size	Learnables	Total Learnables
Image_Input	$256 \times 256 \times 3$	-	0
Convolution	$242 \times 242 \times 32$	Weights $15 \times 15 \times 3 \times 32$ Bias $1 \times 1 \times 32$	21632
Batch_Normalization	$242 \times 242 \times 32$	Offset $1 \times 1 \times 32$ Scale $1 \times 1 \times 32$	64
ReLU	$242 \times 242 \times 32$	-	0
Max_Pooling	$121 \times 121 \times 32$	-	0
Convolution	$115 \times 115 \times 64$	Weights $7 \times 7 \times 32 \times 64$ Bias $1 \times 1 \times 64$	100416
Batch_Normalization	$115 \times 115 \times 64$	Offset $1 \times 1 \times 64$ Scale $1 \times 1 \times 64$	128
ReLU	$115 \times 115 \times 64$	-	0
Max_Pooling	$57 \times 57 \times 64$	-	0
Convolution	$53 \times 53 \times 128$	Weights $5 \times 5 \times 64 \times 128$ Bias $1 \times 1 \times 128$	204928
Batch_Normalization	$53 \times 53 \times 128$	Offset $1 \times 1 \times 128$ Scale $1 \times 1 \times 128$	256
ReLU	$53 \times 53 \times 128$	-	0
Max_Pooling	$26 \times 26 \times 128$	-	0
Convolution	$24 \times 24 \times 128$	Weights $3 \times 3 \times 128 \times 128$ Bias $1 \times 1 \times 128$	147584
Batch_Normalization	$24 \times 24 \times 128$	Offset $1 \times 1 \times 128$ Scale $1 \times 1 \times 128$	256
ReLU	$24 \times 24 \times 128$	-	0
Max_Pooling	$12 \times 12 \times 128$	-	0
Keras Flatten	$1 \times 1 \times 18432$	-	0
Fully_Connected	$1 \times 1 \times 100$	Weights 100×18432 Bias 100×1	1843300
ReLU	$1 \times 1 \times 100$	-	0
Fully_Connected	$1 \times 1 \times 2$	Weights 2×100 Bias 2×1	202
ReLU	$1 \times 1 \times 2$	-	0
Classification_Output	-	-	0

with the existing data [56]. Data augmentation have been used to deform the labelled data without semantic information loss[57]. Here five online data augmentations are used such as horizontal and vertical shifts, horizontal and vertical flips and rotation. The horizontal shift and vertical shifts are of 2% and rotation of 90% is applied. The modified input images are introduced during the training process automatically. For experiment, the dataset is divided in the ratio of 80:20 for the training and testing data.

5.1 Evaluation of the parameters

For evaluation various matrices like accuracy, sensitivity, specificity, precision, recall and F_1-*Score* [58-62] are taken into consideration.The analysis of results is done step by step through different phases as described below:

Phase 1: As described before in the experimental setup, different pooling layers, activation functions, optimization parameters and learning rates are evaluated to understand the behaviour of the model. The obtained results for variation in parameter combination are presented pictorially in the following graphs. The 3-dimensional bar graph representation is used to visualize the behaviour of the model with respect to testing accuracies achieved for different combinational approaches made for the CNN through the process of classification of diabetic retinopathy data.

Four pooling types such as max pooling, min pooling, average pooling and maxmin pooling is considered here. For each pooling, the variation in activation function, learning rate are presented in x-axis and y-axis respectively. Furthermore, four widely used activation functions such as, *LReLu, ReLU*, *sigmoid* and *tanh* are chosen to be studied with combination of learning rate of 0.1, 0.01, 0.001 and 0.0001. Finally, the observed testing accuracy is given in the z-axis. The learning ability of the model with respect to various optimizers such as SGD, ADAM, Adadelta is also observed by the colour indicated in graph respectively. Figure 3 shows the changes in accuracy value for max pooling with different activation functions, learning rate and different optimizer for backpropagation in the CNN model.

The Figure 3 (a), (b), (c), and (d) are the 3-dimensional graphical representation of accuracy values measured using max pooling at epoch 25, 50, 75, and 100 respectively.

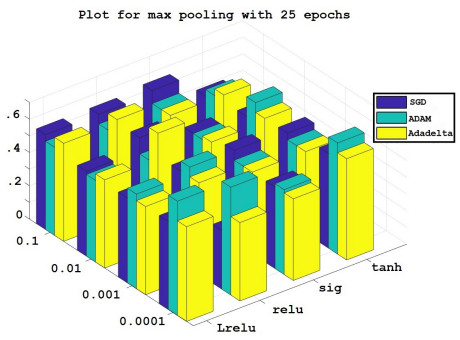
From the Figure 3(a) it is observed that, at epoch 25, a maximum accuracy result of 0.686 is obtained by using LReLu and ReLu activations at 0.01 learning rate with Adadelta optimizer. The same accuracy is observed at 0.0001 learning rate with ADAM optimizer at 100 epochs. The results show that the increase of epoch does not ensure better accuracy here. Moreover, the accuracy can also decrease. This occurs due to overtraining which affects the results or due to improper tuning of the learning rate.

The graphical Figure 3(b) shows increased accuracy values as compared to Figure 3(a). The two activation functions ReLu and LReLu at 0.1 learning rate with Adadelta optimizer present accuracies of 0.725 and0.705, respectively.

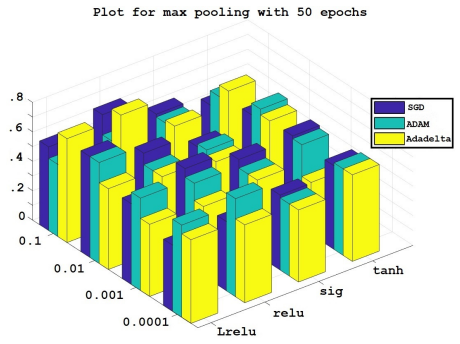
Figure 3(c) shows that the Adadelta optimizer presents an accuracy of 0.725 with tanh activation at 0.01 learning rate for epoch 75. Moreover, the ADAM optimizer also presents reliable results with an accuracy value 0.745 with LReLu at 0.0001 learning rate. Figure 3(d) shows the accuracy of the model by using max pooling layer at epoch 100. The best accuracy of 0.686 is achieved with ReLu activation at 0.01 learning rate and SGD optimization and also by the tanh activation function at learning rate of 0.1 and ADAM optimization. The LReLu activation achieves an accuracy of 0.647 with SGD and ADAM optimizers at 0.1 and 0.0001, respectively. The sigmoid function presents the lowest accuracy results.

Figure 4 represents the 3-dimensional graph for the second type of pooling taken min pooling at different epochs and features. The Figure 4(a) is the plot for accuracy values at epoch 25 using min pooling in the model. Here, the ADAM optimizer at 0.0001 learning rate with ReLu activation function achieves an accuracy of 0.764 and the SGD optimizer at 0.01 with LReLu activation gets a 0.66 accuracy. In Figure 4(b) for epoch 50, the ADAM optimizer at 0.0001 learning rate with ReLu activation gets an accuracy of 0.686. In this same condition LReLu activation increases the accuracy to 0.725. In this specific situation, the LReLu performs better than the ReLu activation. Again, the Adadelta optimizer at 0.0001 learning rate with LReLu gives the same accuracy as 0.725. The Figure 4(c) is the accuracy plot taking epoch value as 75, where only one combination gives a marked accuracy result of 0.686 by Adadelta optimizer at 0.0001 learning rate and tanh activation function. The Figure 4(d) gives the bar graph representation for epoch 100. In this point, the LReLu activation at 0.01 learning rate and ADAM optimizer in the model gets an accuracy of 0.705 and tanh function at 0.1 learning rate and Adadelta optimizer gets an accuracy of 0.686. The observed results shows that, the min pooling and max pooling in the model combined with ADAM or Adadelta optimizer performs better than other optimizer combinations. The following **Figure 5** represents the plotting for another type of pooling called average pooling taken in the model.

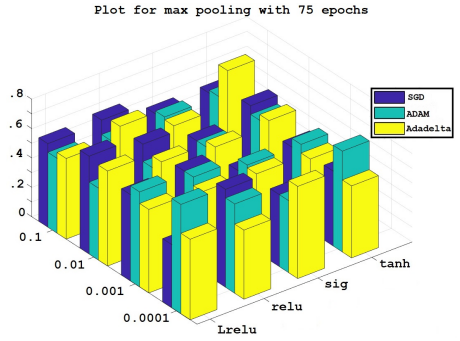
The Figure 5(a) is for epoch 25 which yields two important accuracy observation as 0.686 and 0.705 using ADAM and Adadelta optimizers. In the Figure 5(b), the ADAM optimizer achieves an accuracy of 0.666 with ReLu activation at 0.0001 learning and 0.745 with LReLu at 0.01 learning rate. From the graph it also can be said that the average pooling method in the network model along with SGD optimizer and learning rate of 0.01works well with tanh function resulting an accuracy value of 0.705 compared to ReLu function that yields 0.666 accuracy. The Figure 5(c) is plotted for 75 epochs and yields accuracies of 0.666 and 0.686 by using SGD optimizer at 0.1 and 0.01 learning rate with ReLu and LReLu activations. The ADAM optimizer also reaches an accuracy of 0.764 with tanh activation function at 0.0001 learning rate. But with increased epoch of 100 at Figure 5(d) the accuracy values does not increase further, rather staying at value of 0.666 by SGD and ADAM opti-



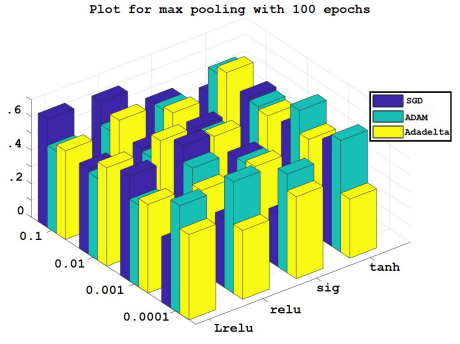
(a)



(b)

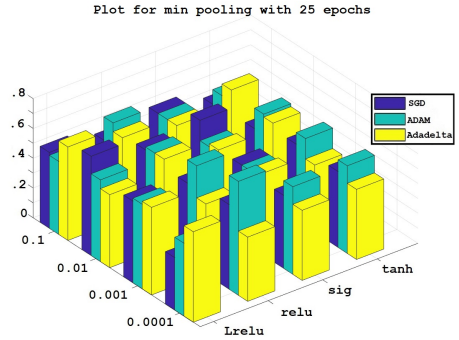


(c)

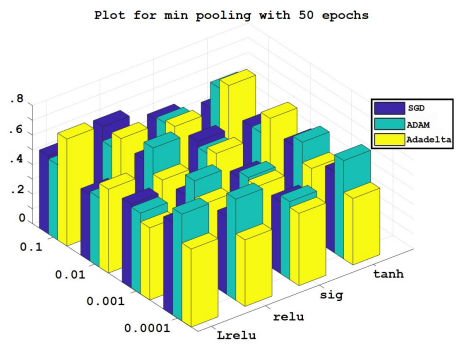


(d)

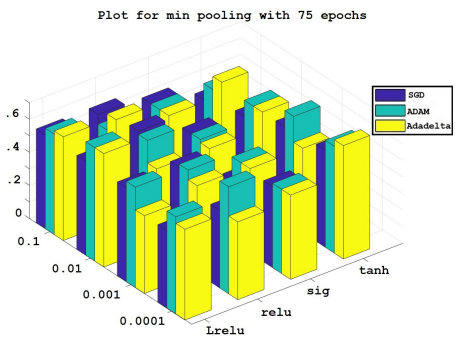
Figure 3: Plotting of accuracy values by taking max pooling in the model. **(a)** is for epoch =25, **(b)** is for epoch =50 **(c)** is for epoch =75 **(d)** is for epoch =100.



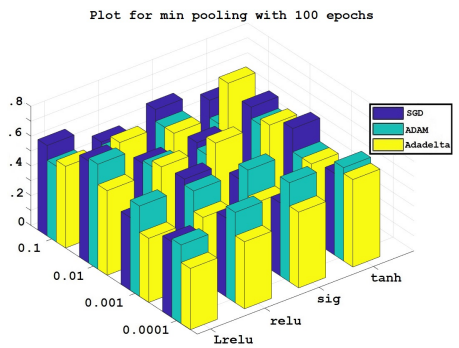
(a)



(b)

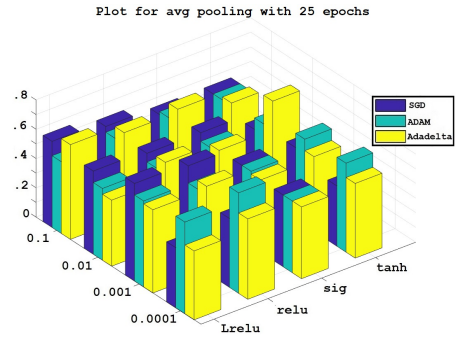


(c)

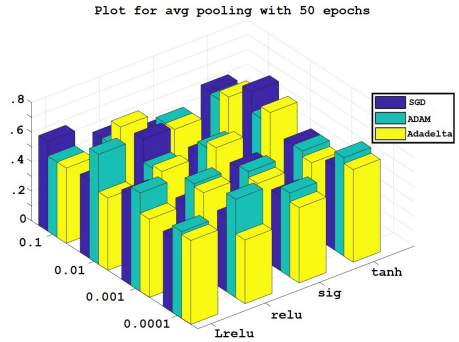


(d)

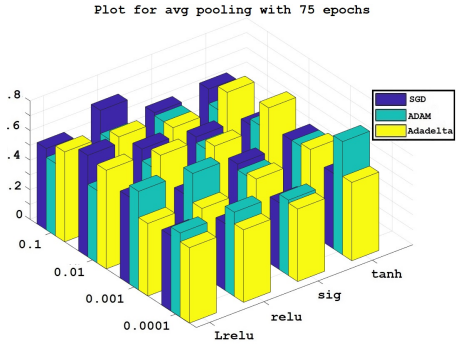
Figure 4: Plotting of accuracy values by taking min pooling in the model. **(a)** is for epoch =25, **(b)** is for epoch =50, **(c)** is for epoch =75 and **(d)** is for epoch =100.



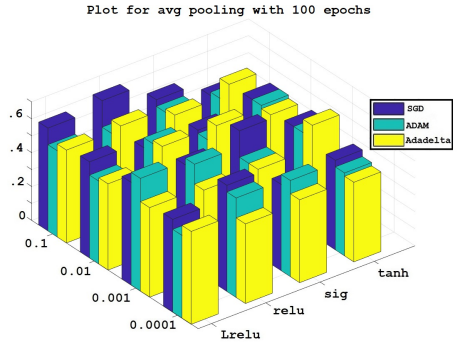
(a)



(b)



(c)



(d)

Figure 5: Plotting of accuracy values by taking average pooling in the model. **(a)** is for epoch =25, **(b)** is for epoch =50, **(c)** is for epoch =75 and **(d)** is for epoch =100.

mizers. Till now with these three pooling layers ReLu and LReLu activation functions performs well on most of the cases, while tanh performs in few and the sigmoid function seems to be a poor performer at this point.

In this similar procedure the accuracy values achieved by using max-min pooling in the model along with different parameters are also analyzed. By taking 25 epochs in the model, it reaches to an accuracy value of 0.803 using ADAM optimizer at 0.0001 learning rate with tanh activation function. This accuracy does not increase in further increase of epochs to 50, 75 or 100, rather falls to 0.666 in some combinational approaches. The experimental results clearly conveys that every condition, every combination of parameter is effecting the model's behaviour. Each combinational approach gives dissimilar results to analyze. From these results, selective combinations, which gives the highest result among all, are chosen to test with more increased epoch and taken to phase two.

Phase 2: In the second phase, the accuracy and other evaluation parameters are analyzed for the chosen combinations. The following **Table 3** shows the selected accuracy values for different pooling with combination of features and parameters chosen along with its sensitivity (Sn), specificity (Sp), precision (Pr), recall (Re) and F_1-score (F) values. Above table contains evaluation parameters for the combinations through which the maximum accuracy values are achieved in phase 1. It shows that, the max pooling combined with Adadelta optimizer at 0.1 learning rate goes well for achieving a good accuracy values. Finally, the ADAM optimizer with LReLu and 0.0001 learning rate attains the maximum accuracy among this for max pooling. At this accuracy value, the sensitivity is 0.76 and the specificity the precision are 0.73 with recall and F_1-score to be 0.76 and 0.745 respectively. From the combinational approaches for min pooling the ADAM optimizer gives better results as compared to other optimizers and the learning rate 0.0001 can be selected for proper tuning here. The LReLu function also proved to be more performable than ReLu and other activation functions in this situation. The actual positive cases that are correctly identified is measured by sensitivity and the maximum of it is 0.79 for the accuracy 0.76 and specificity of 0.74 while the precision, recall and F_1-score are 0.73,0.8 and 0.76 respectively. These are the best values obtained in this spot.

While taking average pooling in this model at various epochs the tanh activation seems to perform better at this state. At learning rate 0.01 accuracies are increased but the maximum accuracy yield here is at 0.0001 learning rate which is 0.764. The sensitivity and specificity are 0.791 and 0.740 while, the precision, recall and F_1-score are 0.73, 0.8 and 0.76 respectively. Here the ADAM optimizer gives the maximum result compared to all other optimizers.

In this case of maxmin pooling, the ADAM optimizer performs very well compared to other optimizers again. It gains an accuracy of 0.803 and a sensitivity of 0.9, which is the maximum among all the previous results. Till now this is the best combinational approach for the model us-

ing this dataset. The specificity, precision, recall and F_1-score are 0.741, 0.692, 0.9 and 0.782 respectively. Here it is visualized how the change in pooling layer along with its parameters and features effects the behaviour of model. The *LReLu activation* works well for *min pooling* and *tanh function* for *average pooling*, but in other pooling they are not the same. The ADAM optimizer gives better result in *min pooling* and *maxmin pooling*, however AdaDelta optimizes well in max pooling for this model with this data. The graphs and tables demonstrates the clear conceptualization of the independent behaviour of parameters and conditions in a particular scenario. Now, all of the above combinations of all four pooling layers are taken to phase 3 for further experiment.

Phase 3: The above combinations are the repeated with increased iteration up to 1500 and the acquired outputs are analyzed. The following tables compares the accuracy calculated at 100 and less than 100 iterations with accuracy values at 1500 iterations. In this the actual set of combinations are selected *i.e.* the set of states which gives higher results are finally selected at top best combinations for this particular model. The accuracy at 1500 iterations are shown in Table 4 for max pooling. From the above table, it is marked that the accuracy value which is actually the maximum among the four values is not increased more with the increase in epoch. Another new combination of features has acquired an increased highest accuracy of 0.84 with the increased epoch of 1500. The sensitivity is 0.87 while the specificity is 0.81. The precision, recall and F_1-score are 0.8, 0.88 and 0.84 respectively. So for max pooling this can the best combination of parameters and features to get a good result using this model. It can be seen in table 4 that in min pooling the expected accuracy value is not achieved, rather the accuracies decreased with the increase in number of iterations. It may happen due to overtraining which affects the result in opposite way.

Table 4 also illustrates the accuracy values and other evaluation parameter values for the average pooling in the model. Here also the increased accuracy value is not the further increased value from the previous phase. The Adadelta optimizer performs well here compared to ADAM as before for this average pooling at learning rate 0.01 rather than 0.0001. It achieves an accuracy of 0.862 with a sensitivity of 0.85 which is smaller than the specificity 0.875. Here some more tuning may increase the accuracy even more by increasing the sensitivity. The precision, recall and F_1-score are 0.888, 0.84 and 0.86 respectively (Table 4). In this table of accuracy representation for max-min pooling layer, the highest accuracy value from the previous phase increased to 0.86 from 0.80. Though the sensitivity was high before, still it maintains a value of 0.88 and specificity of 0.84.

The precision, recall and F_1-score are 0.85, 0.88 and 0.86 respectively. We can see that some accuracy values are increased due to increase in iteration and some of them give a decreased value as they might got stuck in local minima or may be due to overtraining. Only a few among these yields refinement of accuracies after increase in iteration and is marked with bold font. Further, those highlighted accuracy's combinational condition can be taken for experimentation with a more increased value of iteration.

Phase 4: Here the three pooling techniques *i.e.* max pooling, average pooling and max-min pooling and their combinational features are finally selected in **Table 5** which are working well for this model and dataset. Further this combinations can be used either with optimization technique or with some hybridization to achieve more suitable model having higher accuracy values.

This table contains the three maximum accuracy values of 0.843, 0.862 and 0.865 achieved in this experimental approach with other evaluation parameters. These can be taken as the best three combination of activation function, learning rate, optimizer and pooling layer to achieve a good accuracy for this model. In this paper only two class classification of diabetic retinopathy data is considered which is either DR or normal. DR is taken as a five class problem i.e. a normal or non DR class and four stages such as, mild DR, moderate DR, severe DR, and proliferative DR classes [39-40]. This empirical analysis will help us in the further deep classification of DR. Some observations are made in this experiment are stated below:

- Among various types of the pooling used here, max-min pooling gives comparatively better result. The fusion of max and min values of the pixels during the selection through pooling generates a good selection of pixel values for the model's training and testing for classification.
- A combined pooling helps in improving the performance value of model. This fused pooling technique also works well in [62].
- Unlike the ReLU function, the LReLu does not transform all the negative values to zero. Here, the LReLu performs finer than other activation functions.
- The combinations which give rise to good accuracy results, amid of them, 50% of are having LReLu as their activation function.
- Particularly, for this experiment if we grade the activation functions, LReLu takes the first position resulting as the best activation function for image classification using CNN.
- The tanh function comes in the second place and finally the ReLU comes in the third place by providing good result in few combinations. But with increase in iteration, the tanh function wins over the LReLu by obtaining more accurate results.
- Going through the accuracy results, it is noticed that more than 50% good accuracy values are seen in the cases using Adam as an optimizer here. The Adadelta optimizer also give results in less than 40% combinational cases.

Table 3: Selective combinations for various pooling.

Pooling	Activation	Optimizer	LR	Iter.	Acc.	Sn	Sp	Pr	Re	F
Max	ReLU	Adadelta	0.1	50	0.72	0.71	0.73	0.76	0.68	0.74
	LReLu	Adadelta	0.1	50	0.7	0.67	0.75	0.8	0.6	0.73
	LReLu	ADAM	0.0001	75	0.74	0.76	0.73	0.73	0.76	0.74
	Tanh	Adadelta	0.1	75	0.72	0.71	0.74	0.769	0.68	0.74
Min	ReLU	Adadelta	0.1	50	0.72	0.71	0.73	0.76	0.68	0.74
	LReLu	Adadelta	0.1	50	0.7	0.67	0.75	0.8	0.6	0.73
	LReLu	ADAM	0.0001	75	0.74	0.76	0.73	0.73	0.76	0.74
	Tanh	Adadelta	0.1	75	0.72	0.71	0.74	0.769	0.68	0.74
Avg	ReLU	Adadelta	0.01	50	0.745	0.74	0.75	0.769	0.72	0.754
	LReLu	Adadelta	0.01	25	0.7	0.789	0.656	0.57	0.89	0.666
	LReLu	ADAM	0.01	50	0.7	0.761	0.666	0.615	0.8	0.68
	Tanh	Adadelta	0.0001	75	0.764	0.791	0.74	0.73	0.8	0.76
Max-Min	LReLu	ADAM	0.001	50	0.72	0.8	0.677	0.615	0.84	0.695
	Tanh	ADAM	0.0001	25	0.803	0.9	0.741	0.692	0.92	0.782

Table 4: Combinations for several pooling selected from phase 1.

Pooling	Activation	Optimizer	LR	Acc. (<=100)	Acc. (at 1500)	Sn	Sp	Pr	Re	F
Max	ReLU	Adadelta	0.1	0.72	0.666	0.645	0.7	0.769	0.56	0.701
	LReLu	Adadelta	0.1	0.705	0.843	0.875	0.814	0.807	0.88	0.84
	LReLu	ADAM	0.0001	0.745	0.607	0.625	0.592	0.576	0.64	0.6
	Tanh	Adadelta	0.1	0.72	0.588	0.608	0.571	0.538	0.64	0.571
Min	ReLU	Adadelta	0.0001	0.764	0.549	0.565	0.535	0.5	0.6	0.53
	LReLu	Adadelta	0.1	0.725	0.549	0.56	0.538	0.538	0.56	0.549
	LReLu	ADAM	0.0001	0.725	0.627	0.684	0.593	0.5	0.76	0.577
	Tanh	Adadelta	0.0001	0.705	0.588	0.619	0.566	0.5	0.68	0.553
Avg	ReLU	Adadelta	0.01	0.745	0.647	0.63	0.66	0.73	0.56	0.678
	LReLu	Adadelta	0.01	0.7	0.862	0.85	0.875	0.884	0.84	0.867
	LReLu	ADAM	0.01	0.7	0.647	0.66	0.629	0.615	0.68	0.64
	Tanh	Adadelta	0.0001	0.764	0.607	0.615	0.6	0.615	0.6	0.615
Max-Min	LReLu	ADAM	0.001	0.72	0.686	0.66	0.714	0.769	0.6	0.714
	Tanh	ADAM	0.0001	0.803	0.865	0.88	0.846	0.85	0.88	0.867

Table 5: Combinations selected in phase 3 and their accuracy.

Pooling	Activation	Optimizer	LR	Acc. (<=100)	Acc. (at 1500)	Sn	Sp	Pr	Re	F
Max	LReLu	Adadelta	0.1	0.705	0.843	0.875	0.814	0.807	0.88	0.84
Avg	Tanh	Adadelta	0.01	0.7	0.862	0.85	0.875	0.884	0.84	0.867
Max-Min	Tanh	ADAM	0.0001	0.803	0.865	0.88	0.846	0.85	0.88	0.867

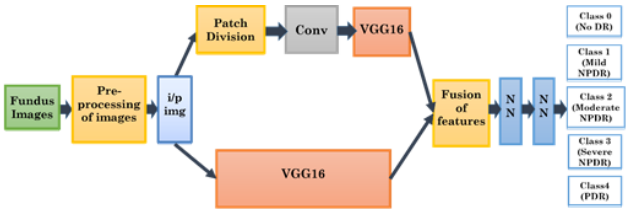


Figure 6: Proposed customized 2-branch CNN model for DR classification.

- It can be concluded that, the Adam and Adadelta works well for image in CNN compared to the SGD optimizer. Many of the cases here shows SGD to be very less powerful optimizer than others.
- The lower learning rate like 0.0001 and 0.001 obtains superior results in half of the cases in this experimental analysis for CNN.

5.2 Evaluation of the model

From the above study of parameters, max-min pooling, LReLu activation function, Adam optimizer and learning rate as 0.0001 are chosen to be applied in the proposed 2-branch CNN model for the five class classification of DR using Kaggle dataset. Each CNN branch of the 2 branch CNN model is inspired by the VGG16 network model [63]. Detailed structure of the proposed 2-branch CNN model is shown in following **Figure 6**.

The proposed model is having two branches to which inputs are supplied. These two branches are trained individually with the Kaggle dataset for thorough learning of features. After that, the output features of each branch are fused and handed over to a fully connected neural network for classification.

After pre-processing, the image of size 256×256×3 is ready to be used in the network for training. Patch size of the image given is 128×128.

For the 1st branch, for extraction of more possible features locally, the image is divided into various patches. For this patch division, functional layer Lambda is used. It is included in the TensorFlow library. These received patches are in one dimension format. So, to rearrange the extracted patches into normal input shape, the Reshape layer is used before going to VGG network through a convolution layer. Here, the convolution layer is used to make the image size suitable to be the input for VGG16 network. The pre-trained VGG16 model from the library is used without its fully connected layers. The pre-trained weights of the network help the network to achieve better accuracy. The output features are stored for further use.

In the 2nd branch, the pre-processed image is directly given as input without patch division. The VGG16 network is used without its fully connected layers and the output features of the branch are stored. Both the output features from the two branches are combined and fed to the fully connected layer. The combined features are of size 1024 (*i.e.* 512+512 from each branch), from which the five class predictions are performed.

Following is the steps used in the proposed 2-branch model for the experiment -

Model flow Steps:

Step 1: Pre-processing of the Fundus Images

- *Cropping*
- *Resizing of image*
- *Contrast Enhancement*
- *Normalization*

Step 2: Loading the pre-processed data

Step 3: Processing for the Branch 1 CNN

- *Patch division and resizing of the input images*
- *Input to the VGG16 network*
- *The output of the network is stored as A*

Step 4: Processing for the Branch 2 CNN

- *Input to the VGG16 network*
- *The output of the network is stored as B*

Step 5: The output features of branch 1 and branch 2 are concatenated

Step 6: The total concatenated output is given to a fully connected Neural Network for classification

6 Results and discussion

The results of the proposed model are discussed and compared with state-of-the-art models. Along with various assessment parameters such as accuracy, sensitivity, specificity, precision, recall, F1-score, AUC [64] and ROC [65] are used here.

6.1 Analysis of results

Following **Table 6** shows the classification result of the five classes of diabetic retinopathy classification in the form of confusion matrix. The above discussed assessment parameters are calculated from the confusion matrix and shown in **Table 7.** To analyze the result in a proper way and to access its performance, a result comparison table has been shown in **Table 8**. The model proposed by S. Qummar *et al.* [66] is also considering the same dataset and performing the five-class classification. So, the comparison between the models can be done.

The accuracy of the model is calculated as 0.966. The highest sensitivity, specificity, precision, recall and F1-score calculated are 0.992, 0.994, 0.978, 0.998 and 0.985. The higher values of sensitivity and specificity show the

Table 6: Confusion matrix from the proposed model.

			Predicted class levels				
			No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR
Actual class	Label	No DR	137	1	3	0	1
		Mild NPDR	5	135	1	0	1
		Moderate NPDR	3	0	138	0	1
		Severe NPDR	0	0	0	139	3
		PDR	1	2	0	1	138

Table 7: Assessment parameter values from the confusion matrix.

	Sensitivity	Specificity	Precision	Recall	F1-Score
No DR (0)	0.964789	0.984155	0.938356	0.991134	0.951389
Mild NPDR (1)	0.950704	0.994718	0.978261	0.987762	0.964286
Moderate NPDR (2)	0.971831	0.992958	0.971831	0.992957	0.971831
Severe NPDR (3)	0.992857	0.994737	0.978873	0.998239	0.985816
PDR (4)	0.958333	0.992933	0.971831	0.989436	0.965035

proportion of actual positives and effected subjects are classified accurately. The high precision rates of 0.978, 0.971, 0.978 and 0.971 in class 1, class 2, class 3 and class 4 shows a low false-positive rate in the classification model. The higher recall values of 0.991, 0.992 and 0.998 for class 0, class 2 and class 3 shows very low false-negative rate. The other two classes also have a false-negative rate in the classification with a recall value of 0.987 and 0.989. The weighted average of precision and recall is the F1-score which is remarkable here. All the five values of F1-score for five classes are more than 0.95 and the maximum is 0.98, which is an indicator as a good classifier.

The probability curve or the ROC curve is calculated and shown in **Figure 7**. This graph shows the capability of our model for distinguishing five classes. The performance of classification at all classification thresholds is depicted in the graph using the two-parameter *i.e.* true positive rate and false-positive rate. The AUC is calculated by measuring the entire two-dimensional area under the entire ROC curve. It provides an aggregate measure of performance for all possible thresholds in classification. AUC is scale-invariant as it measures the ranking of prediction by ignoring the absolute values. It is also a classification threshold invariant, which means it measures the quality of the models predictions without depending on the classification threshold. The AUC for class 0, class 1, class 2, class 3, and class 4 are 0.98, 0.96, 0.98, 0.98 and 0.99 respectively. The maximum AUC achieved is of 0.99 by the PDR class and the minimum achieved is 0.96 by the mild NPDR class.

A comparison between the models AUC and the AUC from S. Qummar *et al.* [66] is shown in **Table 8.** Our model outperforms in every class classification of the DR by distinguishing each class at the highest rates. The maximum AUC of 0.99 shows the distinguishing capacity of the proposed model is at its best.

The AUC values prove the model to be the best performer for the classification of diabetic retinopathy using

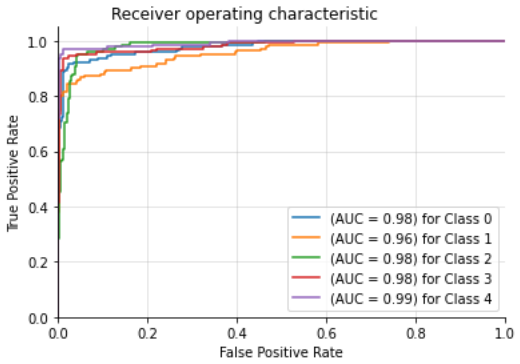


Figure 7: ROC Curve for the model.

fundus images. To give more detail about the model's training capacity, the following curves are presented in **Figure 8**. The graph **(a)** in **Figure 8** shows how the accuracy of the model increases with increasing epochs. The graph **(b)** of **Figure 8** shows that the learning curves are in good fit. The plot of the training loss decreases to a point of stability which is good for the model loss.

From all evaluation parameters calculated and the graph showing model accuracy, model loss with the ROC curve and AUC value; it may be inferred that the model outperforms the existing model and can stand as a good classifier for DR classification.

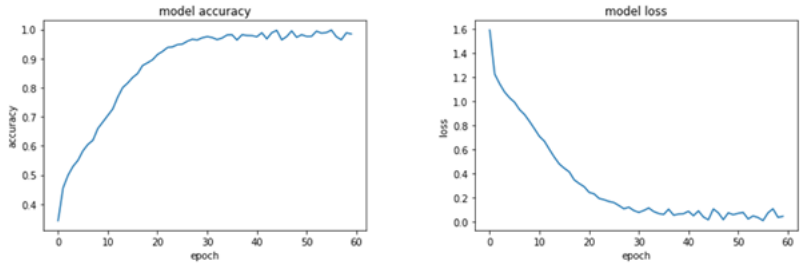
6.2 Novelty of the work

- Analysis of various parameters and using their best values gives an outstanding performance in the proposed 2-branch CNN model.

- The 2-branch classification model of CNN seems to adopt the feature learning very well here.

Table 8: Comparison with state-of-the-art model's output.

Various Class levels	AUC from S. Qummar et al. [81]	AUC of the proposed model
Class 0	0.97	0.98
Class 1	0.91	0.96
Class 2	0.78	0.98
Class 3	0.9	0.98
Class 4	0.94	0.99



(a) Graph showing model accuracy (b) Graph showing model loss

Figure 8: Proposed model's accuracy and loss.

- The detail learning of features *i.e.* the learning of local and global features of fundus images helps a lot to grasp the inter-class level differences of the diabetic retinopathy classification. The probability curve shows it clearly.
- From the obtained result it can be concluded that, use of VGG16 CNN model in this 2-branch classification frame works well for diabetic retinopathy classification.
- A good classification performance from a model can be obtained through correct selection of parameters and hyper-parameters for the model.
- The achieved AUC for each class levels summons the proposed 2-branch model as a good classifier. The accuracy and loss curve of the model also reciprocates it as a good model.

6.3 Case study in IoT and Blockchain based solution for smart healthcare

Blockchain technologies [67] in digital healthcare is very important and showing its importance on routine basis. Blockchain can be described as a powerful method to provide more secure and reliable data sharing between several organizations. Secure and reliable communication between several parties in a major challenge where blockchain technology can be very effective. One of the popular cases uses of blockchain technology is patient-centric electronic health records. In a report published by Johns Hopkins University, it was mentioned that one of the causes of patient's death was the medical errors in the patient records. Therefore, blockchain technology can provide an effective solution by introducing a blockchain based digital system in order to provide a comprehensive and accurate record to the healthcare provides about the patient by linking it with the existing medical records of the patient. IoT in the other hand can incorporate blockchain-based solutions into smart devices that can be available in real time to the patient as well as the health care provider. It will also help to monitor the health of the people in real time basis. One framework is proposed in Figure 9.

The framework consists of vital sign monitoring system, an IoT server, a blockchain network, and a communication interface to collect the information from the healthcare sensors from the patients. All the information is stored securely and communicated to the medical staff for further diagnosis and treatment. The approach once development can be optimized to develop an IoT based smart device and can be used by the medical staff efficiently.

7 Conclusion

Blockchain, AI and IoT based smart healthcare systems needs some prior research that can be optimized further and can be incorporated into portable smart devices. DR is one of the complicated eye disease in all age group people around the world and prior detection and proper treatment is the only way to prevent the high percentage of visual loss through surgical, pharmacological or laser treatment. This is a pragmatic study about the reflected behaviour of a CNN model due to various factors. In this paper, we have investigated different CNN configurations by considering several factors such as the type of pooling layer, activation

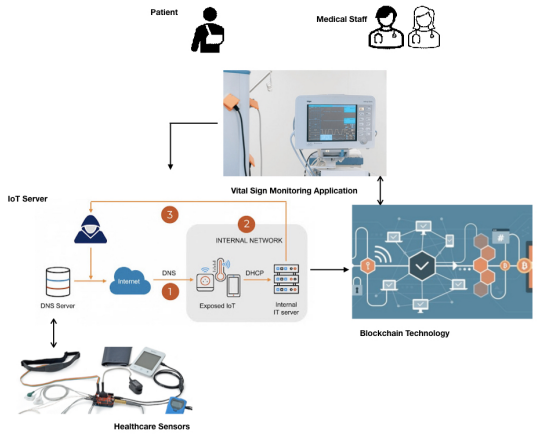


Figure 9: A Typical IoT-Blockchain based smart healthcare system.

used in architecture and the optimizer function, learning rate and iterations for diabetic retinopathy classification. To analyze the performance of different CNN variants, a newly formed dataset is been used. The empirical evaluation suggests that the max-min pooling with hyperbolic tan activation when trained with Adam optimizer with 0.0001 learning rate for 1500 iteration gives the best accuracy of 86.50% among all the tested combination. This study will help the future researchers to get ideas about the various parameters and their effect on the model accordingly to choose best possible combination for an architecture to be designed. Here, to perform the experiment a two-branch CNN model is proposed for the classification of fundus image for diabetic retinopathy and the selected parameters are used. For each class the average sensitivity, specificity, precision, recall, F1-score and AUC are achieved as 0.96, 0.99, 0.96, 0.99, 0.96 and 0.97 respectively from the proposed approach which shows the outperformance of the model.

Data Availability

The fundus images used in this study are available in the following data repository: https://github.com/Manaswinijena/Fundus-images/blob/master/REHLC Dataset.rar

Funding Statement

No funding was provided for this research.

Conflict of interest

The authors do not have any conflicts of interest.

Acknowledgment

We thank Dr. B.N.R. Subudhi (Retd. Professor and HOD of Dept. of Opthalmology, M.K.C.G. Medical College, Berhampur), Dr. Praveen Subudhi and Technician Bhabani Sankar Rath of Ruby Eye Hospital & Lasik Centre (Berhampur) for their precious time to provide the fundus images data for the experiment.

This work is supported by the Ministry of Science and Higher Education of the Russian Federation (Government Order FENU-2020-0022).

References

[1] Kumar S., Tiwari P. & Zymbler, M., “Internet of Things is a revolutionary approach for future technology enhancement: a review”, *J Big Data, Vol.* **6, pp.** 111, 2019.

[2] Mansour R.F., El-Amraoui A., Nouaouri I., Díaz V.G., Gupta D., Kumar S., "Artificial Intelligence and Internet of Things Enabled Disease Diagnosis Model for Smart Healthcare Systems",*IEEE Access,* vol. 9, pp. 45137-45146, 2021.

[3] Sabanayagam C., Yip W., Ting D.S., Tan G., Wong T.Y., “Ten emerging trends in the epidemiology of diabetic retinopathy”, *Ophthalmic epidemiology*, Vol. *23*(4), pp.209-222, 2016.

[4] Piyasena M.M.P.N., Yip J.L., MacLeod D., Kim M., Gudlavalleti V.S.M., “Diagnostic test accuracy of diabetic retinopathy screening by physician graders using a hand-held non-mydriatic retinal camera at a tertiary level medical clinic”, *BMC ophthalmology*, Vol.*19*(1), p.89, 2019.

[5] Paoletti M.E., Haut J.M., Plaza J., Plaza A., “A new deep convolutional neural network for fast hyperspectral image classification”, *ISPRS journal of photogrammetry and remote sensing*, Vol.*145*, pp.120-147, 2018.

[6] Lee H., Kwon H., “Going deeper with contextual CNN for hyperspectral image classification”, *IEEE Transactions on Image Processing*, Vol.*26* (10), pp.4843-4855, 2017.

[7] W. Min, et al "Cross-Platform Multi-Modal Topic Modeling for Personalized Inter-Platform Recommendation," in *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1787-1801, Oct. 2015

[8] Jena M., Mishra S. P., Mishra D., "A survey on applications of machine learning techniques for medical imagesegmentation", International Journal of Engineering & Technology, [S.l], vol.7 (4), pp.4489-4495, nov. 2018. ISSN 2227-524X.

[9] He K., Zhang X., Ren S., Sun, J., "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.770-778, 2016.

[10] Pandey S.K., Sharma, V., "World diabetes day 2018: battling the emerging epidemic of diabetic retinopathy", *Indian journal of ophthalmology*, Vol. *66*(11), p.1652, 2018.

[11] Xu K., Feng D., Mi H., "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image", *Molecules*, Vol. *22*(12), p.2054, 2017.

[12] Otálora S., Perdomo O., González F., Müller H., "Training deep convolutional neural networks with active learning for exudate classification in eye fundus images", *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis,* pp.146-154, 2017.

[13] Quellec G., Charrière K., Boudi Y., Cochener B.,Lamard M., "Deep image mining for diabetic retinopathy screening", *Medical image analysis*, Vol.39, pp.178-193, 2017.

[14] Porwal P., Pachade S., Kamble R., Kokare M., Deshmukh G., Sahasrabuddhe V. Meriaudeau F., "Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research". *Data*, Vol.*3* (3), p.25, 2018.

[15] Jena M., Mishra S. P., Mishra D. "Detection of Diabetic Retinopathy Images Using a Fully Convolutional Neural Network", *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, pp. 523-527, September 2018.

[16] Jena M., Mishra S. P., Mishra D., "A Fully Convolutional Neural Network for Recognition of Diabetic Retinopathy in Fundus Images", Recent Patents on Computer Science, Vol.12 (1), https://doi.org/10.2174/2213275912666190628124008.,

[17] Liskowski P., Krzysztof K., "Segmenting Retinal Blood Vessels with Deep Neural Networks", *IEEE transactions on medical imaging,* Vol.35 (11), pp.2369-2380, 2016.

[18] Picek S., Samiotis I.P., Kim J., Heuser A., Bhasin S., Legay, A., "December. On the performance of convolutional neural networks for side-channel analysis", *International Conference on Security, Privacy, and Applied Cryptography Engineering*, pp. 157-176, 2018.

[19] Yang, X. et al. , "Deep Relative Attributes," IEEE Transactions on Multimedia, vol. 18, no. 9, pp. 1832-1842, Sept. 2016.

[20] Basirat M., Roth P.M., "The quest for the golden activation function". *arXiv preprint arXiv:1808.00783, 2018*.

[21] Pedamonti D., "Comparison of non-linear activation functions for deep neural networks on MNIST classification task", *arXiv preprint arXiv: 1804.02763, 2018*.

[22] Nwankpa C., Ijomah W., Gachagan A., Marshall S., "Activation functions: Comparison of trends in practice and research for deep learning", *arXiv preprint arXiv: 1811.03378, 2018*.

[23] Riegler G., Osman Ulusoy A., Geiger A., "Learning deep 3d representations at high resolutions", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3577-3586, 2017.

[24] Fu H., Gong M., Wang C., Batmanghelich K., Tao D., "Deep ordinal regression network for monocular depth estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002-2011, 2018.

[25] Jena M., Mishra S.P., Mishra D., "Empirical Analysis of Activation Functions and Pooling Layers in CNN for Classification of Diabetic Retinopathy", *International Conference on Applied Machine Learning,* pp.34-39, 2019.

[26] Smith L.N., "A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay", *arXiv preprint arXiv:1803.09820, 2018*.

[27] Dauphin Y., De Vries H., Bengio Y., "Equilibrated adaptive learning rates for non-convex optimization", *Advances in neural information processing systems*, pp. 1504-1512, 2015.

[28] Ranganath R., Wang C., David B., Xing E., "An adaptive learning rate for stochastic variational inference", *International Conference on Machine Learning*, pp. 298-306, 2013.

[29] Zhao H., Fuxian L., Han Z., Zhibing L., "Research on a learning rate with energy index in deep learning", *Neural Networks, Vol.*110, pp.225-231, 2019.

[30] Sharma A., "Guided stochastic gradient descent algorithm for inconsistent datasets", *Applied Soft Computing, Vol.*73, pp.1068-1080, 2018.

[31] Kingma D. P., Jimmy B., "Adam: A method for stochastic optimization", *arXiv preprint arXiv: 1412.6980,* 2014.

[32] Zeiler M. D., "ADADELTA: an adaptive learning rate method", *arXiv preprint arXiv: 1212.5701,2012*.

[33] Lam C., Yi D., Guo M., Lindsey T., "Automated detection of diabetic retinopathy using deep learning", *AMIA Summits on Translational Science Proceedings*, pp.147-155, *2018*.

[34] Xu K., Feng D., Mi H., "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image", *Molecules*, Vol. *22*(12), pp.2054, 2017.

[35] Gharaibeh N., Al-Hazaimeh O.M., Al-Naami B., Nahar K.M., "An effective image processing method for detection of diabetic retinopathy diseases from retinal fundus images", *International Journal of Signal and Imaging Systems Engineering*, Vol.*11*(4), pp.206-216, 2018.

[36] Lin G.M., Chen M.J., Yeh C.H., Lin Y.Y., Kuo H.Y., Lin M.H., Chen M.C., Lin S.D., Gao Y., Ran A., Cheung C.Y., "Transforming retinal photographs to entropy images in deep learning to improve automated detection for diabetic retinopathy", *Journal of ophthalmology*, *2018*.

[37] Adem K., "Exudate detection for diabetic retinopathy with circular Hough transformation and convolutional neural networks", *Expert Systems with Applications*, Vol.*114*, pp.289-295, 2018.

[38] Li Y.H., Ye N.N., Chen, S.J., Chung, Y.C., Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network. Mobile Information Systems, 2019.

[39] Darshit D., Shenoy A., Sidhpura D., Ghapure P., "Diabetic retinopathy detection using deep convolutional neural networks", *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pp.261-266, 2016.

[40] Nikhil M., Rose A., "Diabetic retinopathy stage classification using CNN", *International Research Journal of Engineering and Technology (IRJET)*, Vol.*6*, p.5969, 2019.

[41] Jaiswal A.K., Tiwari P., Kumar S., Al-Rakhami M.S., Alrashoud M., Ghoneim A., "Deep Learning-Based Smart IoT Health System for Blindness Detection Using Retina Images", *IEEE Access, Vol.* **9, pp.** 70606-70615, 2021.

[42] L. Hu, et al, "Software defined healthcare networks," IEEE Wireless Communications, Vol. 22(6), pp. 67-75, December 2015.

[43] Pratt H., Coenen F., Broadbent D. M., Harding S.P., Zheng Y., "Convolutional neural networks for diabetic retinopathy", *Procedia Computer Science*, Vol. *90, pp.* 200-205, 2016.

[44] MM Althobaiti, KPM Kumar, D Gupta, S Kumar, RF Mansour., An Intelligent Cognitive Computing based Intrusion Detection for Industrial Cyber-Physical Systems,Measurement 186, 110145, 2021.

[45] A Tandon, A Dhir, A.K.M. Najmul Islam, M Mäntymäki, Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda, Computers in Industry, Volume 122,103290, 2020

[46] MI Ali, S Kaur, A Khamparia, D Gupta, S Kumar, A Khanna, F Al-Turjman, "Security challenges and cyber forensic ecosystem in IOT driven BYOD environment", IEEE Access 8, 172770-172782, 2020.

[47] Wang, H., "IoT based clinical sensor data management and transfer using blockchain technology", Journal of ISMAC, Vol. 2(03), pp.154-159, 2020.

[48] Poux, F. and Billen, R., 2019. Voxel-based 3D point cloud semantic segmentation: unsupervised geometric and relationship featuring vs deep learning methods. ISPRS International Journal of Geo-Information, 8(5), p.213.

[49] Bengio Y., Courville A.C., Vincent P., "Representation Learning: A Review and New Perspectives", IEEE Trans.Pattern Anal. Mach. Intell., Vol. 35, pp.1798–1828, 2013.

[50] Li B., Lu, H., Zhang H., Tan S., Ji Z., "A multi-branch convolutional neural network for detecting double JPEG compression", arXiv preprint arXiv:1710.05477, 2017.

[51] Feng J., Wu X., Shang R., Sui C., Li J., Jiao L., Zhang X., "Attention multibranch convolutional neural network for hyperspectral image classification based on adaptive region search", IEEE Transactions on Geoscience and Remote Sensing, 2020.

[52] Gao H., Yang Y., Lei S., Li C., Zhou H., Qu X., "Multi-branch fusion network for hyperspectral image classification", Knowledge-Based Systems, Vol. 167, pp.11-25, 2019.

[53] Ge Z., Cao G., Li X., Fu, P., "Hyperspectral Image Classification Method Based on 2D–3D CNN and Multibranch Feature Fusion", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol.13, pp.5776-5788, 2020.

[54] Sahu S., Singh A.K., Ghrera S.P., Elhoseny M., "An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE", Optics & Laser Technology, Vol.110, pp.87-98, 2019.

[55] Ruby Eye Hospital & Lasik Centre (REHLC Dataset). URL: https://github.com/Manaswinijena/Fundus-images/blob/master/REHLC Dataset.rar

[56] Raman R., Srinivasan S., Virmani S., Sivaprasad S., Rao C., Rajalakshmi R., "Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. *Eye*, Vol. *33*(1), pp.97-109, 2019.

[57] Hemanth D.J., Deperlioglu O., Kose, U., "An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network", *Neural Computing and Applications*, Vol.*32* (3), pp.707-721, 2020.

[58] Pawara, P., Okafor, E., Schomaker, L., & Wiering, M. Data augmentation for plant classification. In International Conference on Advanced Concepts for Intelligent Vision Systems (2017, September) 615-626.

[59] Salamon J., Bello, J. P., "Deep convolutional neural networks and data augmentation for environmental sound classification", *IEEE Signal Processing Letters*, Vol. 24(3), pp. 279-283, 2017.

[60] Wong H.B., Lim G.H., "Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV", *Proceedings of Singapore healthcare*, Vol. *20*(4), pp.316-318, 2011.

[61] Zhang D., Wang J., Zhao X., "September. Estimating the uncertainty of average F1 scores", *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pp. 317-320, 2015.

[62] Hamel P., Lemieux S., Bengio Y., Eck D., "October. Temporal Pooling and Multiscale Learning for Automatic Annotation and Ranking of Music Audio", *ISMIR,* pp. 729-734, 2011.

[63] Simonyan K., Zisserman A., "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[64] J. A. Hanley, B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve", Radiology, vol. 143(1), pp. 29-36, 1982.

[65] A. Dalyac M. Shanahan, J. Kelly, "Tackling Class Imbalance With Deep Convolutional Neural Networks", pp. 30-35, 2014.

[66] Qummar S., Khan F.G., Shah S., Khan A., Shamshirband S., Rehman Z.U., Khan I.A., Jadoon W., "A deep learning ensemble approach for diabetic retinopathy detection", IEEE Access, Vol. 7, pp.150530-150539, 2019.

[67] M. A. Rahman et al., "Blockchain-Based Mobile Edge Computing Framework for Secure Therapy Applications," *IEEE Access,* vol. 6, pp. 72469-72478, 2018.

Tele-Collaboration System in CVLab

Djamila Mechta[1*], Saad Harous[2] and Mahieddine Djoudi[3]
E-mail: mechtadjamila@univ-setif.dz, harous@sharjah.ac.ae, mahieddine.djoudi@univ-poitiers.fr
[1]Department of Computer Science, LRSD Lab, College of Sciences, University of Ferhat ABBAS-Sétif-1, Algeria
[2]Department of Computer Science, College of Computing and Informatics, University of Sharjah, Sharjah, UAE
[3]Laboratoire Techne-UR-20297, 1 rue Raymond Cantel Bâtiment A3 TSA 11102 86073 Poitiers Cedex 9, France

Keywords: virtual lab, tele-pw, tele-immersion, tele-collaborative learning, agent approach

Received: June 19, 2019

This work is part of a learning environment that has virtual laboratories that are designed for distant practical work (Tele-PW). In these environments, Tele-PW is performed in two modes: individual and/or collaborative. In this paper, we are concentrating on the tele-collaborative distant practical work model. The work, presented in this paper, proposes an artificial agent called Synchronizer Coordinator Agent (SCA) to synchronize and coordinate the activities of a cognitive process in order to build and maintain a shared conception of a distant practical work between a set of learners. This agent provides certain features such as managing groups of learners, coordinating tasks, shared workspace among members of the Working Group. It is also responsible for the synchronization of workspace agents when they want to manipulate shared virtual objects simultaneously. We have chosen Petri nets to illustrate the principle of granting access to shared objects in the case of simultaneous requests. Experimental results show the effectiveness, of the artificial agent within any tele-collaborative/tele-cooperative learning situation. Several situations describe the geographical/time dispersion of learners and tutors in our system are considered during the system design phase.

Povzetek: V članku je opisan sistem za telekomunikacijsko sodelovanje pri virtualnih laboratorijih.

1 Introduction

The main purpose of practical work (PW) is to compare the knowledge of learners to reality. In this work, we bring an extra dimension: the distance, by focusing on practical remote education. The qualitative "experimental" is applied to training situations where the learner manipulates and interacts with virtual objects and compares the results to theoretical models.

Virtualization [1] associated with remote access provides a relevant solution to the simulation problem of practical work. Not only it allows to reconstruct classical (face to face) teaching situations (traditional rooms for PW), but also to exceed the limitation related to these situations (ex: the large number of students in the lab area with COVID-19 pandemic) [2], and even to consider some uses specific to the exploitation of computers completely paperless, anytime and anywhere.

The environments where we performed the lab experiments are either virtual known as virtual labs [3, 4]; where the objects manipulated in lab activities are totally virtual (virtual resources and remote/local access), or remote labs [5]; where the learners manipulate remotely real devices (Real resources and remote access) or hands-on labs known as traditional labs (Real resources and local access). The three mentioned types of labs have described in [6, 7, 8]. In the review [9], authors presented both virtual and remote labs for different disciplines such as VISIR lab in engineering [10], programming robots [11] and biology [12]. In [13], authors presented an exploratory analysis of two remote labs that are photovoltaic panels and electric machines labs to measure their acceptance, in terms of usability and usefulness, by students in higher education. A virtual biomaterials lab [14] has been proposed during COVID-19 Pandemic to overcome the difficulties when applying the previous models. A virtual pathology lab [15] brings reality, efficiency, and opportunities for collaborative learning that pathology education did not widely used before. An approach that uses virtual and remote labs in Mechatronics Education based on Cloud Services is proposed in [16].

For the three labs categories cited previously, collaboration/cooperation constitutes a key element for conducting lab experiments efficiently.

Tele-collaborative distance Practical Works (tele-collaborative Tele-PW) [17,18,19] are a form of teaching that requires group work where the learner performs tasks, expresses ideas, shares tools, communicates with other learners etc. The evolution of technologies created new forms of collaborative work with their own specificities. Two learners who work together to achieve remote labs share different information about their common goal. But exchanging of information remotely may quickly become laborious and prevent them from accomplishing their

* Corresponding author

activities. Implementing this type of Tele-PW adds new challenges: organizational, human, and of course techniques. As examples, we cited the management of concurrency: Simultaneous execution of transactions over a shared databases or tools between learners can create several data integrity and consistency problems, communication between learners and teachers. In this case, learners/tutors will be confronted with a new work environment. Therefore, personalized solutions must be proposed to respond to their specific needs in terms of tele-collaborative work.

Many scientific productions present the socio-constructivist theories that the learning process is a collaborative and constructive process with directions how to overcome the eventual lack of support of those features in virtual and remote labs as deploying those laboratories into education management systems, supporting the labs to be operated by multiple users at one time, embedding the labs into virtual, etc. [20]

In our previous works, we implemented CVL@b: an agent-based platform of Virtual Laboratory [21]. This platform is a distributed digital environment that contains a space for editing documents describing Tele-PW, an electronic library and a space where the experiment is performed. It enables learners to achieve their distant practical work and teachers to assist the learners during their work sessions. However, the activities carried out in this environment are individual.

To remedy this individual learning situation and the challenges mentioned above, a centralized original solution represented by the creation and integration of the SCA agent in our system will be the aim of this article. This solution enables the synchronization of the Tele-PW that is the synchronization of the concurrent access to virtual objects, which are considered as critical resources. The objective of this work is to show clearly these agents' roles, their functionalities and interactions with the set of existing agents in CVL@b.

The remainder of the paper is organized as follows. In section 2, we describe the collaboration models used. Section 3 presents the context of our work. In section four, we tackle tele-collaborative tele-PW challenges. Our contributions are detailed in Sections 5, 6, and 7. In Sections 8 and 9, we present the implementation. In section 10, we evaluate the performance of the proposed VL plateform CVL@b. Finally, we conclude the paper.

2 Collaboration models

The proposed classification is based on the attributes identified in [22,23]. The attributes that can be used to characterize the collaborative work are the division of tasks, the dimension space/time of interactions. In addition, we considered the number of actors involved in the collaborative work to be important so we include it as an additional criterion of this classification. Indeed, there is a distinction between collaborative situations involving a small group of participants (two to four participants) and those involving a larger number of actors (example projects).

2.1 Division of work

When discussing the collaboration in terms of the division of tasks, there are two forms:

2.1.1 Collective (collaboration)

Learners share the same task. The tasks of each learner in this case are interdependent and learners' roles are interchangeable. Indeed, tasks allocation is dynamic (managed in real time based on the current context). This can be explained by the fact that the collaborative work is dynamic. It is difficult to distinguish the contribution of each participant in the final result.

2.1.2 Distributed (collaboration)

Learners are semi-autonomous. They divide tasks among themselves and work separately. The sub-tasks must be independent of each other. We can distinguish the contribution of each of participants to the final result. In addition, the tasks allocation, in this case, is often static (prior to the action). During the execution of tasks, participants are not necessarily aware of the existence of their partners nor informed of their current activities.

2.2 Space and time dimension of interactions

2.2.1 Space distribution of collaborators

It depends on the location of each participant in the collaborative activity. Two possible forms can be distinguished: face to face situation where group members are in the same location. In this case learners can interact freely. The remote locations (or distributed) where learners of the group are distributed in remote locations, whether located within the same country or spread across continents.

2.2.2 Time distribution of collaborators

This situation is based not only on the activities of the group but also the data transmission techniques (synchronous or asynchronous transmission). This leads to two types of collaboration: real-time collaboration and deferred time collaboration.

In the first form, the groups' activities are performed in real time. Any action taken by a member of the group is immediately sent to other members. It is then a synchronous collaboration. While in the second form, the group members act in deferred time. The actions are spread over the time scale and any member of the group who was not present when the actions were performed; he/she will see the end result of actions performed before his/her arrival. It is then necessary to maintain the system status at any time. Notifications are sent to users to highlight the changes made between two visits made of the same user. In this case we should be to tell who is currently connected. This is called asynchronous cooperation.

2.3 Actors number

Obviously in collaborative practical works, the number of learners is limited. The interactions within a group are affected by its size. This influence concerns in particular the group's performance, decision making and communication. The increase of the group size makes the communication more difficult (less interactive) and can thus affect the working relationship that has been established between the group members.

3 Context

CVL@b is an environment for distance teaching and learning, especially to help learners performing their Tele-PW. It offers on one hand, learners with the opportunity to register, consult the description of practical work, perform the practical work etc., and on the other hand, it offers teachers/tutors a space for the management of educational content, monitoring and evaluation of learners.

CVL@b is structured based on a three levels architecture (GUI, CVL@b kernel and storage space). The kernel is composed of two main systems CVL@b-LCMS (Learning Content Management System) and CVL@b-LMS (Learning Management System). CVL@b-LCMS is designed around a learning content management (LCMS-PWS) and documents management literature embodied in an electronic library (LCMS-EL).

CVL@b-LMS is responsible for the management of learning process through a set of sub-systems: 3D environment for practical work, supervision & control system [21], evaluation system and tele-collaboration system. In this paper, we concentrate on CVL@b-LMS and more precisely on the tele-Collaboration system (LMS-SCA).

4 Tele-collaborative tele-PW teaching problems

For a long time, theories of collaborative learning have focused on the question of how the individual works within a group. Cognition was seen as a product of individual processes and the context of social interaction was seen as a simple background. Recently, the group itself has become the unit of analysis and the perspective of research has shifted to the properties of interaction emerging in the social context [24]. In terms of empirical research, initially the goal was to determine whether collaborative learning was more effective than individual learning and under what conditions. Researchers controlled several independent variables (group size, group composition, task nature, communication media, etc.) that interact with each other in a way it was almost impossible to establish causality links between conditions and effects of tele-collaboration [25,26].

In a conventional PW, learners engage each other in a coordinated effort to solve a practical problem together; they exchange knowledge about the PW and perform the required tasks by PW by sharing the necessary tools for the experiment. So, a space for a collaborative Tele-PW is a digital environment designed to facilitate information exchange, discussion and coordination within a group and the sharing of resources being manipulated. This work focuses on developing a space for tele-collaborative Tele-PW where nature of the tasks accomplished and their complexity differ from distant courses, distant supervised work, etc. This difference requires a specific mode of tele-collaboration, which is the objective of this work.

In our previous work, we proposed an agent-based architecture for a virtual laboratory for the realization of Tele-PW. This architecture is very well suited for individual Tele-PW, however, it has some drawbacks in the case where groups want to cooperate and when the same group of learners wants to work in a collaborative manner. This is a situation where learners share the same workspace. Then, the problem is how to ensure mutual exclusion in accessing shared virtual objects. Also, how to coordinate tasks between learners?

5 Contribution

We had to come up with a new architecture that will solve the problem of simultaneous request of virtual objects and tasks coordination. A new agent called SCA is added whenever a working group is created. This agent is always active as long as there is a learner from any group connected. Its role is to:

- Encapsulate the tasks of group management and coordinate the group implementation of the experiment (each group has its own agent SCA).
- Control the sharing of the work area.
- Allocate and destruct virtual objects.
- Synchronize the access to resources or shared virtual objects in the case of collaborative and/or cooperative Tele-PW.
- Update the overall state of the common workspace.
- Disseminate the overall state to different learners' workspaces.
- The presence of SCA in this architecture has led to several changes:
  1. Implementation of Tele-PW in a collaborative fashion and reduction of the server agent use, that agent is now used only for the extraction of different databases.
  2. When the distant PW is group based, SCA is involved to ensure the synchronization of workspaces agents seeking virtual objects and/or other critical resources. It will make a decision to deny or approve the request.
- The procedure of real-time monitoring has been changed. In the original approach, the supervisor agent communicates directly with the sensor agent to start capturing screen shot of the workspace. There was only one learner and thus a single sensor, but now it is not the case anymore because Tele-PW is group-based work and therefore the number of learners is more than one, which means several sensors (there are as many sensors as the number of members of the group). Therefore, in this case the SCA finds the concerned sensor agent.

- Any change of state of the shared environment of the experiment will be broadcasted by the SCA to the various learners of the same group.

6 SCA structural model

SCA is divided into a set of modules, each module provides some functions: group management, tasks coordination, workspace sharing, virtual objects allocation and synchronization of agents' workspace, update the global state of overall common workspace and its dissemination.

6.1 Group management

When a learner registers, SCA is responsible for assigning it to a group:

- If the learner chooses voluntarily a group with the consensus of other members of the group, then SCA agent only validates this option.
- Otherwise, the choice of the group is systematic.

When a learner connects to the CVL@b, the SCA agent adds her/him to the list of online learners. It then proceeds to announce its presence to the tutors to enable them to supervise the groups of learners and to the members of connected groups so that the collaborative aspect is concretized. The same process is executed if a tutor is connected; an information message is broadcasted by SCA to the connected learners so that they can benefit from her/his assistance.

6.2 Update and dissemination of the workspace global state of the same group

SCA receives changes made to the virtual objects in different workspaces. Then, it saves them (update the global state of the graphical environment) to maintain a global state for workspaces of the group members who decided to share their workspace. After that, it informs the other workspaces of the changes that have happened to the object, then informs the learners who connected and who are interested in sharing of this state. For more details, see section 6.3 (Figures 5 and 6).

6.3 Workspace sharing

A workspace is a graphical environment where experiment takes place. It contains a set of educational elements called virtual objects manipulated by groups of learners during a session of Tele-PW. During this session, members of the same group can collaborate so they share their workspace. This sharing involves the simultaneous manipulation of virtual objects contained in the common workspace. As a result, we are faced with a situation that raises issues related to the synchronization problem of workspace agents, which access concurrently different resources (selection of the same object with the mouse, moving an object and updating of an object). Therefore, one of the main tasks of SCA is to manage and control concurrent access to different virtual objects. The next section describes in detailed how does the SCA agent do the allocation and workspace agents synchronization.

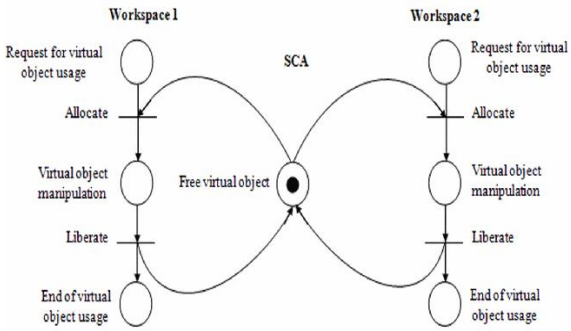


Figure 1: Petri Net for Allocating Common Object.

6.3.1 Design of resources allocation using Petri nets

In the context of concurrent programming, resource allocation is the process of allocating resources to a particular process. This operation is necessary to ensure appropriate access to shared resources between multiple processes. Such an operation is not required for non-shared resources. In our situation, we opted for a formal model for modelling the generic problem of allocating virtual objects between different workspace agents when requesting to manipulate these resources simultaneously. This model is the petri net, which is a directed bipartite graph where tokens constitute the marking.

The Petri net (Figure 1) shows the principle of allocating virtual objects in our system. When learners want to manipulate a common virtual object at the same time, they send a request to SCA agent. This agent acts as an allocator of resources where the token represents the shared virtual object. It allows the use of a common virtual object according to the FIFO strategy: the first request that arrives is the first one served. The mutual exclusion property is guaranteed by the exclusive use of the virtual object. When manipulating a shared virtual object, any modification of its properties will be sent to all other workspaces by the SCA agent.

6.3.2 Virtual objects matrix

It is a static structure of type two-dimensional array whose rows represent the various learners' workspaces of the same group and the columns represent the virtual objects created in each workspace and their properties (objects type, object coordinates on the interface workspace, busy or not, etc.). In addition, this matrix represents the graphical state of the learner's workspace who have started a training session of type Tele-PW. The agent SCA supplies the matrix of virtual objects.

6.3.3 Learning situations

1st case: individual Tele-PW: Figure 2 represents a learning situation in which each learner works locally on her/his own workspace.

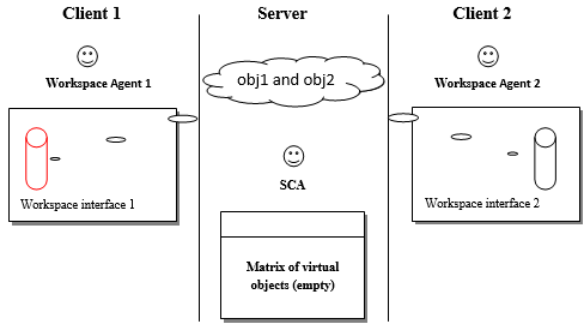


Figure 2: Individual Learning Model.

Workspace 1 and workspace 2 respectively contain the virtual objects obj1 and obj2. In this case, the SCA agent does not intervene, and therefore the matrix of virtual objects is empty.

2nd case: collaborative Tele-PW: Each learner has his\/her own virtual objects in his\/her workspace. If a learner i ($Workspace_i$) wants to share his workspace, he sent his\/her request to the SCA agent who is responsible for updating the matrix of virtual objects and disseminates the new state to the concerned learners (Figure 3).

In this case, the virtual objects will become common so synchronization is needed to ensure they are used properly (mutual exclusion). The SCA agent via its matrix of objects can resolve any problem accessing the same object simultaneously.

This matrix contains a column that shows the state of the object (free or busy) and this column is filled based on the result of the Petri net that we discussed earlier. For the implementation of shared workspace and mutual exclusion, the following algorithms in Figure 4 are used.

6.3.4 Dynamic aspect of SCA

When the workspace agent receives a request to manipulate a virtual object from the learner interface agent, it locates the SCA agent of the concerned to which belongs this learner because each group has its own SCA agent. After, it sends to the SCA agent a message to allocate the desired virtual object and waits for its reply. SCA makes a decision to allocate or not the desired object based on the content of the matrix of virtual objects. If the desired object is busy, the workspace agent receives a negative response and goes in the final state; otherwise,

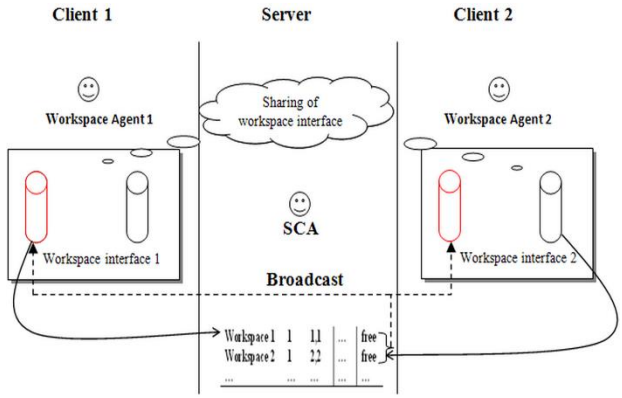


Figure 3: Collaborative Learning Model.

```
get operation()
if (operation = Allocate) then{
  Extract the virtual object if it is free
      state=occupied
  //Authorization to access to this VO
  Send msg to workspace for authorize his use (update)
}
else (operation = liberate) then{
      state= free
  Broadcast msg to workspace agents for information
  that the VO is free and available }
```

Objects synchronization

```
recive msg
if (msg not null){
 Extraction of VO vectors of learner workspace also the
 learner login.
  if (the learner workspace is not empty){
      Add virtual objects to group workspace
      Broadcast the update (the new matrix) to online
      learners'group }
 Send the workspace group to learner
}
```

Shared workspace

Figure 4: Sharing workspace and achieving mutual exclusion.

(when the object requested is available) the SCA agent changes its status in order to lock it and a notification message (positive answer) is sent to the requesting workspace agent. This allows the learner to manipulate the object in its interface (see Figure 5).

During the manipulation of virtual objects, their status will change. These changes are sent by the workspace agent to SCA agent to update the states of the manipulated virtual objects in the matrix of resource allocation. Also, it gets the list of workspaces for different learners in the same group and broadcasts to them the latest updates of the global workspace. Requests for release of virtual objects will be forwarded to the SCA who will execute them (SCA is responsible of adding/releasing virtual objects informs experiment workspace) as shown in Figure 6.

6.4 Coordination task

This task is implicitly the responsibility of the SCA agent throughout its life cycle. It coordinates the various tasks assigned to the artificial agents present in a session of activities related to a Tele-PW.

7 Tele-collaboration/tele-cooperation tools

In tele-Collaboration environment [27], learners meet virtually, is the place where they talk. This is where the group is crystallized and be allowed to exist and live. Tele-Collaboration is essentially made of interaction and dialogue. Learners/tutors can share a digital workspace where they express their ideas, communicate and exchange resources (to share costly equipment and resources, which are otherwise available to limited number of users due to constraints on time and geographical distances). In this space, learners

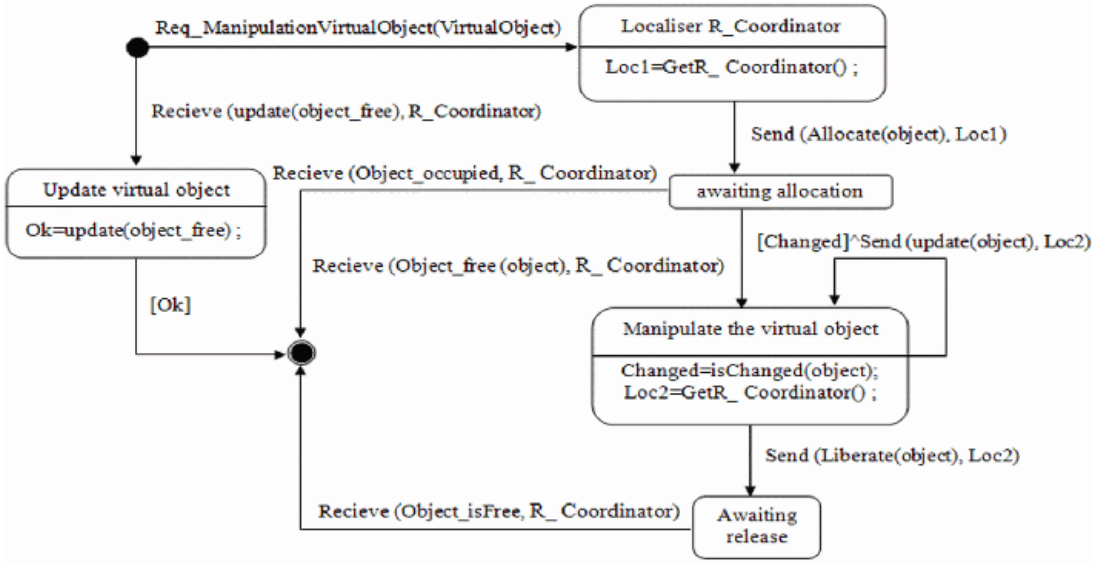


Figure 5: State-transition of Workspace.

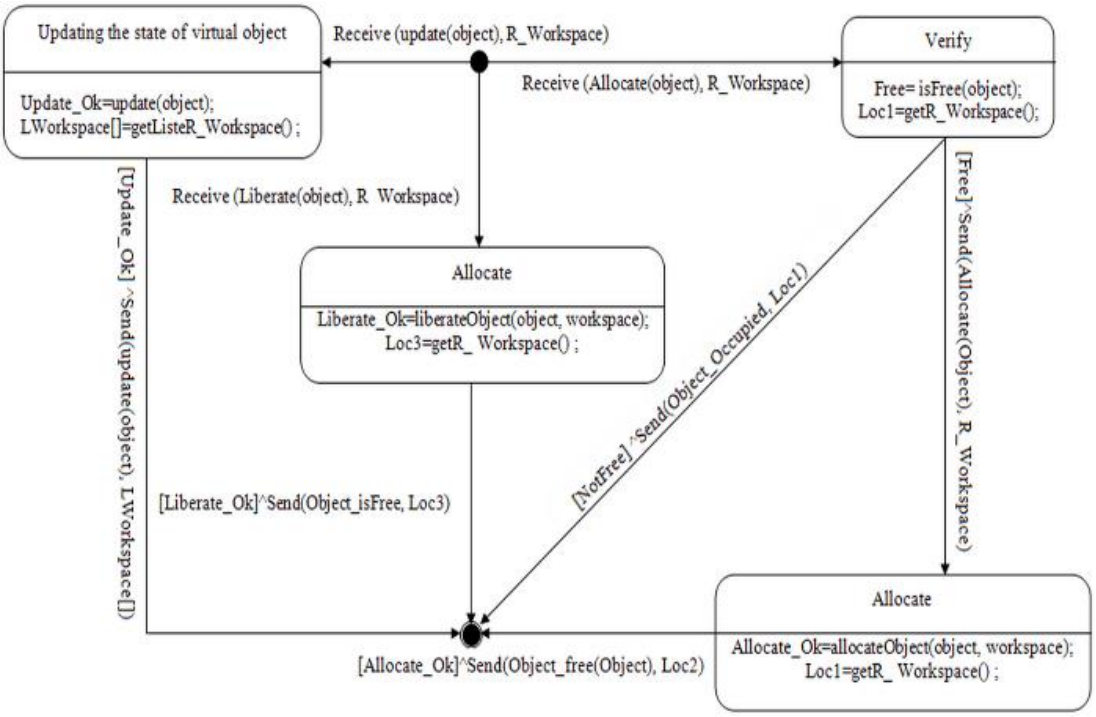


Figure 6: State-transition of SCA.

communicate synchronously or asynchronously, in written or spoken natural language. The communication space includes tools and is prepared for this purpose. The tools most commonly used are chats, email, discussion boards, wikies and on-line forums, etc. In CVL@B, a chat tool based on peer-to-peer architecture is available for the learners so that they can interact and work together which make the resources more efficiently.

All members of the group remain in regular contact; each contributes to the group by performing its actions. Everyone can contribute to the actions, being performed another group member, to increase the performance, interactions are permanent. In addition to communication tools, a shared space for creating/editing reports and a shared workspace favouring the Tele-immersion are available for use by various learners. Regarding tutors, dynamic or evolving situations of teaching/learning require from them a certain number of specific activities to build cognitive skills.

The use of ICT in CVL@B offers the tutor means of supervising, controlling Tele-PW activities, and giving effective feedbacks [21]. The tele-cooperative mode, in CVL@B, is provided by individual learners' workspaces who will act independently. The interactions are concerned to the organization, coordination and monitoring of progress (often under the responsibility of a chosen member of the group who is charged with the responsibility to ensure the individual performance of each group member). Everyone's responsibility is limited to ensure that its actions are achieved: the continuous and

coordinated combination of the sub goals, achieved by each member, will lead to the achievement of the expected goal of Tele-PW.

8 Length of agents' life

The agents' life length (Figure 7) in the system varies based on the agent's role. The connector, archivor and server agents are created when the system is launched and remain during the life cycle of the server application.
These agents communicate with another server of type databases to extract and safeguard data. This communication is done through a well-specified protocol (for example: JDBC for Java).

On the other hand, the supervisor and helper agents are created at the time of, respectively, a request of assistance or supervision. Their deletion from the system is carried out as soon as they finish the task that was assigned to them. The learner interface, teacher interface, SCA, and sensor agents are created when the human actor (Learner/Teacher) has successfully connected. They remain active during this session. SCA remains until all active members of the same group are disconnect.

9 Implementation

Software architecture as illustrates Figure 8 shows the number, types, and locations of agents in our system, each human actor has his own artificial agents: learner/teacher interface, workspace, sensor, assistant, and supervisor. They run on the client and are responsible for the management of user interfaces and interaction with other agents located on the application server. Server agent, archivor agent and SCA remain on the server and are responsible for providing the various resources made available by MySQL database server.

The best approach to construct a multi-agent system (MAS) is to use a platform, which represents a set of tools necessary for the construction and the activation of agents.

JADE (Java Agent DEvelopment framework) platform is deployed for the execution of our MAS. The communication between artificial agents is based on the communication protocols inter-agent, which follows the standard model FIPA ACL.

9.1 Sample interface

When a user wants to connect to his/her personal workspace, he/she enters his/her user's name and password. After its identification, his/her personal workspace (tools bag) is loaded based on the type of user (learner or teacher). This space offers the learner the necessary tools from consulting Tele-PW sheet until the completion of the Tele-PW.

In case the user is not registered, he/she has to register so that he/she can access his/her work later and make use of the available features offered by the platform. In order to check the new design of our SMA, an instance of Tele-PW of collaborative/cooperative nature has been implemented.

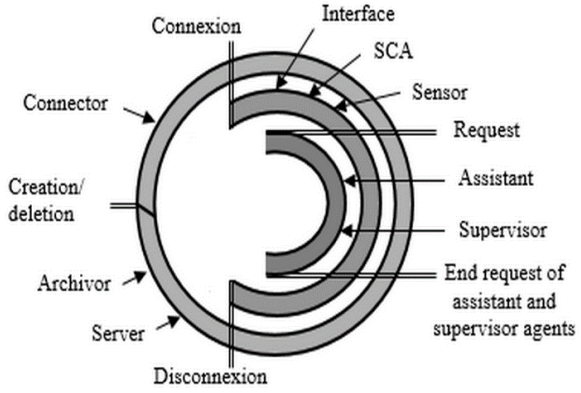


Figure 7: Length of Agents' life.

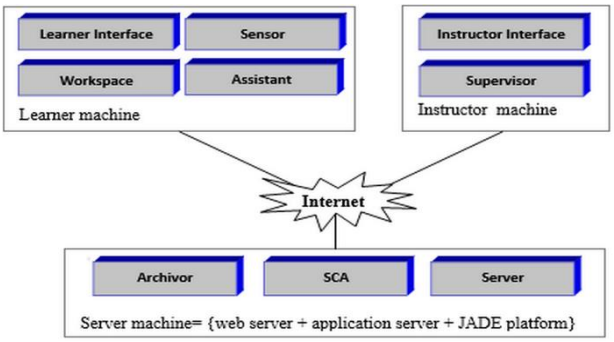


Figure 8: System Logical architecture.

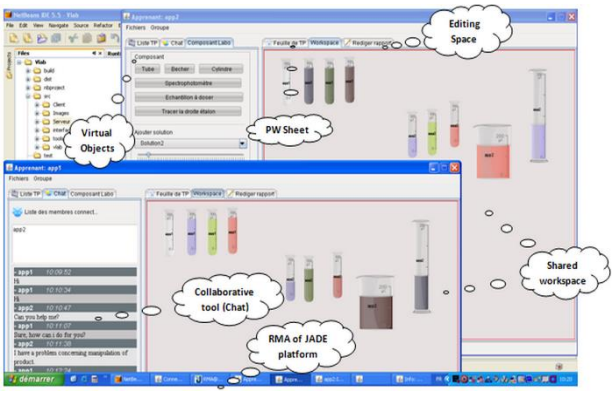


Figure 9: Shared Environment of Tele-PW.

The screen above (Figure 9) represents a distant experiment that takes place in a biochemistry virtual laboratory. Each leaner has a bag of tools.

An interface that contains a shared workspace where the experiment takes place, a panel of virtual objects, a communication space (Chat), list of Tele-PW done earlier and a new one which have been loaded on the system, an editing space, and a display space for the content of Tele-PW documents.

10 Performance evaluation of LMS-SCA

10.1 Parameters to be evaluated

In order to show the effectiveness of LMS-SCA, various parameters were taken into account. In this study, the selected criteria will make it possible to validate the

Situations	Spatial distribution		Temporal distribution	
	Learner/Learner	Tutor/Learner	Learner/Learner	Tutor/Learner
Situation 1	Present	Remotely	Synchronous	Sync/async
Situation 2	Remotely	Present	Sync/async	Sync/async
Situation 3	Remotely	Remotely	Sync/async	Sync/async

Table 1: Pedagogic situations in LMS-SCA.

framework taking into account the quality of training resulting from the use of this system. We describe below the selected criteria:

- Satisfaction: It is important to know if the user is overall satisfied with the use of LMS-SCA.
- Ergonomic: It makes it possible to evaluate the man-machine interaction of a computer system. This criterion takes into account the ease of use, the user-friendliness and the system usefulness. Since LMS-SCA consists of a set of systems, each system will be assessed in an independent way, in addition to the overall assessment.
- Tele-Collaboration/Tele-Cooperation: It is important to measure the level of users' participation.

This factor includes the individual and collective participation in the training. This means that the learner has contributed to the realization of the Tele-PW, participated in the discussion with his/her colleagues about the practical problem and helped them during the activity.

10.2 Experiment Conditions

10.2.1 Population and Groups Structuring

In order to evaluate LMS-SCA, we have used a questionnaire to collect the opinion of 34 participants. Before we start to collect the data, two categories of participants were identified: learners and tutors. The group members and their tutors were distributed in the space. The learners and teachers/tutors gave their impressions about the various functionalities offered by the system.

Each group has between 2 to 3 learners. Tutors have had previous experience on how to form learning groups. Participants (learners/tutors) have never participated previously in a remote experience. The experiment took place in a computer room where all learners were present. The room is equipped with 15 PCs connected by a local area network and an Internet connection. Due to electrical problems, PCs going down, the unavailability of learners at the same time, and each student having a computer, the experience took a long time. To calculate the reliability of the questionnaire, we used the test (Cronbach's Alpha), which gave the value 908 for a population of 34 participants. This value, which showed that there is a harmony between the different items of the questionnaire therefore, the reliability of the result is acceptable.

10.2.2 Pedagogic Situations (tutor versus learner)

According to the dimension space/time of the interactions, there are two distributions: spatial distribution and temporal distribution. Each one contains also two situations. The different educational situations we have considered are summarized in Table 1. These situations allow synchronous and asynchronous tele-collaboration (spatial and temporal dimension), require a limited number of learners per group and the learners work remotely.

10.2.3 Pedagogic Scenarios

The different educational scenarios used in this experimental protocol are summarized in Table 2. From this table, and from the point of view of division of tasks, learners work in tele-collaboration and/or tele-cooperation. But the interaction between tutors and learners is done in a tele-collaborative way only. This is due to the summative/formative evaluation of learners' work.

10.2.4 Pedagogic Content

The choice of the pedagogical scenario for the experiment is not easy at all. This is due to the numerous disciplines that are situated in the experimental sciences. Thus, to choose a discipline, we took the following factors in consideration: degree of interactivity during the activities of Tele-PW and the low level of abstraction of the latter. Biochemistry discipline has met the criteria mentioned in our approach by the nature of its PW which are concrete, do not deal with abstract phenomena and have a very high degree of interactivity. Learners collaborate with each other for the implementation of distant PW tasks. The tele-collaboration/tele-cooperation aspect is one of the main features of distant PW studied. First, learners in the same group cooperate in the implementation of distant PW. Each learner is individually responsible for executing different steps of the affected phase. After the completion of individual tasks, learners, in a collaborative manner, proceed to the combination of the intermediate results to write the distant PW report which will be sent to the tutor for possible evaluation. These distant PWs points of tele-cooperation/tele-cooperation are illustrated in Figure 10.

Scenarios	Learner/Learner	Tutor/Learner
Scenario 1	Tele-collaborative	Tele-collaborative
Scenario 2	Tele-cooperative	

Table 2. Pedagogic scenario in LMS-SCA.

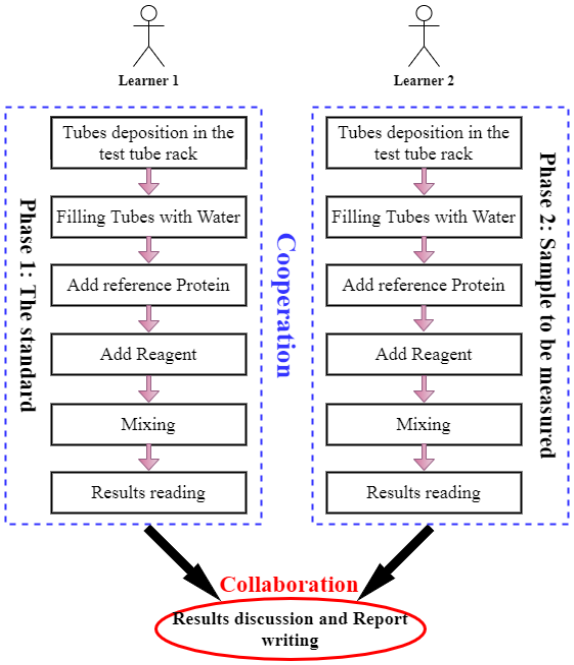


Figure 10: Pedagogic Content.

10.3 Results

To measure each criterion and be able to analyse results, we assumed that the sum of (a lot & complete) gives a positive appreciation and the sum of (not at all & average) gives a negative one. The communication tool used is found to be very useful (71.5 %) and very easy to use by 93 % of the group members and tutors. Most users (76.5 %) find the interface to be friendly. The group that accessed the system remotely believes that the communication delay is acceptable. The results of this evaluation are presented in Table 3.

The statistical results obtained are reported in Figure 11. Analysing the results, for the textual tele-collaboration illustrated by the clarification of the Tele-PW statement (76.3 %), discussion of the results between learners (84.2 %) and drafting of the report (79 %) either Tele-immersive collaboration (68.9 %) by using shared workspace. Thus, in the summary, the level of textual/Tele-immersive collaboration was considered important. In addition, more than 90 % felt the presence of other colleagues using the system that is a good indication that the users of our system did not suffer from loneliness that is a serious problem with online learning.

The space of the experiment was considered very easy to use by 83.44 %, friendly to use by 75.92 % and offers a set of virtual objects that are necessary to the realization of a distant PW as shown in Table 4.

Criteria	Negative		Positive	
	Not at all	Average	A lot	Complete
Easy to use	7	-	26.5	66.5
Utility	2.5	26	21	50.5
Convenient	2.5	21	15.5	61

Table 3. Communication evaluation.

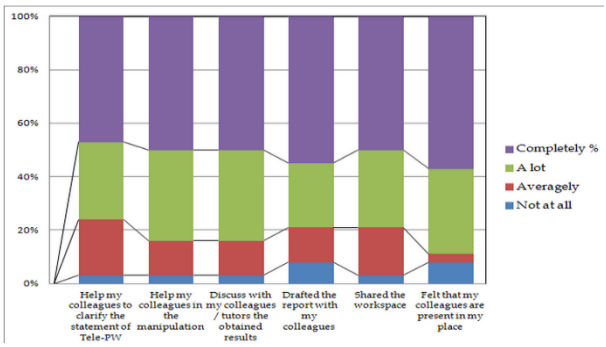


Figure 11: Evaluation of CVL@b tele-collaboration System.

Criteria	Negative		Positive	
	Not at all	Average	A lot	Complete
Easy to use	4.73	11.9-	18.94	64.5
Convenient	4.73	19.34	30.92	45
Provided tools	2.36	19	23.95	54.6

Table 4. Workspace evaluation.

10.4 Discussion

The overall result of the evaluation shows that LMS-SCA is an acceptable tool. The majority participants find it easy to use, useful and convenient. LMS-SCA was well appreciated and it answers the requirements of learners/tutors. The results were very satisfactory and validated that our design was well adapted to the collaborative Tele-PW. The degree of interaction is considered to be important. Almost all participants think LMS-SCA could be used to help learners/tutors in Tele-PW laboratory. But 28.9 % of the participants think that it does not improve the quality of collaboration and does not replace the classical collaboration at all.

11 Conclusion

Our work is in the context of a tele-collaborative learning system for Tele-PW. The tele-collaborative/tele-cooperative aspects relevant to the learning process, presents difficulties in terms of their implementation, the management of concurrent access and shared workspace, and sharing the communication tools among learners etc. In this paper, we presented an approach for synchronization of tele-collaborative Tele-PW. It is an artificial agent SCA, which ensures the smooth running of collaborative activities of a Tele-PW. This agent manages

and controls the various accesses of competing workspace agents to the shared virtual objects and shared virtual workspace while coordinating the tasks performed by the set of agents involved in this process. In order to validate our work, we tested the system by implementing a biochemistry experiment. The results show that the tele-collaboration increases the flexibility of the process of accomplishing a Tele-PW. Nevertheless, in the future, we plan to make our agent more intelligent by developing their skills and make our platform more efficient.

References

[1] Ray, S., and Srivastava, S. (2020). Virtualization of science education: a lesson from the COVID-19 pandemic. Journal of proteins and proteomics,1-4. https://doi.org/10.1007/s42485-020-00038-7.

[2] Shambare, B., and Simuja, C. (2022). A Critical Review of Teaching with Virtual Lab: A Panacea to Challenges of Conducting Practical Experiments in Science Subjects beyond the COVID-19 Pandemic in Rural Schools in South Africa. Journal of Educational Technology Systems, 50(3), pp. 393–408. https://doi.org/10.1177/00472395211058051.

[3] De Jong, T., Linn, M. C., and Zacharia, Z.C. (2013). Physical and virtual laboratories in science and engineering education. *Science*, 6130(340), *pp.305-308*. https://doi.org/10.1126/science.1230579.

[4] Soriano, A., Ponce, P., and Molina, A. (2019). A Novel Design of Virtual Laboratory. 20th International Conference on Research and Education in Mechatronics (REM), IEEE, Wels, Austria, pp. 1-6. https://doi.org/10.1109/rem.2019.8744115.

[5] Post, L. S., Guo, P., Saab, N., and Admiraal, W. (2019). Effects of remote labs on cognitive, behavioral, and affective learning outcomes in higher education, Computers & Education, Volume 140, pp. 103596. https://doi.org/10.1016/j.compedu.2019.103596

[6] Reeves, S.M., and Crippen, K.J. (2020). Virtual Laboratories in Undergraduate Science and Engineering Courses: A Systematic Review, 2009–2019. Journal of Science Education and Technology, volume=30, pp.16-30. https://doi.org/10.1007/s10956-020-09866-0

[7] Hernandez-de-Menéndez, M., Vallejo Guevara, A., and Morales-Menendez, R. (2019). Virtual reality laboratories: a review of experiences. Int J Interactive Design and Manufacturing (IJIDeM), pp.947–966. https://doi.org/10.1007/s12008-019-00558-7.

[8] Ortelt, T.R., Haertel, T., and Frye, S. (2021). Remote Labs in Germany—An Overview About Similarities and Variations. In: Auer M., May D. (eds) Cross Reality and Data Science in Engineering. REV 2020. Advances in Intelligent Systems and Computing, Springer, vol 1231, pp. 143–153. https://doi.org/10.1007/978-3-030-52575-0_11.

[9] Alkhaldi, T., Pranata, I., and Athauda, R. (2016). A review of contemporary virtual and remote laboratory implementations: observations and findings. Journal of Computers in Education, Springer, vol 3, pp.329–351. https://doi.org/10.1007/s40692-016-0068-z.

[10] Marques, M. A., Viegas, M. C., Costa-Lobo, M. C., Fidalgo, A. V., Alves, G. R., Rocha, J. S., et al. (2014). How remote labs impact on course outcomes: Various practices using VISIR. IEEE Transactions on Education, 57(3), pp. 151–159. https://doi.org/10.1109/te.2013.2284156.

[11] Chaos, D., Chacon, J., Lopez-Orozco, J. A., and Dormido, S. (2013). Virtual and remote robotic laboratory using EJS, MATLAB and LabVIEW. Sensors, MPDI, 13(2), pp.2595–2612. https://doi.org/10.3390/s130202595.

[12] Hossain, Z., Jin, X., Bumbacher, E., Chung, A., Koo, S., Shapiro, J. D., et al. (2015). Interactive cloud experimentation for biology: An online education case study. In Proceedings of the 33rd ACM Annual Conference on Human Factors in Computing Systems, Seoul Republic of Korea, pp. 3681–3690. https://doi.org/10.1145/2702123.2702354.

[13] Tirado-Morueta, R., Herrera, R., Márquez, M.A., Borrero, A., and Andujar Marquez, J. (2015). Comparing Remote Laboratories from the Student Perspective. IFAC-Papers on Line, vol 48, pp.176-181. https://doi.org/10.1016/j.ifacol.2015.11.233.

[14] Gerstenhaber, J. A., and Har-el, Ye. (2021). Virtual Biomaterials Lab During COVID-19 Pandemic. Biomedical Engineering Education, Vol. 1, No. 2, pp. 353–358. https://doi.org/10.1007/s43683-020-00045-6.

[15] Erana Rojas, I.E., Barbosa Quintana, A., Pérez Saucedo, J.E., Segura-Azuara, N., and Cabrera, M. (2018). The virtual pathology lab experience. International Journal on Interactive Design and Manufacturing (IJIDeM), vol 12, pp. 1299–1308. https://doi.org/10.1007/s12008-017-0397-9

[16] Vitliemov, P., Bratanov, D., and Marinov, M. (2020). An Approach to Use Virtual and Remote Labs in Mechatronics Education Based on Cloud Services. 7th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE), IEEE, Ruse, Bulgaria, pp. 1-4. https://doi.org/10.1109/eeae49144.2020.9279072.

[17] Huertas Barros, E. (2011). Collaborative learning in the translation classroom: preliminary survey results. Journal of Specialized Translation, Spain, page 42. http://www.jostrans.org/issue16/art_barros.pdf.

[18] Carlos A. J., Francisco A. C., Fernando T., Sebastian D., Francisco E., and Oscar, R. (2009). Real-time

collaboration of virtual laboratories through the internet. Computers & Education, ScienceDirect, 52(1):126-140. https://doi.org/10.1016/j.compedu.2008.07.007.

[19] Heinimäki, O.P., Volet, S., Jones, C., Laakkonen, E., and Vauras, M. (2021). Student participatory role profiles in collaborative science learning: Relation of within-group configurations of role profiles and achievement. Learning, Culture and Social Interaction, vol 30, pp. 100539. https://doi.org/10.1016/j.lcsi.2021.100539.

[20] Dukhanov, A., Karpova, M., and Bochenina, K. (2014). Design Virtual Learning Labs for Courses in Computational Science with Use of Cloud Computing Technologies. Procedia Computer Science, vol 29, pp. 2472-2482. https://doi.org/10.1016/j.procs.2014.05.231.

[21] Mechta, D., Harous, S., Djoud, M., and Douar, A. Supervision and control tool for collaborative virtual l@aboratory (2012). Journal of Software (JSW), 7(7), pp.1496-1504, ISSN 1796-217X. https://doi.org/10.4304/jsw.7.7.1496-1504.

[22] Allwood, J., Traum, D., and Jokinen, K. (2000). Cooperation, dialogue and ethics. International Journal of Human-Computer Studies, vol 53, pp.871-914. https://doi.org/10.1006/ijhc.2000.0425.

[23] Bruffee, K.A. (1995). Sharing our toys-Cooperative learning versus collaborative learning. Change: The Magazine of Higher Learning, Routledge, vol 27, pp.12-18. https://doi.org/10.1080/00091383.1995.9937722.

[24] Dillenbourg, P., Järvelä, S., Fischer, F. (2009). The Evolution of Research on Computer-Supported Collaborative Learning. In: Balacheff, N., Ludvigsen, S., de Jong, T., Lazonder, A., Barnes, S. (eds) Technology-Enhanced Learning. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-9827-7.

[25] Akiyama, Y., & Cunningham, D. J. (2018). Synthesizing the Practice of SCMC-based Telecollaboration: A Scoping Review. CALICO Journal, 35(1), pp.49-76. https://doi.org/10.1558/cj.33156.

[26] Lamy, M., and Goodfellow, R. (2010). Telecollaboration and learning 2.0. In S. Guth, & F. Helm (Eds.), Telecollaboration 2.0: Languages, literacies and intercultural learning in the 21st century, Bern, Switzerland: Peter Lang, pp. 107-138. https://doi.org/10.5565/rev/jtl3.498.

[27] Kyungwon, G., Taejin, H., and Woontack, W. (2013). Digilog space generator for tele-collaboration in an augmented reality environment. R. Shumaker (Ed.): VAMR/HCII, Part I, LNCS 8021, Springer-Verlag Berlin Heidelberg, pp. 343-350. https://doi.org/10.1007/978-3-642-39405-8_38.

A Hesitant Fuzzy Multiplicative Base-criterion Multi-criteria Group Decision Making Method

Monika Narang[1*], M.C. Joshi[1] and A.K. Pal[2]
E-mail: monikanarang5@gmail.com, mcjoshi69@gmail.com, arun_pal1969@yahoo.in
[1]Department of Mathematics, D.S.B. Campus, Kumaun University, Nainital, India
[2]Department of Mathematics, Statistics & Computer Science, College of Basic Sciences & Humanities
Govind Ballabh Pant University of Agriculture & Technology, Pantnagar, India

Keywords: multi-criteria group decision-making, hesitant fuzzy sets, base-criterion method, pairwise comparisons

Received: February 22, 2021

Hesitant fuzzy sets have a unique characteristic that its basic element could manifest the assessment values of different decision makers on the same option under a certain criterion. Base-criterion method is a very significant tool for calculating the weights of the criteria in multiple criteria decision-making. In this paper, we developed a novel approach hesitant fuzzy BCM based on hesitant fuzzy multiplicative preference relation for multiple criteria group decision making. The base-comparison of the preferential criterion relative to other criteria is expressed as linguistic terms, which might be converted into hesitant multiplicative elements (HMEs). HMEs are extended along the same length according to the attitude of the decision makers. Then normalized optimal hesitant fuzzy weights are calculated. The normalized optimal hesitant fuzzy weights of criteria may be transferred to crisp values by employing score function. To illustrate the applicability and suitability of hesitant fuzzy BCM, we analyse the optimal transportation mode selection problem and car selection problem under hesitant fuzzy environment. The outcomes of the proposed model indicate that the hesitant fuzzy BCM is highly consistent and can yield appreciable preference ranking of criteria and alternatives.

Povzetek: Predstavljena je metoda odločanja v skupinah s pomočjo obotavljivih mehkih kriterijev.

1 Introduction

Decision making is the process of selecting an optimal alternative from a set of alternatives. Multi-criteria decision-making is an important tool of decision-making process that explicitly evaluates multiple conflicting criteria in decision making (both in daily life and in settings such as business, government and medicine) [1]. MCDM methods are divided into two main categories: Multi-Objective Decision-Making (MODM) and Multi-Attribute Decision-Making (MADM). The fundamental difference between MODM and MADM is that MODM have no predetermined alternatives and MADM have limited number of alternatives [2,3]. MODM methods are employed to handle continuous problems, on the other hand, multi-attribute decision making (MADM) methods are used to solve discrete problems. MCDM is commonly used to describe the discrete MADM.

Over the previous years, many MCDM methods have been introduced by researchers such as VIKOR [4], ELECTRE (Elimination and Choice Expressing Reality) [5], TOPSIS (Technique for order preference by similarity to an ideal solution) [6,7], COPRAS (Complex Proportional Assessment) [8,9], SWARA (step-wise weight assessment ratio analysis) [10], ANP (Analytic Network Process) [11,12], AHP (Analytic Hierarchy Process) [13], BWM (Best-Worst Method) [14,15].

In a practical problem, MCDM consists of two parts: (a) obtaining decision information, including criterion weight, (b) ranking the alternatives by a certain approach. The most important part is how we calculate the criterion weights, which has been the foundation for the introduction of many MCDM methods. Recently, Haseli et al [16] developed a novel Base-Criterion Method (BCM) which is a better route to determine the weight of the criteria. First, the decision maker selects the base-criterion (preferential, selective) and then a pairwise comparison is made between the base-criterion and other criteria. This technique is much clearer and more accurate because the execution of secondary comparisons is not necessary. This can achieve the weight of the criterion with less pairwise comparisons than existing MCDM methods. The final weights determined through BCM are very authentic as the comparison is completely consistent while other traditional MCDM techniques such as BWM and AHP have low inconsistency ratio.

It is hard to recognize all the facets of a decision-making problem for a single decision maker. The decisions made by groups are mostly different from those made by individuals. So, it is essential to have opinions from group of experts/decision makers. When more than one expert evaluates an option, it is very possible for them

* Corresponding author

to have different opinions. In today's environment, group decision making (GDM) methods pay more attention to the fuzzy information contained in group decision making problems.

Due to the hesitant of the purpose as well as the fuzziness of human mind, human decisions generally collide the characteristics of opacity. Prof. Zadeh in 1965 proposed the fuzzy set theory [17]. Fuzzy mention the things which are obscure. Working system of fuzzy sets is retraced uncertainty and lack of clarity in daily life. After that, several tools have been introduced to handle the fuzzy information such as interval valued fuzzy set [18,19], type-2 fuzzy set [20], intuitionistic fuzzy set [21], and hesitant fuzzy sets [22].

People usually hesitate about one thing or the other while making decisions, which makes it difficult to reach a final agreement. Hesitant Fuzzy Set (HFS) [22,23,24] is very powerful tool for obtaining the optimum alternative in a decision-making process with multi-criteria and multiple individuals. HFS allows the membership having a set of few values. Generally, the grades of preference are not symmetric, but distributed asymmetrically around some value. Saaty's 1/9-9 scale is a useful tool to deal with such a situation, especially in expressing the multiplicative preference relation applied in many areas. A hesitant fuzzy multiplicative preference relation (HFMPR), which is developed based on the fuzzy preference relation [25,26], is the most usual tool for expressing DMs' preferences over decision-making alternatives.

In the present study, using hesitant fuzzy multiplicative numbers the reference comparisons of BCM are executed from group decision making process to have more real and convincing ranking results. So, a novel approach, including BCM under hesitant fuzzy environment has been developed and employed for the first time. We expended the BCM to more authentic HF-BCM.

We systematize the rest of paper as follows: Section 2 demonstrates the basic concepts. Section 3 introduces the proposed methodology. Section 4 describes the application of our proposed method by considering simple example of decision-making problems. Section 5 shows findings and conclusion.

2 Preliminaries

Definition 2.1 ([22],[24]) Let X be a fixed set, HFS on X is in terms of a function that when applied to X returns a subset of [0,1].

Xia and Xu [27] expressed the HFS by a mathematical symbol:

$$A = \{\langle x, h_A(x)\rangle | x \in X\} \quad (1)$$

Where $h_A(x)$ is a set of some values in [0,1], denoting the possible membership degrees of the element $x \in X$ to the set A and $h = h_A(x)$ is a hesitant fuzzy element (HFE).

Definition 2.2 [27] For a HFE $h, s(h) = \frac{1}{l_h}\sum_{p\in h} p$, is called the score of h, where l_h is the number of the elements in h.

For two HFEs, h_1 and h_2, if $s(h_1) > s(h_2)$, then $h_1 > h_2$.

Definition 2.3 [27] Let h_1 and h_2 be two HFEs, then basic operations on HFEs are as follows:

$$h_1 \oplus h_2 = \bigcup_{p\in h_1, p\in h_2}\{p_1 + p_2\};$$
$$h_1 \ominus h_2 = \bigcup_{p_1\in h_1, p_2\in h_2}\{p_1 - p_2\};$$
$$h_1 \otimes h_2 = \bigcup_{p_1\in h_1, p_2\in h_2}\{p_1 . p_2\};$$
$$h_1 \oslash h_2 = \bigcup_{p_1\in h_1, p_2\in h_2}\left\{\frac{p_1}{p_2}\right\}$$
(2)

Definition 2.4 [13] Let $A = \{A_1, A_2, .., A_n\}$ be a set of n alternatives, then $B = \left(b_{ij}\right)_{n\times n}$ is called a multiplicative preference relation on $A \times A$, whose element b_{ij} estimates the preference of the alternative A_i over A_j , and is characterized by a ratio scale such as Saaty's ratio scale such that $b_{ij} \in \left[\frac{1}{9}, 9\right]$, and $b_{ij}. b_{ji} = 1, i, j = 1,2,.., n$.

Where b_{ij} unfolds indifference between A_j and A_i ; $b_{ij} > 1$ unfolds that A_i is preferred to A_j , especially, $b_{ij} = 9$ unfolds that A_i is absolutely preferred to A_j ; $b_{ij} < 1$ unfolds that A_j is preferred to A_i , especially, $b_{ij} = \frac{1}{9}$ unfolds that A_j is absolutely preferred to A_i.

Definition 2.5 [28] Let $X = \{x_1, x_2, \dots . x_n\}$ be a fixed set, then a hesitant fuzzy multiplicative preference relation on the set X is represented by the matrix $H = \left(h_{ij}\right)_{n\times n} \in X \times X$, where $h_{ij} = \left\{p_{ij}^l,\ l = 1,2, \dots \left|h_{ij}\right|\right\}$ is a HME which expresses all possible preference degrees of the alternative x_i over x_j given by the DMs. h_{ij} (HME) should satisfies the conditions:

$$p_{ij}^{\sigma(l)} . p_{ji}^{\sigma(|h_{ij}|-l+1)} = 1, \quad p_{ii} = 1, \quad \left|h_{ij}\right| = \left|h_{ji}\right|, \quad i, j = 1,2, \dots, n. \quad (3)$$

Where all p_{ij} are ranked in ascending order, $p_{ij}^{\sigma(l)}$ represents the l^{th} smallest value in h_{ij} and $\left|h_{ij}\right|$ unfolds the number of elements in h_{ij}. In particular, $p_{ii} = 1$ unfolds the indifference between x_i and x_j, $p_{ij} > 1$

indicates that x_i is preferred to x_j, and $p_{ij} < 1$ indicates that x_i is not preferred to x_j or x_j is preferred to x_i.

In fact, if the preference degree of the alternative x_i over x_j is p, then the preference degree of the alternative x_j over x_i should be $1/p$. Thus, in the hesitant multiplicative circumstance, the product of the l^{th} smallest value in h_{ij} and the l^{th} largest value in h_{ji} should be 1. The second condition defines that the preference degree of the alternative x_i over itself should be 1. The third one states that the lengths of h_{ij} and h_{ji} should be the same.

Example 2.1 If a group of decision makers is asked to give the estimation of the degree to which A_i is preferred to A_j $(i \neq j)$, some DMs give p_{ij}^1, some give p_{ij}^2 and others give p_{ij}^3, where $p_{ij}^1, p_{ij}^2, p_{ij}^3 \in \left(\frac{1}{9}, 9\right)$. Then the preference information h_{ij} that A_i is preferred to A_j is given by $h_{ij} = \left\{p_{ij}^1, p_{ij}^2, p_{ij}^3\right\}$. For alternatives A_i and A_k $(i \neq j \neq k)$, some DMs in a group may give p_{ik}^1 and the others give p_{ik}^2.

Then the preference information h_{ij} that A_i is preferred to A_k is given by $h_{ij} = \{p_{ik}^1, p_{ik}^2\}$.

Definition 2.6 [29] According to the definition of HME, it can be seen that the number of assessment values in different HMEs may vary and the assessment values in each HME are usually out of order.

To guarantee that the number of values in different HMEs be equal, we can add elements by the linear combination of the maximal and minimal element in h_{ij} with a parameter λ, shown as:

$$\overrightarrow{h_{ij}} = \lambda h_{ij}^{max} + (1-\lambda)h_{ij}^{min} \quad (4)$$

where $\lambda = 1$ and $\lambda = 0$ means the optimistic attitude and pessimistic attitude of DMs, respectively. $\lambda = \frac{1}{2}$ indicates the neutral attitude of the DMs and the adding element h_{ij} is the average value of h_{ij}.

3 Hesitant fuzzy BCM (HF-BCM)

Because of the hesitancy of the purpose as well as the fuzziness of human mind, human decisions generally clasp the characteristics of opacity. To elaborate hesitancy and vagueness involved in decision making, Hesitant fuzzy sets (HFS) have been introduced by Torra and Narukawa [22,24] as an extension of fuzzy sets [17]**.** HFS allows the membership having a set of several possible values to deal with uncertain information. when the assessment values given by experts are different, we could unify them into an HFE. The features of HFE are very compatible with GDM problems. BCM [16] is a novel MCDM method to calculate the weights of the criteria and alternatives. In the BCM, First, the decision maker selects the base-criterion (preferential) and then a pairwise comparison is made between the base-criterion and other criteria. This technique is much clearer and more accurate because the execution of secondary comparisons is not necessary.

In our proposed methodology, hesitant fuzzy BCM method is developed based on hesitant fuzzy multiplicative preference relation for multi-criteria group decision making (MCGDM).

3.1 Hesitant fuzzy BCM with hesitant fuzzy multiplicative preference relations

In a MCGDM, it is too difficult to determine the weights of the criteria because DMs often have different preferences for criteria and also uses natural language. Linguistic terms such as “Equally important”, “Extremely important” and “Strongly important” are used to make the hesitant pairwise comparisons. The decision makers provide the hesitant fuzzy multiplicative preference relations via pairwise comparison on the n criteria by using the Saaty’s 1/9-9 [11] scale. The rules of transformation are listed in Table 1. Suppose an HFMPRs be

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{n1} \end{pmatrix} \quad (5)$$

Where$h_{ij} = \{p_{ij}^l, l = 1,2, \dots |h_{ij}|\}$ (hesitant multiplicative element (HME)) unfolds the relative hesitant fuzzy preference (HFP) of criterion i to the criterion j given by the decision makers; $p_{ii} = 1$, when $i = j$.

In hesitant fuzzy BCM, Hesitant fuzzy pairwise comparisons are divided into two parts:

Definition 3.1 If i is the base-criterion, then h_{ij} is called hesitant fuzzy base comparison.

Definition 3.2 If i and j are not base-criterion, then h_{ij} is called hesitant fuzzy final comparison.

The basic principle of BCM [16] tells us that not all hesitant fuzzy pairwise comparisons are required to obtain a complete matrix. There is total n^2 HFMNs in the matrix $H = (h_{ij})_{n\times n}$. But we only require $n-1$ hesitant fuzzy pairwise comparisons (hesitant fuzzy base-comparison). The hesitant fuzzy final comparisons are taken from the hesitant fuzzy base comparisons. Without making the hesitant fuzzy final comparisons, optimal hesitant fuzzy weight values are obtained. The HFWs of criteria and alternatives with respect to various criteria could be derived using HF-BCM.

Now, we elaborate the steps of hesitant fuzzy BCM to determine the optimal hesitant fuzzy weights.

Step 1 Determine the decision criteria set and group of experts

Determine a set of decision criteria $\{C_1, C_2, C_3, \dots, C_n\}$ and group of experts $\{DM_1, DM_2, .., DM_k\}$ on the basis of which decision is taken.

Step 2 Determine the base-criteria (preferential, selective).

Decision makers select one of the criteria as a base-criteria (preferential) from a set of decision criteria $\{C_1, C_2, C_3, \dots, C_n\}$ but no comparison is performed in this step.

Step 3 Execute the hesitant fuzzy base-comparisons.

Based on Table 1, the relative hesitant fuzzy preference of the base criteria at the other criteria is derived. The linguistic terms are transformed into HMEs. The resulting vector of hesitant fuzzy base-comparisons as follows.

$$H_B = (h_{B1}, h_{B2}, h_{B3}, \dots, h_{Bn})$$

Table 1: The Saaty’s Scale.

1/9-9 scale	0.1-0.9 scale	Linguistic terms
1/9	0.1	Extremely not important
1/7	0.2	Very strongly not important
1/5	0.3	Strongly not important
1/3	0.4	Moderately not important
1	0.5	Equally important
3	0.6	Moderately important
5	0.7	Strongly important
7	0.8	Very strongly important
9	0.9	Extremely important
Other values b/w 1/9-9	Other values b/w 0-1	Intermediate values used to present compromise

where H_B unfolds the hesitant fuzzy base-criteria at the other criteria vector, and h_{Bj} unfolds the hesitant fuzzy preference (HFP) of the base-criteria over the j criteria.

Step 4 Normalization of the hesitant fuzzy base-comparison vector.

According to the definition of HME, it can be seen that the number of assessment values in different HMEs may vary and the values in each HME are usually out of order.

Using Equation 4, the hesitant fuzzy base-comparison vector is normalized according to decision-makers' attitude.

Step 5 Derive the normalized hesitant fuzzy optimal weights.

The normalized optimal hesitant fuzzy weights for each $\frac{h_{wB}}{h_{wj}}$ will be equal to h_{Bj} for all j. The optimal hesitant fuzzy weight values can be determined by absolute differences $\left|\frac{h_{wB}}{h_{wj}} - h_{Bj}\right|$ for all j. Regarding the weight values are non-negative, the normalized optimal hesitant fuzzy weights can be determined by deriving the problem as follows:

$$\text{Min max}\left|\frac{h_{wB}}{h_{wj}} - h_{Bj}\right| \le \xi$$

$$\text{Such that} \quad \begin{cases} \sum_{j=1}^{n} R\left(h_{w_j}\right) = 1 \\ p_j^l \ge 0 \; for \; all \; j \end{cases} \qquad (6)$$

Where $h_{wB} = \{p_B^l, l = 1,2,..,|h_{ij}|\}$, $h_{wj} = \{p_j^l, l = 1,2,..,|h_{ij}|\}$, $\xi = \{p^{l\xi}, l = 1,2,..,|h_{ij}|\}$

The Equation 6 can be rewritten as the nonlinearly constrained problem.

min ξ

Such that

$$\begin{cases} \left|\frac{h_{wB}}{h_{wj}} - h_{Bj}\right| \le \xi \\ \sum_{j=1}^{n} R\left(h_{w_j}\right) = 1 \\ p_j^l > 0 \; for \; all \; j; l = 1,2,..,|h_{ij}| \end{cases} \qquad (7)$$

Regarding the HMEs and $\xi = \{k^*\}$, Equation 7 can be rewritten as:

min ξ

Such that

$$\begin{cases} \left|\frac{(p_B^l)}{(p_j^l)} - (p_{Bj}^l)\right| \le (k^*) \\ \sum_{j=1}^{n} R\left(h_{w_j}\right) = 1 \\ p_j^l > 0 \; for \; all \; j; l = 1,2,..,|h_{ij}| \end{cases} \qquad (8)$$

Solution of the Equation 8 gives the normalized optimal hesitant fuzzy weights $\{h_{w_1}, h_{w_2}, \dots, h_{w_n}\}$, then these normalized optimal HFWs can be converted to crisp numbers by employing score function. Figure 1 represents the flowchart of the proposed methodology.

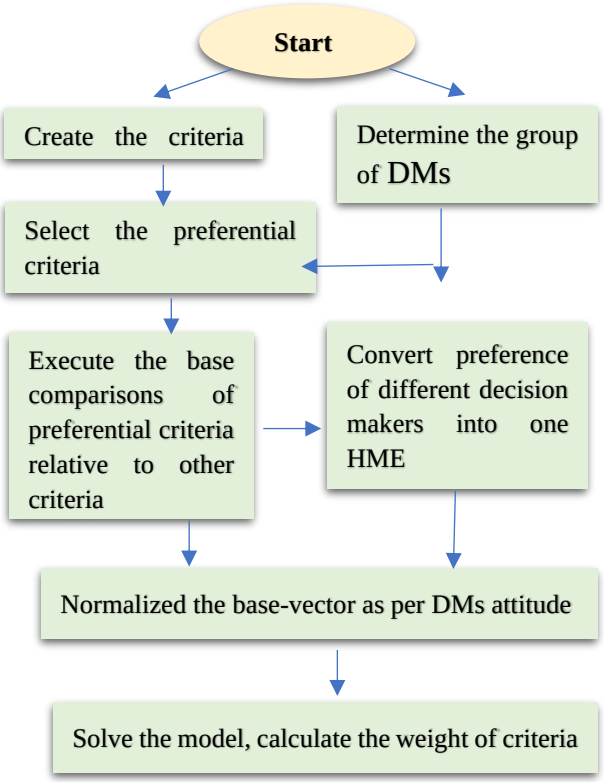


Figure 1: Flowchart of the proposed methodology.

3.2 Consistency for HF-BCM

The hesitant fuzzy pairwise comparison is fully consistent if

$$h_{Base.i} \times h_{ij} = h_{Base.j} \quad \text{for all } i \text{ and } j.$$

The decision maker should pursuance the following principle in entrusting the hesitant fuzzy multiplicative numbers for hesitant fuzzy base-comparisons.

$$\left(\frac{1}{9}, \frac{1}{8}, \dots\right) \le \{p_{ij}^l, l = 1,2, \dots |h_{ij}|\} \le (8,9, \dots)$$

$$\left(\tfrac{1}{9}, \tfrac{1}{8}, \dots\right) \le \frac{(p_{Bj}^l, l=1,2,..,|h_{ij}|)}{(p_{Bi}^l, l=1,2..,|h_{ij}|)} \le (8,9, \dots) \qquad (9)$$

4 Case study

In this section, we describe the application of hesitant fuzzy BCM by considering simple examples of decision-making problems.

4.1 Case study 1

A company wants to select the best transportation mode to deliver the product to the market place. As a case study, we adopted the example of mode of transport described in [16] and deal the problem by using our proposed hesitant fuzzy BCM method from group decision making process.

There are three criteria chosen for the optimal transport mode selection issue: (a) load flexibility (b) accessibility (c) cost. The group of decision makers chooses cost criterion as the base-criterion. DMs executes hesitant fuzzy base-comparisons based on HMEs using Table 1 group of decision makers provide the estimation of the degree to which cost is preferred to load flexibility, some DMs provide the preference value 6, and others

Table 2: Hesitant fuzzy pairwise comparisons.

Transportation criteria	Load flexibility	Accessibility	Cost
Base-criterion (Cost)	(6,8)	(1,2,3,4)	(1,1,1,1)

Table 3: Scores of criteria for all the three cases.

Cases	$S(h_{w_1})$	$S(h_{w_2})$	$S(h_{w_3})$	ξ
$\lambda = 1$	0.256	0.725	2.016	0.0000008
$\lambda = 1/2$	0.267	0.717	2.010	0.0000003
$\lambda = 0$	0.290	0.711	1.997	0.0000002

provide the preference value 8. Group of decision makers provide the estimation of the degree to which cost is preferred to accessibility, some DMs provide the preference value 2, some DMs provides the preference value 4, some DMs provides the preference value 1 and other provides 3. On the basis which the hesitant fuzzy base-comparison vector can be obtained as:

$$H_B = [(6,8),(1,2,3,4),(1,1,1,1)]$$

Now, the normalized hesitant fuzzy base-comparison vector can be computed by using Equation 4 as follows:

Case-1 Optimistic attitude of DMs ($\lambda = 1$).

$$H_B = [(6,8,8,8),(1,2,3,4),(1,1,1,1)]$$

Case-2 Neutral attitude of DMs ($\lambda = \frac{1}{2}$).

$$H_B = [(6,7,7,8),(1,2,3,4),(1,1,1,1)]$$

Case-3 Pessimistic attitude of DMs ($\lambda = 0$).

$$H_B = [(6,6,6,8),(1,2,3,4),(1,1,1,1)]$$

The maximum length in the normalized hesitant fuzzy preference is 4. Thus, the number of elements in the normalized hesitant fuzzy preference are 4, and the number of several possible values in the normalized hesitant fuzzy optimal weights is four as well.

Based on Table 2 and Case-1, the normalized optimal hesitant fuzzy weight for each criterion can be obtained by solving the non-linear optimization problem as follows:

$$\min \xi$$

Such that

$$\begin{cases} \left|\frac{(p_3^1,p_3^2,p_3^3,p_3^4)}{(p_1^1,p_1^2,p_1^3,p_1^4)} - (p_{31}^1,p_{31}^2,p_{31}^3,p_{31}^4)\right| \le (k^*,k^*,k^*,k^*) \\ \left|\frac{(p_3^1,p_3^2,p_3^3,p_3^4)}{(p_2^1,p_2^2,p_2^3,p_2^4)} - (p_{32}^1,p_{32}^2,p_{32}^3,p_{32}^4)\right| \le (k^*,k^*,k^*,k^*) \\ \left|\frac{(p_3^1,p_3^2,p_3^3,p_3^4)}{(p_3^1,p_3^2,p_3^3,p_3^4)} - (p_{33}^1,p_{33}^2,p_{33}^3,p_{33}^4)\right| \le (k^*,k^*,k^*,k^*) \\ \sum_{j=1}^{n} R\left(h_{w_j}\right) = 1 \\ p_j^l > 0 \; for \; all \; j. \end{cases} \quad (10)$$

By putting HMEs of hesitant fuzzy base-comparison vector in the Equation 10, the nonlinearly constrained optimization problem as follows:

$$\min k^*$$

Such that

$$\begin{cases} p_3^1 - 6*p_1^1 \le k*p_1^1; p_3^1 - 6*p_1^1 \ge k*p_1^1; \\ p_3^2 - 8*p_1^2 \le k*p_1^2; p_3^2 - 8*p_1^2 \ge k*p_1^2; \\ p_3^3 - 8*p_1^3 \le k*p_1^3; p_3^3 - 8*p_1^3 \ge k*p_1^3; \\ p_3^4 - 8*p_1^4 \le k*p_1^4; p_3^4 - 8*p_1^4 \ge k*p_1^4; \\ p_3^1 - p_2^1 \le k*p_2^1; p_3^1 - p_2^1 \ge k*p_2^1; \\ p_3^2 - 2*p_2^2 \le k*p_2^2; p_3^2 - 2*p_2^2 \ge k*p_2^2; \\ p_3^3 - 3*p_2^3 \le k*p_2^3; p_3^3 - 3*p_2^3 \ge k*p_2^3; \\ p_3^4 - 4*p_2^4 \le k*p_2^4; p_3^4 - 4*p_2^4 \ge k*p_2^4 \\ \left(\frac{1}{12}\right) * \begin{pmatrix} p_1^1+p_1^2+p_1^3+p_1^4+p_2^1+p_2^2+ \\ p_2^3+p_2^4+p_3^1+p_3^2+p_3^3+p_3^4 \end{pmatrix} = 1; \\ p_1^1 \le p_1^2 \le p_1^3 \le p_1^4; \\ p_2^1 \le p_2^2 \le p_2^3 \le p_2^4; \\ p_3^1 \le p_3^2 \le p_3^3 \le p_3^4; \\ p_1^1, p_2^1, p_3^1, p_4^1 > 0; \\ k^* \ge 0; \end{cases} \quad (11)$$

After solving the Equation 11, the normalized optimal HFWs of criteria are obtained, which are:

$$h_{w_1} = \{0.081,0.162,0.292,0.492\};$$
$$h_{w_2} = \{0.489,0.649,0.779,0.985\};$$
$$h_{w_3} = \{0.489,1.299,2.337,3.941\};$$

By employing score function, the crisp weights of normalized optimal HFWs are obtained.

Load flexibility $S(h_{w_1}) = 0.256$;

Accessibility $S(h_{w_2}) = 0.725$; Cost $S(h_{w_3}) = 2.016$;

Similarly, the normalized optimal HFWs of criteria are obtained for all the cases. Table 3 shows the calculations.

It can be noted from Table 4 that for BCM [16] and HF-BCM criteria have the same preference order, but there is a slight difference in the criteria weights. Due to $\xi = 0$ (0.00000008), regardless of any values for the consistency index, the consistency ratio is optimal. Also, the proposed method is even better than the HF-BWM [27] in terms of consistency. Since the consistency ratio is the closest value to zero. This example unfolds that the hesitant fuzzy BCM method could consider the ambiguity of DMs in the process of group decision making.

Table 4: Comparison of results.

Methods	**Pairwise comparisons**	**Weights of criterion**	**Consistency**
HF-BCM (Proposed)	HMPRs (1/9-9 scale)	0.267 0.717 2.010	0.0000
BCM [16]	1/9-9 scale	0.076 0.307 0.615	0.0000
BWM [14]	1-9 scale	0.071 0.338 0.589	0.0580
HF-BWM [27]	HMPRs (1-9 scale)	0.413 1.120 1.465	0.0558

4.2 Case study 2

We consider the example of car selection which is handled by using BCM method [16]. In this case study, we solve the same problem by our proposed method HF-BCM from group decision making process. As the number of criteria increases, it becomes difficult to specify values for relative preference in base-comparison. There are six criteria chosen for the car selection problem: (1) convenience (2) fuel consumption (3) safety (4) style (5) acceleration (6) consumer price. The group of decision makers chooses safety criterion as the base-criterion. DMs executes hesitant fuzzy base-comparisons based on HMEs using Table 1. On the basis which the hesitant fuzzy base-comparison vector can be obtained as:

$$H_B = \left[(1,2), \left(\frac{1}{3}, \frac{1}{2}\right), (1,1,1), (2,4,6), \left(\frac{1}{2}\right), \left(\frac{1}{4}, \frac{1}{3}\right)\right]$$

The maximum length in the normalized hesitant fuzzy preference is 3. So, the number of elements in the normalized hesitant fuzzy preference are 3.

The normalized hesitant fuzzy base-comparison vector

can be computed by using Equation 4 as follows:

Case-1 Optimistic attitude of DMs ($\lambda = 1$).

$$H_B = \begin{bmatrix} (1,2,2), \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{2}\right), (1,1,1), (2,4,6), \\ \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{4}, \frac{1}{3}, \frac{1}{3}\right) \end{bmatrix}$$

Case-2 Neutral attitude of DMs $\left(\lambda = {}^{1}/_{2}\right)$.

$$H_B = \begin{bmatrix} \left(1, \frac{1}{2}, 2\right), \left(\frac{1}{3}, \frac{5}{12}, \frac{1}{2}\right), (1,1,1), (2,4,6), \\ \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{4}, \frac{7}{24}, \frac{1}{3}\right) \end{bmatrix}$$

Case-3 Pessimistic attitude of DMs ($\lambda = 0$).

$$H_B = \begin{bmatrix} (1,1,2), \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{2}\right), (1,1,1), (2,4,6), \\ \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{3}\right) \end{bmatrix}$$

Based on Table 5 and Case-1, the normalized optimal hesitant fuzzy weight for each criterion can be obtained by solving the non-linear optimization problem as follows:

$\min \xi$

Such that

$$\begin{cases} \left|\frac{(p_3^1,p_3^2,p_3^3)}{(p_1^1,p_1^2,p_1^3)} - (p_{31}^1,p_{31}^2,p_{31}^3)\right| \le (k^*,k^*,k^*) \\ \left|\frac{(p_3^1,p_3^2,p_3^3)}{(p_2^1,p_2^2,p_2^3)} - (p_{32}^1,p_{32}^2,p_{32}^3)\right| \le (k^*,k^*,k^*) \\ \left|\frac{(p_3^1,p_3^2,p_3^3)}{(p_3^1,p_3^2,p_3^3)} - (p_{33}^1,p_{33}^2,p_{33}^3)\right| \le (k^*,k^*,k^*) \\ \left|\frac{(p_3^1,p_3^2,p_3^3)}{(p_4^1,p_4^2,p_4^3)} - (p_{34}^1,p_{34}^2,p_{34}^3)\right| \le (k^*,k^*,k^*) \\ \left|\frac{(p_3^1,p_3^2,p_3^3)}{(p_5^1,p_5^2,p_5^3)} - (p_{35}^1,p_{35}^2,p_{35}^3)\right| \le (k^*,k^*,k^*) \\ \left|\frac{(p_3^1,p_3^2,p_3^3)}{(p_6^1,p_6^2,p_6^3)} - (p_{36}^1,p_{36}^2,p_{36}^3)\right| \le (k^*,k^*,k^*) \\ \sum_{j=1}^{n} R\left(h_{w_j}\right) = 1 \\ p_j^l > 0 \; for \; all \; j. \end{cases} \quad (12)$$

By putting HMEs of hesitant fuzzy base-comparison vector in the Equation 12, the nonlinearly constrained optimization problem as follows:

$\min k^*$

Such that

$$\begin{cases} p_3^1 - 1 * p_1^1 \le k * p_1^1; p_3^1 - 1 * p_1^1 \ge k * p_1^1; \\ p_3^2 - 2 * p_1^2 \le k * p_1^2; p_3^2 - 2 * p_1^2 \ge k * p_1^2; \\ p_3^3 - 2 * p_1^3 \le k * p_1^3; p_3^3 - 2 * p_1^3 \ge k * p_1^3; \\ p_3^1 - \frac{1}{3} * p_2^1 \le k * p_2^1; p_3^1 - \frac{1}{3} * p_2^1 \ge k * p_2^1; \\ p_3^2 - \frac{1}{2} * p_2^2 \le k * p_2^2; p_3^2 - \frac{1}{2} * p_2^2 \ge k * p_2^2; \\ p_3^3 - \frac{1}{2} * p_2^3 \le k * p_2^3; p_3^3 - \frac{1}{2} * p_2^3 \ge k * p_2^3; \\ p_3^1 - 2 * p_4^1 \le k * p_4^1; p_3^1 - 2 * p_4^1 \ge k * p_4^1; \\ p_3^2 - 4 * p_4^2 \le k * p_4^2; p_3^2 - 4 * p_4^2 \ge k * p_4^2; \\ p_3^3 - 6 * p_4^3 \le k * p_4^3; p_3^3 - 6 * p_4^3 \ge k * p_4^3; \\ p_3^1 - \frac{1}{2} * p_5^1 \le k * p_5^1; p_3^1 - \frac{1}{2} * p_5^1 \ge k * p_5^1; \\ p_3^2 - \frac{1}{2} * p_5^2 \le k * p_5^2; p_3^2 - \frac{1}{2} * p_5^2 \ge k * p_5^2; \\ p_3^3 - \frac{1}{2} * p_5^3 \le k * p_5^3; p_3^3 - \frac{1}{2} * p_5^3 \ge k * p_5^3; \\ p_3^1 - \frac{1}{4} * p_6^1 \le k * p_6^1; p_3^1 - \frac{1}{4} * p_6^1 \ge k * p_6^1; \\ p_3^2 - \frac{1}{3} * p_6^2 \le k * p_6^2; p_3^2 - \frac{1}{3} * p_6^2 \ge k * p_6^2; \\ p_3^3 - \frac{1}{3} * p_6^3 \le k * p_6^3; p_3^3 - \frac{1}{3} * p_6^3 \ge k * p_6^3; \\ \left(\frac{1}{18}\right) * \begin{pmatrix} p_1^1 + p_1^2 + p_1^3 + p_2^1 + p_2^2 + p_2^3 + p_3^1 \\ + p_3^2 + p_3^3 + p_4^1 + p_4^2 + p_4^3 + p_5^1 + p_5^2 \\ + p_5^3 + p_6^1 + p_6^2 + p_6^3 \end{pmatrix} = 1; \\ p_1^1 \le p_1^2 \le p_1^3; \\ p_2^1 \le p_2^2 \le p_2^3; \\ p_3^1 \le p_3^2 \le p_3^3; \\ p_4^1 \le p_4^2 \le p_4^3; \\ p_5^1 \le p_5^2 \le p_5^3; \\ p_6^1 \le p_6^2 \le p_6^3; \\ p_1^1, p_2^1, p_3^1 > 0; \\ k \ge 0 \end{cases} \quad (13)$$

Table 5: Hesitant fuzzy pairwise comparisons.

Criteria	**Base-criterion (Safety)**
Convenience	(1,2)
Fuel	(1/3,1/2)
Safety	(1,1,1)
Style	(2,4,6)
Acceleration	(1/2)
Price	(1/4,1/3)

Table 6: Scores of criteria for all the three cases.

Cases	$\boldsymbol{\lambda = 1}$	$\boldsymbol{\lambda = 1/2}$	$\boldsymbol{\lambda = 0}$
ξ	0.00000	0.00000	0.00000
$S(h_{w_1})$	0.247	0.343	0.283
$S(h_{w_2})$	0.937	0.921	0.984
$S(h_{w_3})$	0.440	0.428	0.412
$S(h_{w_4})$	0.102	0.088	0.093
$S(h_{w_5})$	0.881	0.856	0.824
$S(h_{w_6})$	1.384	1.360	1.400

After solving the Equation 13, the normalized optimal HFWs of criteria are obtained, which are:

$$h_{w_1} = \{0.161,0.210,0.370\};$$
$$h_{w_2} = \{0.487,0.841,1.483\};$$
$$h_{w_3} = \{0.160,0.420,0.741\};$$
$$h_{w_4} = \{0.080,0.105,0.123\};$$
$$h_{w_5} = \{0.321,0.841,1.483\};$$
$$h_{w_6} = \{0.643,1.271,2.242\};$$

By employing score function, the crisp weights of normalized optimal HFWs are obtained.

Convenience = 0.247, fuel consumption = 0.937, safety = 0.440, style = 0.102, acceleration = 0.881, price = 1.384, ξ = 0.00000007;

Consistency ratio is optimal because ξ = 0 (0.00000007).

Table 6 shows the normalized optimal HFWs of criteria for all the cases.

5 Conclusion

Multi-criteria group decision making problems often have strong uncertainty, which is characterized as fuzziness within the group decision-making. When more than one expert evaluates an option, it is very possible for them to have different opinions. In our proposed methodology, we have developed a unique approach hesitant fuzzy Base-criterion method which is an extension of latest MCDM method BCM in the hesitant situation for multi-criteria group decision making. HFS has its unique advantages in a decision-making problem with multi-criteria and multiple individuals. Employing linguistic variables is more worthy than crisp values, in order to make a base-comparison for criteria and alternatives in the group decision-making process.

The base-comparisons values manifested by linguistic terms could be converted into HMEs that are used in the nonlinear optimization problem and then normalized optimal HFWs of criteria can be obtained. The normalized optimal hesitant fuzzy weights of criteria can be transferred to crisp values by employing score function. The suitability and applicability of the developed HF-BCM is verified by discussing the optimal transportation mode selection problem and car selection problem. In terms of strength and direction, the results of HF-BCM are completely consistent if decision makers have selected HMEs for pairwise-comparisons based on Equation 9. The proposed method HF-BCM has several important features as follows:

- Integration of Base-criterion method with hesitant fuzzy circumstances is novel and provides more reliable and accurate weights of criteria and alternatives for MCGDM.
- Hesitant fuzzy set has its unique advantages over the other methods. In MCGDM, when the assessment values given by DMs vary, we can unify them into one HME. Also, BCM requires fewer comparisons to calculate the weights of criteria and alternative. In this way, the process of calculating lots of group decision making problems is effectively simplified.
- Compared to the BCM method, the hesitant fuzzy BCM method also uses linguistic terms to do base-comparisons. Use of linguistic variables makes pairwise-comparisons more accurate and easier. HF-BCM can gain the more reliable weights.
- The hesitant fuzzy BCM needs less number of pairwise comparisons and highly consistent rather than the hesitant fuzzy AHP and hesitant fuzzy BWM methods.

References

[1] Parreiras, R., Pedrycz, W., Ekel, P. (2011). Fuzzy Multicriteria Decision-Making: Models, Methods and Applications. John Wiley & Sons. http://dx.doi.org/10.1109/IFSA-NAFIPS.2013.6608469

[2] Yoon, K., Hwang, C.L. (1981). Multiple Attribute Decision Making: Methods and Applications. A State-of-the-Art Survey. New York, Springer-Verlag.

[3] Rao, R.V. (2013). Decision Making in The Manufacturing Environment Using Graph Theory and Fuzzy Multiple Attribute Decision Making. 2. London, Springer-Verlag.

[4] Duckstein, L., Opricovic, S. (1980). Multiobjective optimization in river basin development. Water Resources Research, vol 16, no. 1, pp. 4-20. https://doi.org/10.1029/WR016i001p00014

[5] Benayoun, R., Roy, B., Sussman, N. (1996). Manual de, du programme ELECTRE. Note Synth. Form. Vol, 25.

[6] Hwang, C.L., Yoon, k. (1981). Multiple attribute decision making, methods and applications. Springer, New York

[7] Yoon, K. (1987). A reconciliation among discrete compromise solutions. Journal of the Operational Research Society, vol 38, no. 3, pp. 277-86. https://doi.org/10.2307/2581948

[8] Zavadskas, E.K. (1994). "The new method of multicriteria complex proportional assessment of projects," Technological and Economic Development of Economy, pp. 131-139.

[9] Zavadskas, E.K., Kaklauskas, A. & Kvederytė, N. (2001). Multivariant design and multiple criteria analysis of building life cycle. Informatica, vol 12, no. 1, pp. 169–188. https://doi.org/10.3233/INF-2001-12111

[10] Keršuliene, V., Zavadskas, E.K., Turskis, Z. (2010). Selection of rational dispute resolution method by applying new step-wise weight assessment ratio analysis (SWARA). Journal of Business Economics and Management, vol 11, no. 2, pp. 243-258. https://doi.org/10.3846/jbem.2010.12

[11] Satty, T.L. (1996). Decision making with dependence and feedback: The analytic network process. RWS Publication

[12] Saaty, T.L. (2005). Theory and applications of the analytic network process: decision making with benefits, opportunities, costs, and risks. RWS publications.

[13] Saaty, L. (1980). The Analytical Hierarchy Process. McGraw-Hill, New York

[14] Rezaei, J. (2015). Best-worst multi-criteria decision-making method. Omega, vol 53, pp. 49–57. https://doi.org/10.1016/j.omega.2014.11.009

Rezaei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. Omega, vol 64, pp. 126–130. https://doi.org/10.1016/j.omega.2015.12.001

[15] Haseli, G., Sheikh, R., Sana, S.S. (2019). Base-criterion on multi-criteria decision-making method and its applications. International journal of management science and engineering management, vol 15, no. 2, pp. 79-88. https://doi.org/10.1080/17509653.2019.1633964

[16] Zadeh, L.A. (1965). Fuzzy sets. Information Control, vol 8, pp. 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

[17] Pramanik, T., Samanta, S., Pal, M., Mondal, S., Sarkar, B. (2016). Interval-valued fuzzy φ-tolerance competition graphs. SpringerPlus, vol 5, no. 1. https://doi.org/10.1186/s40064-016-3463-z

[18] Rashmanlou, H., Pal, M., Borzooei, R.A., Mofidnakhaei, F., Sarkar, B. (2018). Product of interval-valued fuzzy graphs and degree. Journal of Intelligent Fuzzy System, vol 35, no. 6, pp. 6443–6451. https://doi.org/10.3233/JIFS-181488

[19] Dubois, D., Prade, H. (1980). Systems of linear fuzzy constraints. Fuzzy Sets and System, vol 3, no. 1, pp. 37–48.

[20] Atanassov, K.T. (1999). Intuitionistic fuzzy sets. In: Intuitionistic fuzzy sets. Physica Heidelberg, pp. 1-137. https://doi.org/10.1007/978-3-7908-1870-3_1

[21] Torra, V. (2010). Hesitant fuzzy sets. International Journal of Intelligent Systems, vol 25, pp. 529–539. https://doi.org/10.1002/int.20418

[22] Torra, V., Narukawa, Y. (2007). Modeling decisions: Information fusion and aggregation operators. Springer http://dx.doi.org/10.1007/978-3-540-68791-7

[23] Torra, V., Narukawa, Y. (2009). On hesitant fuzzy sets and decision. In: The 18th IEEE International Conference on Fuzzy Systems, Jeju Island, Korea, pp. 1378–1382. https://doi.org/10.1109/FUZZY.2009.5276884

[24] Xia, M.M., Xu, Z.S. (2013). Managing hesitant information in GDM problems under fuzzy and multiplicative preference relations. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, vol 21, pp. 865–897. https://doi.org/10.1142/S0218488513500402

[25] Xia, M.M., Xu, Z.S. (2011a). Studies on the aggregation of intuitionistic fuzzy and hesitant fuzzy information. Technical Report.

[26] Ali, A., Rashid, T. (2019). Hesitant fuzzy best worst multi-criteria decision-making method and its applications. International Journal of Intelligent Systems, vol 34, pp. 1953-1967. https://doi.org/10.1002/int.22131

[27] Xu, Z., Zhang, S. (2019). An overview on the applications of the hesitant fuzzy sets in group decision-making: Theory, support and methods. Frontiers of Engineering Management, vol 6, no. 2, pp. 163-182. https://doi.org/10.1007/s42524-019-0017-4

[28] Xu, Z. S., & Zhang, X. L. (2013). Hesitant fuzzy multi-attribute decision making based on TOPSIS with incomplete weight information. Knowledge-Based Systems, vol 52, pp. 53–64. https://doi.org/10.1016/j.knosys.2013.05.011

Formal Approach to Data Accuracy Evaluation

Belkacem Athamena[1] and Zina Houhamdi[2]
E-mail: athamena@gmail.com, belkacem.athamena@aau.ac.ae, z_houhamdi@yahoo.fr, zina.houhamdi@aau.ac.ae
[1]Business Administration Department, College of Business, Al Ain University, United Arab Emirates
[2]Cybersecurity Department, College of Engineering, Al Ain University, United Arab Emirates

Keywords: data integration systems, data quality, data accuracy

Received: December 20, 2019

Usually, data quality is defined by multiple attributes that allow classifying the output data (such as completeness, freshness, and accuracy) or the methods exploiting these data (such as dependability, performance, and protection). Among the suggested quality attributes, we will discuss one of the principal categories: data accuracy. Scientific experiments, decision–making, and data retrieval are examples of situations that require a formal evaluation approach to data accuracy. The evaluation approach should be adaptable to distinct understandings of data accuracy and distinct end-user expectations. This study investigates data accuracy and defines dimensions and metrics that affect its evaluation. The investigation of data accuracy generates problems in the user expectation specification and database quality models. This work describes our proposed approach for data accuracy evaluation by defining an evaluation algorithm that considers the distribution of inaccuracies in database relations. The approach decomposes the query output in accordance with data accuracy, labels every part with its accuracy value, and addresses the possibility of enforcing data accuracy by using these values. This study mainly contributes by proposing an explicit evaluation of quality attributes of data accuracy, a formal evaluation approach to data accuracy, and suggesting some improvement actions to reinforce data accuracy.

Povzetek: Opisana je formalna metoda preverjanja točnosti podatkov.

1 Introduction

Data quality has increasingly become an essential characteristic required by users particularly for data integration systems (DISs), which involve combining data residing in multiple databases and providing the user with a unified view of these data as answers to their queries [4]. Because of the growth of retrieved data, users are becoming increasingly worried about data quality. On the other hand, the number of quality attributes and their relationships are huge. Accordingly, data quality evaluation is considered as a complex problem involving multiple variables. In DIS context, data quality evaluation is exceptionally complicated because of the combination of data derived from different databases with possibly distinct qualities. Because of the large number and high heterogeneity of databases that are independent, it is crucial to precisely determine their quality and to consider it in the design phase of DISs.

System quality improvement is associated with optimization of a problem with multiple variables, which may be very complex specifically in an indefinite context [3, 26]. Consequently, it is arduous to consider all quality attributes at the same time. To investigate data quality thoroughly, it is inevitable to investigate each quality attribute independently besides the factors of the context that affect it. Dependencies will be investigated later. Among the proposed quality attributes, we opt for the main category: data accuracy. Currently, many systems consider the need to have reliable measurements of data accuracy as a crucial and decisive requirement. These systems are numerous and are used in diverse fields, such as customer relationship management, web-services integration, scientific experiments, decision–making, and data retrieval.

This study discusses data accuracy analysis in DIS and considers the relational scenario. Explicitly, it treats user queries that are composed of projections, selections, and joins (PSJ) operators over a collection of database relations. It addresses the problem of accuracy evaluation of data delivered to the user in response to user queries and decides if the expectations of the user on the data accuracy can be achieved or not.

Our approach consists of fragmenting the query output in sets of tuples (called areas), which have uniform accuracy values, and marking each area with its accuracy value. Consequently, the user can do the following:

- extract only the most precise data by selecting the area that possesses high accuracy values,
- exclude data that do not satisfy the accuracy threshold by ignoring areas possessing low accuracy values, or
- classify data by sorting areas based on their accuracy values. Moreover, if we desire displaying additional data (because the first accurate area is incomplete), we can show the next area and so on. Thus, this will represent an added value to the delivered data.

This paper illustrates our proposed accuracy evaluation algorithm. We present the values of semantic accuracy using the Boolean metric where each cell of every area of any relation is assigned a value accurate or inaccurate. However, the accuracy of the whole area is calculated as an aggregation of the cells' accuracy values by dividing the number of accurate cells by the total number of cells in the area. The query results are sorted by the areas' accuracy values. All areas with accuracy value bigger than the user threshold are considered as area with high accuracy value and on the other hand, the areas with accuracy value less than the user threshold are considered as area with low accuracy value.

The fragmentation of query output and evaluation of data accuracy requires an in–depth analysis of the inaccuracy distribution in database relations and their union to generate the query output [24]. For this purpose, we split database relations into areas possessing uniform accuracies. As clarified in [15], databases are usually of heterogeneous quality; consequently, assigning an accuracy value for the entire relation is considered as an imprecise accuracy calculation of particular data [14]. The area is described as a view (selection and projection) over the relations of the database. That is to say, a fragment aggregates a group of relations specified by the predicates characterizing the fragment [8, 9, 25].

This paper presents a formal approach to evaluate data accuracy that considers the portions of database relations and processes them to generate an output for user query. It focuses on pre–evaluation, that is, the data accuracy is estimated before the execution of the user query. The outputs of our evaluation approach will be useful to make a comparison between different query plans to select the plan with the maximal accuracy value. Furthermore, these outputs are beneficial during the design phase (to determine which databases will be included in the DIS) and monitoring phase (to calculate the query accuracy). Eventually, this study discusses the issue of accuracy improvement. It proposes the usage of output fragments to choose the areas with high accuracy values. We use the portions of a relation having the highest accuracy value instead of using the whole relation. This distinguishes our approach from existing approaches in the literature.

2 Background

Data accuracy symbolizes a set of quality attributes. This section describes the three accuracy attributes listed in the literature [23, 27]:

- *Syntactic accuracy* represents the level at which data do not contain syntactic mistakes such as format inconsistencies and spelling errors. It expresses the interval between descriptions of the data in the DIS and conventional descriptions of these data (syntactic gap).
- *Semantic accuracy* defines how adequately the data describe the environment state. It represents the interval between the data described in the DIS and real–world data (semantic gap).
- *Precision* describes the level of data details. It expresses the gap between the detail level of the DIS data and its planned detail level.

Apropos of accuracy measurements, three metrics are mentioned in the literature [12]:

- *Boolean metric* uses a Boolean value to indicate if the data detail is accurate (1 or True) or inaccurate (0 or False).
- *Degree metric* uses a dimension to capture the principle of how precise the data are; this usually belongs to a $[0-1]$ interval.
- *Value–deviation* is defined as an integer to capture the gap between data item in the system and the original one; this is usually normalized to a $[0-1]$ interval.

This section reviews the main concepts used in this study. First, it discusses current approaches for evaluating data accuracy as adapted from our proposed model, particularly, the fragmentation technique, and it recalls the characteristics of partitioning relational databases. It reviews the concept of query rewriting and explains the *bucket* algorithm. Finally, it comments on selectivity estimation techniques.

Our accuracy evaluation approach is supported by two techniques: *prior evaluation* approach, which presumes homogenous distribution of errors [16] and *posterior evaluation*, which uses the accuracy homogeneity for partitioning database relations [19].

Prior Evaluation Method: Naumann et al. [16] proposed a method that propagates quality factors (encompassing data accuracy) in conformity with query operators. The method estimates the query output accuracy based on the accuracy of the database. The database relation accuracy is the percentage of syntactic accuracy (rate of accurate cells).

The query is a PSJ. The method assumes that inaccuracies are homogeneously spread in the database relations; thus, regardless of the selected tuples or projected attributes, the database accuracy is conserved. Concerning the join operation, the joined data accuracy is computed by multiplying the accuracy of input relations [20].

The disadvantage of this method is the strong assumption on a homogenous error distribution (usually inapplicable to real–world data). Generally, query operations do not maintain accuracies and accordingly, this method does not obtain an exact calculation of the accuracy. This deficiency is due to the absence of knowledge on inaccuracy distribution (where the errors are concentrated). Supplementary knowledge characterizing instances of relation is mandatory to obtain outputs that are more accurate.

Posterior Evaluation Method: This method uses a partitioning algorithm [15, 19] for fragmenting the database relations into areas that have extremely uniform accuracy. An area is described as a view, which involves projection and selection operations. The accuracy is calculated by considering the portion of each relation and then computing the cell accuracy of that portion. The area accuracy is the ratio of semantic accuracy (rate of correct cells). The accuracy values will be exploited in portion fragmentation by applying a computerized algorithm for fragmentation that evaluates a set of criteria. After that, the same fragmentation is performed over the complete database relation.

Usually, the query is a PSJ and relational algebra is employed to operate with the partitions, i.e., the operator takes as inputs the set of relations with their respective partitions. Its output is a single relation with its partitions: the partition of projection (selection) is determined as the intersection of the operation output and the partition of the input relation (intersection of selection conditions with projected attributes). In this case, the data accuracy is conserved owing to accuracy uniformity. Finally, when the query output is ready, we calculate the accuracy value as the weighted sum of area accuracy (weight is equal to the number of cells in the area). Note that a single value of accuracy is computed for the entire output.

Partition Algorithm: We will concisely describe Rakov's algorithm for sampling a relation [19]. Figure 1 shows the related pseudo–code. It is an iterative function using a classification tree. It takes a relation as input and splits it into two blocks (vertical or horizontal splitting but never both) and it tries to identify the block with the highest uniformity; then it iterates the procedure for each block.

The block partitioning terminates if it provides a negligible amelioration in uniformity. A threshold x notes a fair uniform distribution of accuracies in the block and it serves as a stopping condition. Uniformity is estimated using Gini indices [6]. The Gini index $GI(P)$ formula is given by equation 1:

$$GI(P) = 2r(1-r) \tag{1}$$

where r represents the rate of accurate cells in partition P. The equation 2 defines the halt constraint that estimates the reduction of the partition of the Gini index:

$$\Delta GI = GI(P) - \alpha_1 GI(P_1) - \alpha_2 GI(P_2) \tag{2}$$

where $\{P_1, P_2\}$ are partitions of P and $\alpha_i = \frac{|P_i|}{|P|}, i = 1, 2$.

Because consideration of the complete possible relation partitions is excessively costly, Rakov suggests some heuristics to decrease the number of treated partitions. In addition, we distinguish two types of attributes: categorical and ordered [1]. In the case of horizontal partitioning, if the attribute is ordered and possesses K different values $(A_1 \leq \cdots \leq A_k)$, the partitions are identified as $(K-1)$ binary conditions $t \leq A_i$. Otherwise, in the case of a categorical attribute that possesses K different values, these values are sorted based on the number of incorrect cells. After that, they are considered as ordered attributes. For vertical partitioning, we consider all partitions for a small number of attributes; otherwise, for a large number of attributes, we apply the same method employed for categorical attributes. In the following, we discuss the characteristics of well–constructed partitions.

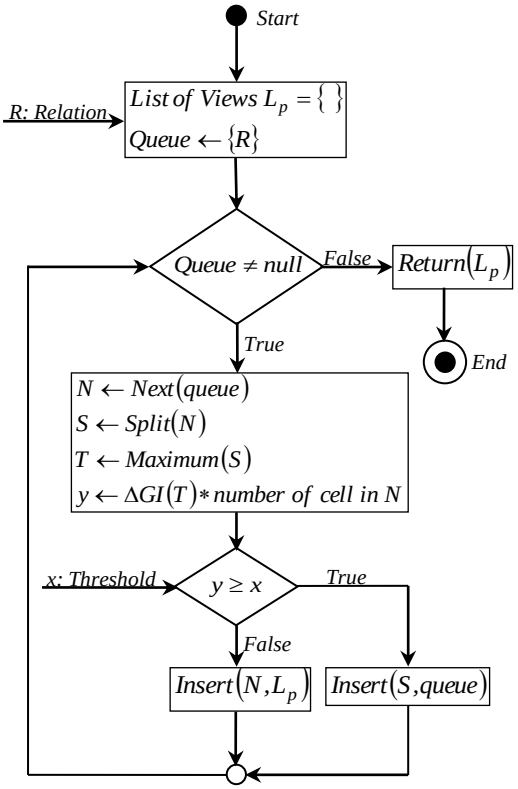


Figure 1: Relation Partitioning Algorithm.

Partition correctness : There are three correctness constraints that a partition must satisfy to guarantee database coherence [17]. The constraints are related to the following:

- *Disjunction*: The horizontal decomposition of a relation R into partitions R_1, ..., R_n, verifies that $\forall cell\ d_i/d_i \in R_j \Rightarrow d_i \notin R_k, k \neq j$ (i.e., each cell in the partition R_j does not appear in any other partitions R_k where $k \neq j$). This property ensures that the partitions are disjoints. This rule guarantees that there is no intersection between all horizontal partitions. In the case of vertical decomposition, the disjunction is limited to non-primary key attributes of the relation because the primary key attributes are basically duplicated in all partitions.

- *Restoration*: In the case of decomposing a relation R into partitions R_1, ..., R_n, there is always a way to

find an operator $\Omega/R = \Omega R_i$, $i = 1 \dots n$. This rule assures the preservation of data restrictions defined as dependencies. Normally, the union operator is used for horizontal partitioning whereas the join operator is used for vertical partitioning.

– *Completeness*: Assuming that the decomposition of a relation R into partitions $R_1, \dots, R_n$, all items belonging to relation R also belong to one or more R_i partitions. This characteristic ensures data preservation, i.e., there is no data loss (all data in global relation are projected in partitions). For horizontal partitioning, items denote tuples; however, for vertical partitioning, they denote attributes.

Query rewriting: is the reformulation of a user query (defined by the global relation) to a possible analogous scheme, known as *rewriting*, which concerns exclusively the database framework [11]. In the local–as–viewed (LAV) strategy, the global relation is defined without referring to the databases, and after that, a mapping between them is made by expressing each database relation as a view over the global relation [21]. Thus, query rewriting is comparable to the application of views to answer a user query.

Definition: For a query Q over relations R_i defining the global relation and views V_i referring to database relations over R_i, the query Q_r is called a rewriting of Q if:

– $Q_r \subset Q$

– Q_r concerns exclusively the views.

In general, the query is a PSJ and is defined using a datalog–like language. Thus, query Q is expressed by equation 3:

$$Q(X) = R_1(Z_1) \wedge \dots \wedge R_n(Z_n) \wedge C_Q \qquad (3)$$

where

– R_i *is a relation* and $Z_i = \{A_i\}/A_i$ *is an attribute* of R_i

– C_Q is a *conjunction* predicate, $C_Q = u\theta v/\theta \in \{=, <, >, \leq, \geq\}$, and $u, v \in \bigcup_{1 \leq i \leq n} Z_i$

– $X \subseteq \bigcup_{1 \leq i \leq n} Z_i = \{A_i\}/A_i$ projected by Q

Numerous algorithms for query rewriting can be found in the literature and [7] presents a review of these algorithms. The most popular is the bucket algorithm [13], which we will describe and use in this study.

The bucket algorithm computes all possible rewritings that are included in (but not imperatively analogous to) the initial query [5]. The algorithm prunes the area of candidate rewritings in two phases:

– $\forall t \in Q$: Construct a container called bucket that includes the contributing database relations for t, i.e., the database relations containing tuples of t.

– Construct candidate rewritings (query integration by joining one database relation from each bucket) and preserve only the rewritings belonging to Q.

The first phase possesses a polynomial complexity with respect to the database number. However, the second phase reduces the complexity considerably by diminishing the number of possibilities. Despite the fact that containment is generally manageable, its resolution is equal to the query size (usually small) and happens when the query has many occurrences of exact schemas; accordingly, the containment complexity is not a challenge in real applications [5].

Selectivity Evaluation: Techniques for selectivity evaluation are widely practiced in query optimization. The statistics of data saved in the database provides an estimation to the optimizer [22]. The most popular statistics are histograms, which are adopted by several business database systems. This section shows their current application to evaluate the selectivity of a complex query.

The histogram of attribute A is defined as a list of buckets, where bucket b_i defines r_i, which is a subarea of A's area, and possesses two attributes:

– periodicity p_i, which represents the number of tuples x verifying $x.A \in r_i$, and

– discrete value dv_i, which represents the number of different values of $x.A$ in all tuples $x/x.A \in r_i$.

The query selectivity is calculated by dividing the query cardinality by the relation cardinality. In order to evaluate the query cardinality, we sum the periodicity of all buckets included (totally or partially) in the predicate. If the query has many extended predicates, the selectivity is calculated as the product of all selectivities.

For a random PSJ query, there is a new challenge: cardinality evaluation needs statistical information propagation over predicates, i.e., we must create histograms for in–between outputs using the histograms of database relations. The propagation of in–between statistics in addition to multiple query operators can considerably decrease the accuracy. A possible solution is a precalculation of statistics for a subset of query results that are exceptional and to use them in selectivity calculation of in–between outputs.

The remainder of this paper describes the proposed model for evaluation of data accuracy. Algorithms and methods reviewed previously are modified and adopted in data accuracy estimation.

3 Formal model

This section formalizes the proposed approach for accuracy assessment. The approach considers DIS in relational context (the system combines data from relational databases

and allows users to execute queries over a global schema). Each query is reformulated according to the database relations in order to obtain a set of rewritings that select data to answer the query.

The accuracy of data delivered to a user as answer to the user's query is addressed in this section. We arrange data in areas with uniform accuracy to apprise the user on the inaccuracy distribution. To achieve this goal, database relations are decomposed into partitions (called areas) possessing uniform accuracy. Note that areas are views (virtual relations) determined by the distribution rules (projection attributes and selection conditions). The user query is reformulated over the areas (rather than in the global relation). The proposed approach uses an *a priori* assessment method: before performing the query rewriting, we assess the data accuracy, depending only on the operations (PSJ) that define the rewritings and the area accuracy. We can state the problem as follows:

- *Input:*
  - User query
  - A collection of database relations decomposed into areas having uniform accuracy
- *Output:*
  - A rewriting set answering the user query
  - Accuracy value of rewritings
  - Accuracy value of the query answer

Our proposed scenario for accuracy estimation contains three phases:

- *Database relation fragmentation based on accuracy uniformity*: This stage estimates the accuracy of a portion of individual database relation and uses the estimation results for relation fragmentation. The fragmentation phase is performed at the initial stage (DIS development or regularly) but is isolated from the query appraisal stage.
- *Query rewriting with respect to fragments*: Every query is reformulated using the areas of the database relations. This stage generates a set of rewritings using the areas. The output of the query is the union of the created rewritings.
- *Data accuracy estimation of query outputs*: This stage estimates the accuracy of data generated by the rewritings, using the area accuracy, and aggregates them to calculate the accuracy value for the whole user query output.

To reinforce the data accuracy, a simple adjustment consists of rejecting rewritings or areas with small accuracy values. The rejection can be done at the rewriting generation time (early stage). Thus, our approach can be considered as selective. The next subsections provide the details of each phase.

3.1 Database relation fragmentation based on accuracy uniformity

Using the partitioning approach suggested by Rakov [19], the data accuracy is used for fragmenting each database relation. The purpose of fragmentation is the manipulation of pieces of database relation that are uniformly accurate; in other words, if we try to fragment the accuracy value again, it will approximately stay unchangeable [2]. This subsection defines the fragmentation process and discusses how to extract appropriate fragments.

To simplify the partition usage, the approach suggests decomposing the relation into areas using a horizontal partition (selection predicates) and after that, it decomposes each area into subareas using a vertical partition (subsets of attributes). Fragments should verify the three completeness constraints discussed previously: each tuple exists in a unique area, key attributes appear in all subareas, and non–key attributes appear in a single subarea. To manipulate all attributes (key and non–key) in a similar way (projection of the attributes in unique subarea), the key attributes are duplicated. Consequently, a new key for the subarea (composite key) is generated. This approach is called key expansion.

Key expansion definition: If $R(A_1, \ldots, A_m, \ldots, A_n)$ is a relation, where $m \leq n$, and $(A_1, \ldots, A_m)$ form the key of R, $\bar{R}$ denotes the key expansion of R generated by duplicating the key attributes: $\bar{R}(K_1, \ldots, K_m, A_1, \ldots, A_m, \ldots, A_n)$. The duplicated attributes $(K_1, \ldots, K_m)$ are the expanded keys of $\bar{R}$.
In vertical fragmentation, we project the expanded key in all subareas, and we project all attributes of the initial relation in a unique subarea. The fragmentation process (area and subarea) can be formalized as follows:

Horizontal fragmentation: A horizontal fragmentation of a relation R, represented by $H_{P_1,\ldots,P_m}(R)$ (see equation 4), consists of defining a set of subrelations $\{R_1, \ldots, R_m\}$, named areas created by the application of predicates $P_1, \ldots, P_m$ to $\bar{R}$, where a conjunctive predicate set $\bar{P} = \{P_1, \ldots, P_m\}$ over R, disjoint and complete (each tuple of R verifies a unique P_i):

$$H_{P_1,\ldots,P_m}(R) = R_1, \ldots R_m = \sigma_{P_1}(\bar{R}), \ldots, \sigma_{P_m}(\bar{R}) \quad (4)$$

The area is denoted by $< Name, Predicate, N, Key_{Accuracy} >$:

- *Name* distinguishes areas within the relation,
- *Predicate* is the conjunction that determines the area,
- *N* represents the number of tuples in the area (verify the *Predicate*), and
- $Key_{Accuracy}$ is the accuracy value of the key attributes of R.

Vertical fragmentation: For n attribute subsets of a relation R, where $\frac{R(S_1,\ldots,S_n)}{\bigcap_i^n S_i} = \phi$: the *vertical fragmentation* of an area R_i, $R_i \in R$, is expressed as $V_{S_1 \ldots S_m}(R_i)$ (see equation 5), which defines a list of views $\{V_{i1}, \ldots, V_{in}\}$, named subareas generated by projecting the attribute subsets to R_i. The subareas are disjointed (each attribute of R belongs to one subarea). However, the expanded key K of $\bar{R}$ belongs to all subareas:

$$\begin{aligned} V_{S_1 \ldots S_n}(R_i) =& V_{i1}, \ldots, V_{in} \\ =& \pi_{k,S_1}(R_i), \ldots, \pi_{k,S_n}(R_i) \end{aligned} \quad (5)$$

The subarea is defined as a three–tuple $< Name, Attributes, Accuracy >$, where *Name* distinguishes the subarea within the relation, *Attributes* define the list of subarea attributes, and finally, *Accuracy* denotes the calculated subarea accuracy.

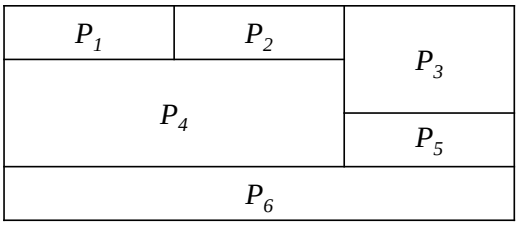
Again, subareas and areas are views determined by fragmentation rules (projection attributes and selection conditions); in other words, the database relations are not really partitioned and saved as independent partitions. Database relations are preserved unvaried in databases and the description models of subareas and areas are kept in the DIS.

After the horizontal and vertical fragmentation of the relation, any relation cell appears in a single subarea belonging to the single area. To fragment database relations, Rakov's algorithm described earlier can be applied. However, it should be noted that the approach alternately executes horizontal and vertical partitioning. Accordingly, the resulting fragments can be different from the areas and their subareas. As a possible solution, we propose to reorganize the fragment in accordance with the horizontal and vertical fragments.

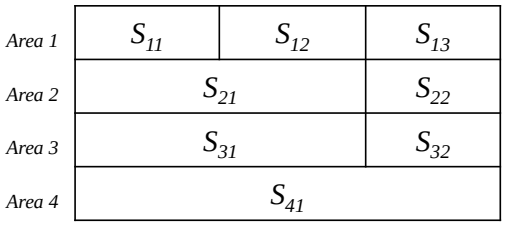
For any hybrid fragmentation (horizontal and vertical) of a relation R, which consists of a set of areas that verify the correctness criteria, it is always possible to obtain different fragments of R by additional fragmentation of some of the areas, i.e., the algorithm obtains the subareas $\{S_{11}, \ldots, S_{1m_1}, \ldots, S_{n1}, \ldots S_{nm_n}\}$, where the subset $\{S_{i1}, \ldots, S_{1i_1}\}$ represents the vertical fragmentation of some area A_i ($1 \leq i \leq n$), and $\{A_1, \ldots, A_n\}$ set represents the horizontal fragmentation of R. To this end, we follow a process that contains four steps:

- Find the selection predicates $P = \{p_1, \ldots, p_r\}$ defining the partition $T_1, \ldots, T_k / r \leq k$ (because multiple partitions possess identical selection predicates).
- Search for two non-disjunctive predicates p_i and p_j (i.e., $p_i \bigcap p_j \neq \phi$) and then replace p_i and p_j by $p_i - p_j$, $p_j - p_i$, and $p_i \bigcap p_j$. This step will terminate because the predicates are the unions of inequalities on R's attributes. The output is a set of predicates $P' = \{p'_1, \ldots, p'_n\}$, which are disjoint.
- Determine the area set $A = \{A_1, \ldots, A_n\}$. For each predicate $p'_i \in P'$, there is an area $A_i \in A$, where A corresponds to the horizontal fragmentation of R; in other words, A verifies the completeness constraints, which implies that predicates sub-expressions are not lost.
- Intersect every area $A_i \in A$ with partitions $T_1, \ldots, T_k$ to obtain a new set of subareas $\{S_{i1}, \ldots, S_{1i_1}\}$. This new set corresponds to the vertical fragmentation of A_i; in other words, it verifies the completeness criteria.

Consequently, each hybrid fragmentation of a relation that satisfies the correctness constraints can be reorganized in areas and subareas; particularly, the fragments generated by Rakov's algorithm. Figure 2 shows an example of a reorganization.



(a) Before Reorganization



(b) After Reorganization

Figure 2: Fragments Reorganization.

To fragment database relations, we suggest applying Rakov's algorithm. Nevertheless, the fragmentation can be done manually by taking advantage of knowledge concerning the database relations collected from the user, DIS expert, or IT administrator.

Rakov's algorithm calculates the area accuracy as a percentage of accurate cells. Equation 6 calculates the Gini indices:

$$G(V) = 2k(1-k) \quad (6)$$

where k is the accuracy value, and the indices will be used to estimate the accuracy uniformity, and thereafter, to select the fragment representing the highest value of uniformity (i.e., the fragmentation function is parameterized to compute the area accuracy). Each fragment is reorganized as illustrated earlier. The area and subarea schema are also deduced from the related fragment using the following steps:

1. Name the areas sequentially and then their associated subareas.
2. Define the area predicate and the list of attributes describing subareas from the fragment.

3. Estimate the subarea accuracy by calculating the cell accuracy average (computed by Rakov's algorithm).

4. Estimate the key accuracy by calculating the average of accuracies of attributes conforming to the key (the accuracy of attributes is equal to their subarea accuracy owing to accuracy uniformity). Note that the key accuracy corresponds to the subarea accuracy if all key attributes are projected in a single subarea.

5. At the end, as Rakov's algorithm fragments a sample of database relation, the number of tuples in the sample is used to deduce the number of tuples in an area, i.e., $number\ of\ tuples \times \frac{relation\ size}{sample\ size}$.

The fragmentation is executed one time at DIS development or regularly. However, it is not related to the accuracy estimation of the query. The following section describes the query reformulation and its impact on accuracy estimation.

3.2 User query rewriting

Our approach proposes to reformulate the user query according to database relation areas. We will take into consideration the option of reformulation of user query in reference to subareas and we clarify the reason for discarding this alternative.

To describe a user query based on areas that constitute database relations, our approach uses the classic rewriting algorithm, i.e., the bucket algorithm. Hence, the LAV technique is applied to define the areas as views of a global schema by replacing the database relation with its description over the global schema, i.e., it unfolds views over the database relation. Note that this operation is unrelated to the user query because it is performed after fragmentation of database relations.

The query rewriting applies the bucket algorithm that creates buckets, compares predicates of the query and the area, and then determines possible rewritings and checks query containment. The number of rewritings increases with the number of areas in a polynomial manner. Note that the rigidity of the rewriting algorithm is not due to the number of relations, but to the query size (which is approximately small) and exists only if the query has more than one occurrence of the same relations [13].

Now, we discuss the option to rewrite a query over subareas and we explain why this option is rejected. Areas contain all tuples of database relations. Consequently, the rewriting algorithm of a query over areas joins all tuples of database relations (similar to query rewriting over database relations). Nevertheless, a subarea decomposes tuples as it projects a part of attributes of the relation. First, some versions of the bucket algorithm generate the total required attributes by joining multiple relations of a bucket. Consequently, query rewriting over subareas is not impossible. On the other hand, assembling a small number of attributes belonging to a large number of relations augments hazardously the risk of interpolating semantic inaccuracies (generation of tuples without meaning in real world). By way of illustration, the output can be the student name and the phone number of a different student who possess identical identifiers in distinct databases. This problem is intrinsic to DIS, but it remarkably grows if the number of joins increases (particularly when tuples are split). Moreover, the rewriting algorithm avoids splitting tuples specifically to reduce this risk.

Accordingly, our approach rewrites the query over areas and inevitably uses a subarea structure for computing data accuracy. This also explains why the proposed approach fragments relations in a horizontal and then vertical manner rather than by random management of hybrid fragmentations as Rakov's approach. The following section describes the estimation of data accuracy for each rewriting.

3.3 Data accuracy estimation

To compute area accuracies and aggregate them to calculate the accuracy value of the entire user query output, the proposed approach proceeds in three steps:

- identification of areas and subareas constituting the rewriting,
- estimation of the rewriting accuracy and its key accuracy, and
- calculation of the number of tuples in each area to perform the aggregation.

Identification of areas and subareas: An area contains at least one subarea and the rewriting is expressed as the join of area sets. The joining output is a unique area, which consists of the union of all subareas possessing part of the projected attributes and verifies the selection predicates of all input areas. The following example shows three areas, namely P_1, P_2, and P_3, decomposed to subareas (P_{11}, P_{12}), (P_{21},P_{22}), and (P_{31},P_{32},P_{33}) respectively (see Figure 3). The rewriting QR is defined as one area joining all subareas, which include part of the projected attributes. Note that, P_{12} is not included in QR because it does not contain any of the projected attributes in QR.

Because all tuples in the output of the rewriting fulfill all area predicates, a single area is built for the rewriting by joining the area predicates and rewriting predicates. In addition, because the bucket algorithm generates only relevant rewritings, the predicates cannot be contradictory. Note that we list only the predicates that are more constraining than others (e.g., "$age \geq 18$" is less constraining than "$age = 20$").

The rewriting subareas are defined as the aggregation of all subareas belonging to input areas, by considering just the common attributes between the subarea and rewriting. If the intersection between the rewriting and subarea is null, the subarea is discarded. The resulting rewriting

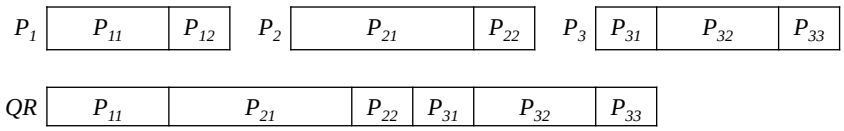


Figure 3: Rewriting Joins Multiple Areas.

subareas represent the vertical fragmentation of the rewriting area. The completeness constraint is satisfied because rewriting attributes appear in certain subareas of the input areas. However, for the disjunction constraint, the natural join emerges as a problem related to joining attributes existing in two different input subareas and appear only once in the rewriting. To solve this problem, a new subarea is created and its accuracy value will be computed as the average of both subarea accuracies (this will be discussed later).

Estimation of rewriting accuracy: If fragments are adequately determined, they logically conserve the accuracy of subareas because they are affected by the projection of predicates and attributes (owing to the accuracy uniformity). However, for the joining operation, each tuple in the subarea is composed of two input tuples; consequently, the produced subarea accuracy depends on both input areas. Particularly, the accuracy propagation can be different and depends on the accuracy factor. Recall that during the evaluation of semantic correctness, if the key of a particular tuple is inaccurate (does not reflect the real–world object), the entire tuple is considered as incorrect. Actually, semantic correctness evaluates the correspondence of an attribute (the key) to the real–world entity. On the other hand, the syntactic correctness evaluates the cell accuracy without regard to the key attributes. Thus, during area joining, the cell accuracy is computed differently from semantic correctness. Consequently, the subarea accuracy in the join output is computed as the product of the input subarea accuracy and key accuracy of the remaining areas. However, during syntactic correctness evaluation, the cell accuracy is equal to the cell accuracy of the input subarea.

Thus, our approach proceeds as follows:

- *Semantic correctness*: subarea accuracy is computed as the product of the input subarea accuracy and the key accuracy of the remaining areas.
- *Syntactic correctness*: subarea accuracy is calculated as the input subarea accuracy.

In the case of subareas that include join attributes (which were separated in the prior step), we calculate the accuracy for each input subarea separately and then take the average. The subareas with identical accuracy values are merged in a single subarea.

At the end, the accuracy of the key is calculated by multiplying the accuracy of keys in all areas. Then, the rewriting accuracy is calculated as the average of the cell accuracies (weighted average of subarea accuracy, where the weight is the number of attributes in the subarea).

Selectivity estimation: Because the query result is determined as the fusion of multiple rewritings, its accuracy is estimated as the weighted aggregation of the rewriting accuracy, where the weight is equal to the rewriting size (number of tuples in the rewriting). Thus, the estimation of the number of tuples in each rewriting is necessary.

To this end, we suggest estimating the rewriting selectivity to deduce its number of tuples. In general, a join corresponds to the comparison between two keys (primary and foreign), and then histograms are used for the calculation. Note that query optimization algorithms or statistical information about the prior execution of the same/similar query can be applied to estimate the selectivity. Particularly, experts also can estimate the selectivity. Remember that our approach does not depend on the estimation algorithm but certainly, the resulting accuracy value depends on it. The selectivity is defined as follows:

Given a query rewriting Q_R over areas $\{R_1, \ldots, R_k\}$, the selectivity of Q_R, expressed as $S(Q_R)$, represents the number of tuples in the Cartesian product of all areas that satisfy the rewriting predicates. If Q_R does not have a selection (or join) predicate, the whole set of tuples is considered and in this manner, $S(Q_R) = 1$. Concerning the rewriting query, the number of tuples is calculated as:

$$Number\ of\ tuples\ in\ the\ input\ areas\ \times \\ Rewriting\ selectivity$$

3.4 Quality graph

After generation of the query rewritings, we build a quality graph to calculate the user query as the union of all rewritings [10, 18]. We can have multiple rewritings or a unique one.

The creation of the quality graph shown in Figure 4 follows these definitions:

- *Source* nodes represent database relations.
- *Target* node represents user query.
- Activity nodes represent areas of database relations, denoted as Area. Edges connect the area to its source.
- Activity nodes represent the rewritings, denoted by *Rewriting*. Edges connect the rewriting to the area indicated by the rewriting.

– One activity node represents the union of rewritings (possibly single rewriting), denoted by the *Union* node. This node is the successor of all rewriting nodes and the predecessor of the *Target* node.

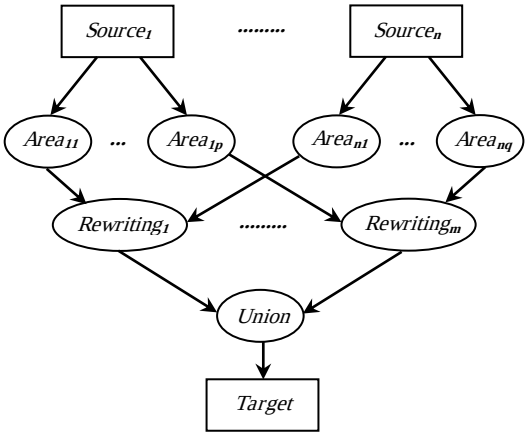


Figure 4: Quality Graph.

3.5 Accuracy estimation algorithm

We propose an algorithm for data accuracy estimation. The pseudo–code is shown in Figure 5. It implements the approach previously described (section 3.3) consisting of three phases:

1. For each *Rewriting*, the algorithm generates an area, defines its subareas, and computes the characteristic values (cardinality, predicates, subarea accuracy, and key accuracy).

2. The algorithm assembles all rewritings to determine the *Union* node.

3. The algorithm performs the accuracy value aggregation for all data edges.

In the first phase, the algorithm executes two loops over area nodes that are inputted to each rewriting node. During the first loop, the key accuracy feature is calculated by multiplying the accuracy of all areas; the predicate feature is determined by merging the predicates of input areas and the rewriting predicate, and eliminating worthless constraints; finally, the cardinality feature is computed as the product of the rewriting selectivity and the input area cardinalities.

During the second loop, to estimate the subarea feature, we just insert subareas of all input areas. This should be performed in another loop as the input of subarea accuracy is the key accuracy computed in the first loop. Then the algorithm adds subareas to the rewriting node and joins subarea attributes with the rewriting attributes. The natural join attributes are stored in new subareas. We calculate their accuracy using equation 7:

$$Accuracy = \frac{\sum_1^n subarea_i}{n} \tag{7}$$

where n is the number of subareas. At the end, the subareas with identical accuracy values are merged. U_{areas} is a list that contains all generated areas for each rewriting; consequently, the second phase requires setting U_{areas} as the value of the union node.

Finally, for each source or activity node, the accuracy aggregation is calculated as the weighted sum of the subarea accuracy and the weight is calculated by multiplying the number of subarea attributes by the area cardinality. The resulting output is associated with all edges outgoing the node.

4 Case study

This example is used to demonstrate the proposed approach for data accuracy estimation. The Boolean metric is used to illustrate the evaluation of semantic correctness; nevertheless, alternative accuracy metrics and factors can be used.

Assume that the DIS global schema contains two relations that hold data concerning students and their marks respectively:

– $Student(ID, name, level, exam, phone, address, city)$

– $Mark(ID, mark, year)$

Attributes of the relation student are ID (the student identifier), $name$, $level$ (first level defined by initial interview and its value can be “low”, “medium”, or “high”), $exam$ (initial exam result; its value $\in [0, 1]$), $phone$, $address$, and $city$.

Attributes symbolizing the relation mark are ID (the student identifier), $mark \in [0 - 20]$, where 20 is the highest mark), and $year$. The relation keys are ID and $ID, year$ respectively.

Suppose that two databases provide data concerning students and their marks as presented in Table 1 and Table 2:

– $S(ID, name, level, exam, phone, address)$ // students residing at Al–Ain (UAE).

– $M(ID, mark, year)$ // the student marks.

The colored cells represent the inaccuracy. The aggregated accuracy values for S and M relations (calculated as the average of cell accuracies) are 0.6 $(40/66)$ and 0.77 $(37/48)$ respectively. The key expansions are:

$$\bar{S}(K_{ID}, ID, name, level, exam, phone, address)$$

and

$$\bar{M}(K_{ID}, K_{year}, ID, mark, year)$$

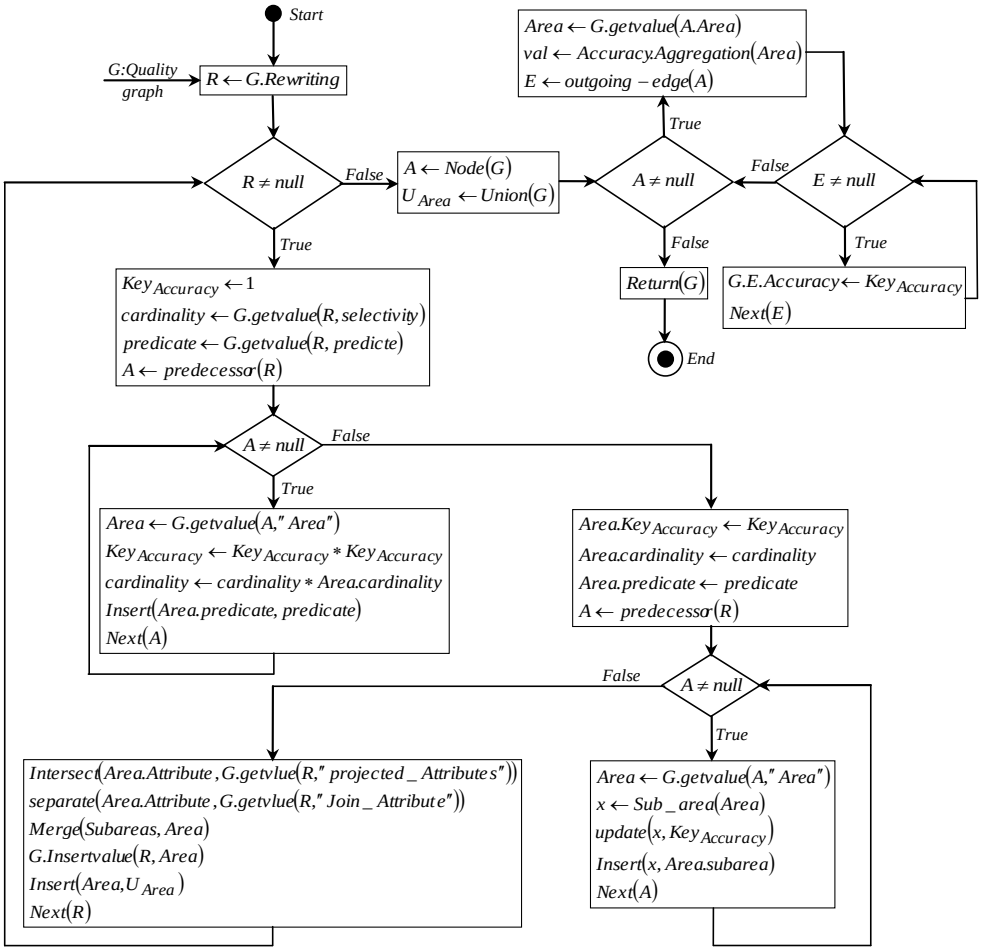


Figure 5: Accuracy Propagation Algorithm.

where $\{K_{ID}\}$ and $\{K_{year}, K_{ID}\}$ define the expanded key. Now S and M are fragmented to determine their areas and subareas; then, we calculate the number of tuples and their accuracies. An acceptable fragmentation of S is:

- Area S_1; $[ID < 300]$; 6 tuples; $keyaccuracy = 0.50$
  - Subarea S_{11}; $\{ID, name, level, exam\}$; $accuracy = 0.50$
  - Subarea S_{12}; $\{address, phone\}$; $accuracy = 0.25$
- Area S_2; $[ID \geq 300]$; 5 tuples; $keyaccuracy = 1.00$
  - Subarea S_{21}; $\{ID\}$; $accuracy = 1.00$
  - Subarea S_{22}; $\{name, level, exam, phone, address\}$; $accuracy = 0.80$

Similarly, a possible fragmentation of the M relation is:

- Area M_1; $[year < 2014]$; 1 tuple; $keyaccuracy = 0$
  - Subarea M_{11}; $\{ID, mark, year\}$; $accuracy = 0.00$ $(0/3)$
- Area M_2; $[year \geq 2014 \wedge id < 300]$; 6 tuples; $keyaccuracy = 0.50$
  - Subarea M_{21}; $\{ID, mark, year\}$; $accuracy = 0.50$ $(9/18)$
- Area M_3; $[year \geq 2014 \wedge ID \geq 300]$; 8 tuples; $keyaccuracy = 0.93$
  - Subarea M_{31}; $\{ID, mark, year\}$; $accuracy = 0.93$ $(25/27)$

After horizontal and vertical fragmentation of the relation, each cell belongs to a specific subarea included in a unique area; consequently, the fragments are represented by different colors describing different subareas. Table 3 presents the fragments of the student relation. The expanded key is not colored because it belongs to all subareas and it can be left out in the graphical illustration.
The respective schemas of S and M are:

- $S(ID, name, level, exam, phone, address, city) \leftarrow Student(ID, name, level, exam, phone, address) \wedge city = \text{“}Al - Ain''$

Table 1: S Relation.

ID	Name	Level	Exam	Phone	Address
120	Rani	High	1	6001104	12
123	Nada	Medium	0.465	99628734	Chiab al Alashekhar
141	Areej	Low	0.987		Al–Ain
154	Hanan	Low	0.1234	9023365	Palmier 13
155	Zineb	Medium	0.61	3364244	502 logts 13 n4
157	Deena	Low	0.2	7091232	Annaba 69
300	Assala	High	0.97	4112533	
301	Alae	High	0.92	5437898	
302	Raid	High	0.78		Abu Dhabi
303	Med	Low	0.2	3248673	Algeria 1280/12
304	Sarah	Medium	0.67	231987253	

Table 2: M Relation.

ID	Mark	Year
120	17	2015
141	10	2015
141	18	1014
154	9	2014
155	13	2014
155	4	2015
157	5	2015
300	10	2014
300	19	2015
301	17	2014
301	12	2014
302	18	2015
302	6	2014
303	9	2015
304	11	2014
304	13	2015

– $M(ID, mark, year) \leftarrow Mark(ID, mark, year)$

The areas of S and M are expressed in datalog–like notation:

– $S_1(ID, name, level, exam, phone, address) \leftarrow S(ID, name, level, exam, phone, address) \wedge ID < 300$

– $S_2(ID, name, level, exam, phone, address) \leftarrow S(ID, name, level, exam, phone, address) \wedge ID \geq 300$

– $M_1(ID, mark, year) \leftarrow M(ID, mark, year) \wedge y < 2014$

– $M_2(ID, mark, year) \leftarrow M(ID, mark, year) \wedge y \geq 2014 \wedge ID < 300$

– $M_3(ID, mark, year) \leftarrow M(ID, mark, year) \wedge y \geq 2014 \wedge ID \geq 300$

The substitution of S and M by their expressions results in the area definition in terms of the global schema:

– $S_1(ID, name, level, exam, phone, address) \leftarrow Student(ID, name, level, exam, phone, address) \wedge city = \text{“}Al - Ain'' \wedge ID < 300$

– $S_2(ID, name, level, exam, phone, address) \leftarrow Student(ID, name, level, exam, phone, address) \wedge city = \text{“}Al - Ain'' \wedge ID \geq 300$

– $M_1(ID, mark, year) \leftarrow Mark(ID, y, m) \wedge y < 2014$

– $M_2(ID, mark, year) \leftarrow Mark(ID, y, m) \wedge y \geq 2014 \wedge ID < 300$

– $M_3(ID, mark, year) \leftarrow Mark(ID, y, m) \wedge y \geq 2014 \wedge ID \geq 300$

Suppose the user query Q in datalog–like notation:

– $Q(ID, name, year, mark) \leftarrow S(ID, name, level, exam, phone, address, city) \wedge M(ID, mark, year) \wedge y = 2015$

The output of query Q (extracted from the S and M relations) is given in Table 4.
The query rewriting using S_1, S_2, M_1, M_2, and M_3 creates the following buckets: $Buck(S) = \{S_1, S_2\}$ and $Buck(M) = \{M_2, M_3\}$. The area M_1 is omitted in $Buck(M)$ because it does not satisfy the Q predicate (“$year = 2015$” in Q and “$year < 2014$” in M_1).
Then the query rewritings are produced, considering one area for each bucket:

– $QR_1(ID, year, name, mark) \leftarrow S_1(ID, name, level, exam, phone, address) \wedge M_2(ID, mark, year) \wedge y = 2015$

Table 3: Student Relation Fragmentation.

KID	ID	Name	Level	Exam	Phone	Address
120	120	Rani	High	1	6001104	Zakhir 12
123	123	Nada	Medium	0.465	99628734	Chiab al Alashekhar
141	141	Areej	Low	0.987		Al–Ain
154	154	Hanan	Low	0.1234	9023365	Palmier 13
155	155	Zineb	Medium	0.61	3364244	502 logts 13 n4
157	157	Deena	Low	0.2	7091232	Annaba 69
300	300	Assala	High	0.97	4112533	
301	301	Alae	High	0.92	5437898	
302	302	Raid	High	0.78		Abu Dhabi
303	303	Med	Low	0.2	3248673	Algeria 1280/12
304	304	Sarah	Medium	0.67	231987253	

Table 4: Query Result.

ID	Name	Year	Mark
120	Rani	2015	17
141	Areej	2015	10
155	Zineb	2015	4
157	Deena	2015	5
300	Assala	2015	19
302	Raid	2015	18
303	Med	2015	9
304	Sarah	2015	13

– $QR_2(ID, year, name, mark) \leftarrow S_2(ID, name, level, exam, phone, address) \wedge M_3(ID, mark, year) \wedge y = 2015$

Because the areas, S_1 and M_3 are contradictory ("$ID \geq 300$" and "$ID < 300$"), they are not joined; the same applies for areas S_2 and M_2. Consequently, $Q \supseteq QR_1 \bigcup QR_2$.
Let us focus on the rewriting of QR_1:
QR_1 and QR_2 areas and subareas are defined as creations of a unique area for each rewriting of QR_1 and QR_2 respectively, because the rewriting of QR_1 combines areas S_1(its subareas are S_{11}, S_{12}) and M_2(its subarea is M_{21}). The intersection of subarea S_{11} and M_{21} attributes with the rewriting attributes results in two subareas QR_{11} and QR_{12}; the join attribute is isolated in subarea QR_{13}. In this case there is no projection of the subarea S_{12} attribute. QR_2 subareas are defined similarly. The final result is:

– Area $QR_1\{ID < 300 \wedge year = 2015\}$; ? tuples; $keyaccuracy$ =?; inputs: S_1, M_2
  – Subarea $QR_{11}\{name\}$; $accuracy$ =?; input: S_{11}
  – Subarea $QR_{12}\{year, mark\}$; $accuracy$ =?; input: M_{21}
  – Subarea $QR_{13}\{ID\}$; $accuracy$ =?; inputs: S_{11} and M_{21}
– Area $QR_2\{ID \geq 300 \wedge year = 2015\}$; ? tuples; $keyaccuracy$ =?; inputs: S_2, M_3
  – Subarea $QR_{21}\{name\}$; $accuracy$ =?; input: S_{22}
  – Subarea $QR_{22}\{year, mark\}$; $accuracy$ =?; input: M_{31}
  – Subarea $QR_{23}\{ID\}$; $accuracy$ =?; inputs: S_{21}, M_{31}

After the estimation of semantic accuracy of subareas, we obtain

– Area $QR_1\{ID < 300 \wedge year = 2015\}$; ? tuples; $keyaccuracy = 0.25$ (0.50×0.50); inputs: S, M_2
  – Subarea $QR_{11}\{name\}$; $accuracy = 0.25 = (averageof(0.50 \times 0.50))$; input: S_{11}
  – Subarea $QR_{12}\{year, mark\}$; $accuracy = 0.25 = 0.50 \times 0.50$; input: M_{21}
  – Subarea $QR_{13}\{ID\}$; $accuracy = 0.25(average(0.50 \times 0.50, 0.50 \times 0.50)$; input: S_{11}, M_{21}
– Area $QR_2\{ID \geq 300 \wedge year = 2015\}$; ? tuples; $keyaccuracy = 0.93(1.00 \times 0.93)$; inputs: S_2, M_3
  – Subarea $QR_{21}\{name\}$; $accuracy = 0.74(0.80 \times 0.93)$; input: S_{22}
  – Subarea $QR_{22}\{year, mark\}$; $accuracy = 0.93(0.93 \times 1.00)$; input: M_{31}
  – Subarea $QR_{23}\{ID\}$; $accuracy = 0.84(average 0.8 \times 0.93, 0.93 \times 1.00)$; inputs: S_{21}, M_{31}

Because subareas QR_{11}, QR_{12}, QR_{13} possess the same accuracy, they are merged to subarea QR_{11}.

- Area $QR_1\,\{ID < 300 \wedge year = 2015\}$; ? tuples; $keyaccuracy = 0.25$; inputs: S_1, M_2
  - Subarea $QR_{11}\,\{ID, name, year, mark\}$; $accuracy = 0.25$
- Area $QR_2\,\{ID \geq 300 \wedge year = 2015\}$; ? tuples; $keyaccuracy = 0.93$; inputs: S_2, M_3
  - Subarea $QR_{21}\,\{name\}$; $accuracy = 0.74$; input: S_{22}
  - Subarea $QR_{22}\,\{year, mark\}$; $accuracy = 0.93$; input: M_{31}
  - Subarea $QR_{23}\,\{ID\}$; $accuracy = 0.84$; inputs: S_{21} and M_{31}

Remember that the rewriting accuracy is the average of the cell accuracies (accuracy aggregation). In this manner, the QR_1 accuracy is 0.25 and the QR_2 (which represents the accuracy value of its unique subarea) accuracy is $0.86(0.74 \times 1 + 0.93 \times 2 + 1.00 \times 3)$.
Thus, the QR_1 and QR_2 selectivity is calculated as $\frac{1}{9}(\frac{4}{6\times 6})$ and $\frac{1}{10}(\frac{4}{5\times 8})$ respectively. Subsequently, the partition metadata is

- Area $QR_1\,\{ID < 300 \wedge year = 2015\}$; $\frac{1}{9}(6\times 6) = 4$ tuples; $keyaccuracy = 0.25$; inputs: S_1, M_2
  - Subarea $QR_{11}\,\{ID, name, year, mark\}$; $accuracy = 0.25$
- Area $QR_2\,\{ID \geq 300 \wedge year = 2015\}$; $4(= 5 \times 8 \times 1/10)$ tuples; $keyaccuracy = 0.93$; inputs: S_2, M_3
  - Subarea $QR_{21}\,\{name\}$; $accuracy = 0.74$
  - Subarea $QR_{22}\,\{year, mark\}$; $accuracy = 0.93$
  - Subarea $QR_{23}\,\{ID\}$; $accuracy = 0.84$

The global user query Q accuracy value is estimated as the sum of the weighted and rewriting accuracies $(weight = tuplenumbers)$; thus, we obtain $(4\times 0.25 + 4\times 0.93)/(4+4) = 0.59$.
Figure 6 shows the quality graph for user query Q defined by merging its rewritings QR_1 and QR_2.

In the following section, we will discuss some data accuracy improvements by presenting a possible approach, which discards the areas or subareas causing exaggeration in accuracy prediction.

5 Accuracy improvement

To enforce the data accuracy, some fundamental amendment actions can be taken if the user expectations on data accuracy cannot be achieved. Because the proposed estimation approach arranges the query outputs in subareas possessing uniform accuracy, there is no guarantee that the

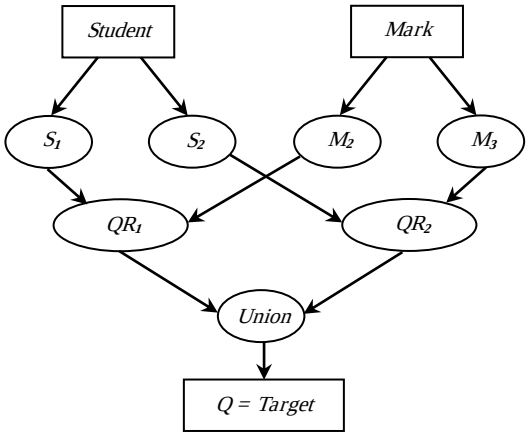


Figure 6: Quality Graph.

total cells in the query output satisfy the target expectations. However, the delivered data, in general, satisfy the user target.

We propose a simple improvement action consisting of discarding pieces of the query output with small accuracy values. We will present different ways and different times of executing such improvement actions. Particularly, we distinguish three accuracy levels:

- *Cell level*: Accuracy of cell set, in general, must be less than a specific threshold.
- *Tuple level*: Accuracy of tuple set, in general, must be less than a specific threshold.
- *Output level*: Accuracy of the whole output, must be less than a specific threshold.

The *cell level* is the most constraining. It reflects the fact that the user accepts only tuples containing accurate values. Moreover, the accuracy restrictions may be related to specified attributes; as an example, the user requires accurate phone numbers and disregards the accuracy of other attributes. For this type of restriction, we suggest filtering subareas with small accuracy values because the query output is partitioned to subareas with uniform accuracy. If the restriction concerns only specific attributes, filtering concerns only the subareas involving those attributes. Such action can be performed in the initial evaluation phase, exactly during insertion of areas in the buckets. We call this improvement action selective rewriting.

The *tuple level* accepts the existence of a group of attributes having low accuracy with the condition that the aggregate accuracy of the tuple is sufficiently high. Some users may tolerate acquiring tuples containing inaccuracies in certain attributes as long as the remaining cells are correct (e.g., if different attributes hold possible manner to contact clients such as phone, address, and email). Therefore, mistakes in some attributes are tolerable if the remaining attributes have high accuracy values. In this case, we

suggest area filtering rather than subarea filtering; in other words, the accuracy value is the aggregation of subarea accuracy, which is then compared to user expectations. Note that area filtering should be performed after rewriting computation (contrary to subarea filtering), as the aggregation is communicated to several subareas belonging to different database relations.

Finally, the *output level* does not consider the tuple or attribute accuracy; however, it signifies that the final output must reach a specific accuracy level. Remember that query outputs containing a mixture of tuples having high accuracy and tuples having very low accuracy are tolerable. This reflects people requesting all possible output data, without ignoring accuracy (e.g., the user sends publicity message to clients, accepting up to 6% undelivered message because of mistakes in contact attribute). To satisfy this constraint type, we must generate all possible data without considering their accuracy. To this end, we suggest sorting areas by their accuracies, aggregate the accuracy values gradually, and then stop if the correctness constraints are unsatisfied. Additionally, the data delivery can be incremental, allowing the users to stop when they obtain the necessary data or data containing many mistakes. This approach is executed only after aggregating the accuracy of all rewritings.

Therefore, three improvement actions are proposed:

- *Selective rewriting* is the generation of query rewritings that achieve a specific quality level, particularly "all subarea accuracy should be greater than the threshold." In this case, the rewriting algorithm can be modified to exclude from buckets the areas related to subareas having low accuracy. As the subarea accuracy was precomputed during the fragmentation process, this strategy can be implemented easily.
- *Rewriting filtering* can be implemented directly during the aggregation of rewriting accuracy.
- *Incremental data delivery*. After the aggregation of rewriting accuracy, we suggest sorting them according to descending order of accuracies. The algorithm is shown in Figure 7. The inputs of the algorithm are the expected values of accuracy and a list of areas in the rewritings; the output is a sorted area list. The list represents the data that satisfy user expectations.

Note that if the constraints are more restrictive, the output size will be small. It is worthy to mention that we should have a balance between accuracy and completeness expectations to deliver useful and valuable tuples to the user and avoid filtering too much data.

Figure 8 shows an example that illustrates the proposed improvement actions. Assume that QR_1, QR_2, QR_3, and QR_4 represent four rewritings. The number inside each subarea indicates the accuracy value of that subarea and the aggregated accuracy value for each area is written in front of QR_i.

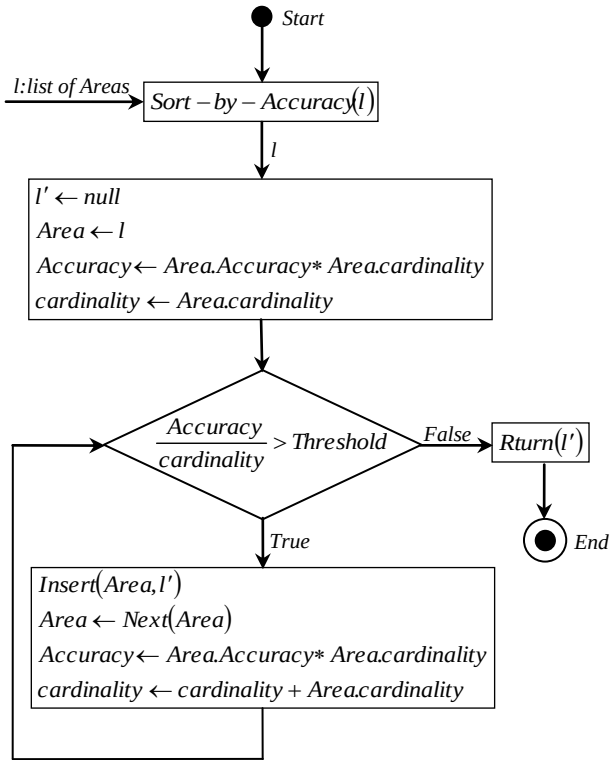


Figure 7: Incremental Data Delivery.

Assume also that C_1, C_2, and C_3 describe three constraints as follows:

- C_1: All subarea accuracy values must be more than or equal to 0.8.
- C_2: All area accuracy values must be more than 0.8.
- C_3: The query accuracy value must be more than 0.7.

Only QR_3 satisfies constraint C_1. Thus, the algorithm will generate only QR_3. Furthermore, QR_4 will be discarded because one of its subareas S_{41} does not satisfy the constraint. However, QR_3 and QR_4 satisfy constraint C_2. Despite the fact that subarea S_{41} possesses a low accuracy value than 0.8, the whole area accuracy is accepted.
To check constraint C_3, we sort QR_i values by accuracy. Thus, we obtain this order: $QR_3 \rightarrow QR_4 \rightarrow QR_2 \rightarrow QR_1$. We compute the aggregate accuracy gradually:

- $\{QR_3\}$ accuracy is 0.9.
- $\{QR_3, QR_4\}$ accuracy is $(0.9 \times 20 + 0.8 \times 10)/(20 + 10) = 0.86$.
- $\{QR_3, QR_4, QR_2\}$ accuracy is $(0.9 \times 20 + 0.8 \times 10 + 0.6 \times 10)/(20 + 10 + 10) = 0.8$.
- $\{QR_3, QR_4, QR_2, QR_1\}$ accuracy is $(0.9 \times 20 + 0.8 \times 10 + 0.6 \times 10 + 0.5 \times 20)/(20 + 10 + 10 + 20) = 0.7$. Because constraint C_3 is not satisfied, QR_1 is discarded.

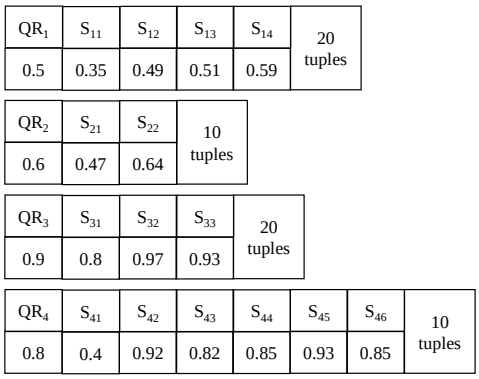


Figure 8: Rewritings Filtering.

6 Conclusion

This study investigates the data accuracy estimation and proposes some improvement actions. We suggested the fragmentation of database relations according to data accuracy by applying Rakov's algorithm and the projection of such fragments to query outputs to report the inaccuracy distribution in a better way. We rewrote the user query according to the fragments, and then we aggregated the accuracy values for the rewritings. We defined the query output by uniting the output tuples of rewritings. A basic algorithm to estimate the data accuracy was presented. In contrast to the existing approaches, the areas with low accuracy are identified explicitly by our proposed algorithm, so that data unsatisfying user expectations are discarded.

Furthermore, we suggested three primary improvement actions in order to filter data having lower accuracy to meet different types of user expectation accuracy. This approach is applicable in multiple stages of DIS development (design, production, or maintenance) to inform users about data accuracy, compare databases, check expectation satisfaction, or analyze enforcement actions for improving data accuracy.

Our objective in the near future is the development of a prototype implementing the proposed accuracy evaluation algorithm and the illustration of its usage in real–world applications. For this purpose, the prototype should display and edit the different components (such as databases, quality graph, characteristics, accuracy values, etc.) of the framework in addition to the execution of the quality evaluation algorithms. Furthermore, the prototype should evaluate the data accuracy in different applications for validation purposes by describing several tests in order to assess the approach performance and limitations. Finally, we suggest to improve some features of the quality evaluation tool and perform additional performance tests. In addition, we aim to analyze further quality factors and their inter–relationships.

References

[1] Abdalla, H., and Artoli, A. M. Towards an efficient data fragmentation, allocation, and clustering approach in a distributed environment. *Information*, 10(3), 112, 2019. DOI: 10.3390/info10030112

[2] Agrawal, S., Narasayya, V., and Yang, B. Integrating vertical and horizontal partitioning into automated physical database design. *In Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 359–370, 2004. DOI: 10.1145/1007568.1007609

[3] Alsaif, S. A., and Hidri, A. Impact of Data Balancing During Training for Best Predictions. *Informatica*, 45(2), 2021. DOI: 10.31449/inf.v45i2.3479

[4] Azeroual, O., Ershadi, M. J., Azizi, A., Banihashemi, M., and Abadi, R. E. (2021). Data Quality Strategy Selection in CRIS: Using a Hybrid Method of SWOT and BWM. *Informatica*, 45(1), 2021. DOI: 10.31449/inf.v45i1.2995

[5] Bai, Q., Hong, J., and McTear, M. F. Some modifications of bucket-based algorithms for query rewriting using views. *In International Conference on Advances in Information Systems*, Springer, Berlin, Heidelberg, 57–67, 2004. URL: https://rb.gy/leqpjc

[6] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. Classification and Regression Trees, *: Wadsworth*, 368, 1984. DOI: 10.1201/9781315139470-8

[7] Calvanese, D., Lembo, D., and Lenzerini, M. Survey on methods for query rewriting and query answering using views. *Integrazione, Warehousing e Mining di sorgenti eterogenee*, 25, 2001. URL: https://rb.gy/05cken

[8] Ezéchiel, K. K., Kant, S., and Agarwal, R. A systematic review on distributed databases systems and their techniques. *Journal of Theoretical and Applied Information Technology*, 96(1), 236–266, 2019. URL: https://rb.gy/gs00qk

[9] Fuaad, H. A., Ibrahim, A. A., Majed, A., and Asem, A. A Survey on Distributed Database Fragmentation Allocation and Replication Algorithms. *Current Journal of Applied Science and Technology*, 27(2), 1–12, 2018. DOI: 10.9734/CJAST/2018/37079

[10] Gao, J., Li, X., Xu, Y. E., Sisman, B., Dong, X. L., and Yang, J. Efficient knowledge graph accuracy evaluation. *arXiv preprint arXiv:1907.09657*, 2019. DOI: 10.48550/arXiv.1907.09657

[11] Halevy, A. Y. Answering queries using views: A survey. *The VLDB Journal*, 10(4), 270–294, 2002. DOI: 10.1007/s007780100054

[12] Hill, G. A Framework for valuing the quality of Customer Information *Doctoral dissertation, Melbourne University*, 1–194, 2009. URL: http://hdl.handle.net/11343/35567

[13] Levy, A., Rajaraman, A., and Ordille, J. Querying heterogeneous information sources using source descriptions. *In Proceedings of 22th International Conference on Very Large Data Bases*, 1, 1–26, 1996. DOI: 10.1049/tpe.1981.0030

[14] Ma, H., Noack, R., Schewe, K. D., and Thalheim, B. Using meta–structures in database design. *Informatica*, 34(3), 387—403, 2010. https://rb.gy/cbonrz

[15] Motro, A., and Rakov, I. Estimating the quality of databases. *In International Conference on Flexible Query Answering Systems*, Springer, Berlin, Heidelberg, 298–307, 1998. URL: https://rb.gy/iriobk

[16] Naumann, F., Leser, U., and Freytag, J. C. Quality–driven integration of heterogeneous information systems. *In Proceedings of the International Conference on Very Large Data Bases*, 447–458, 2005. URL: https://rb.gy/bgzhyn

[17] Özsu, M. T., and Valduriez, P. Principles of Distributed Database Systems, *Management, Springer, New York*, 12, 657–722, 2011. DOI: 10.1007/978-1-4419-8834-8

[18] Peng, P., Zou, L., Chen, L., and Zhao, D. Adaptive distributed RDF graph fragmentation and allocation based on query workload. *IEEE Transactions on Knowledge and Data Engineering*, 31(4), 670–685, 2018. DOI: 10.1109/TKDE.2018.2841389

[19] Rakov, I. Data quality and its use for reconciling inconsistencies in multidatabase environment.*PhD dissertation, George Mason University*, 1998.

[20] Raman, V., Swart, G., Qiao, L., Reiss, F., Dialani, V., Kossmann, D., Narang, I. and Sidle, R. Constant–time query processing. *In 2008 IEEE 24th International Conference on Data Engineering*, 60–69, 2008. DOI: 10.1109/ICDE.2008.4497414

[21] Riezler, S., and Liu, Y. Query rewriting using monolingual statistical machine translation. *Computational Linguistics*, 36(3), 569–582, 2010. DOI: 10.1162/coli_a_00010

[22] Silberschatz, A., Korth, H. F., and Sudarshan, S. Database system concepts. *McGraw-Hill, New York*, 5, 2002. URL: https://rb.gy/ttrvjt

[23] Stvilia, B. A workbench for information quality evaluation, *In Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital libraries*, 469–469, 2008. DOI: 10.1145/1378889.1379014

[24] Tarun, S., Batth, R. S., and Kaur, S. A Review on Fragmentation, Allocation and Replication in Distributed Database Systems. *In International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 538–544, 2019. DOI: 10.1109/ICCIKE47802.2019.9004233

[25] Várkonyi, G. G., and Gradišek, A. Data protection impact assessment case study for a research project using artificial intelligence on patient data. *Informatica*, 44(4), 2020. DOI: 10.31449/inf.v44i4.3253

[26] Wang, R. Y., and Strong, D. M. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4), 5–33, 1996. DOI: 10.1080/07421222.1996.11518099

[27] Zozus, M. N., Pieper, C., Johnson, C. M., Johnson, T. R., Franklin, A., Smith, J., and Zhang, J. Factors affecting accuracy of data abstracted from medical records. *PloS one*, 10(10), e0138649, 2015. DOI: 10.1371/journal.pone.0138649

A Novel Term Weighting Scheme for Imbalanced Text Classification

Tanapon Tantisripreecha[1] and Nuanwan Soonthornphisaj[2*]
E-mail: tanapon.tan@mahidol.ac.th, nuanwan.s@ku.th
[1]Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand
[2]Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand

Keywords: text classification, imbalance problem, term weighting schemes, TFIDF, SVM, logistic regression

Received: April 29, 2021

High dimensional feature is the main problem of text domain. If imbalance class is also found in the context, the classifier's performance is worsen. Moreover, solving imbalance problem by oversampling method in this circumstance is very difficult to get performance improvement. In this paper, a new term weighting scheme is proposed by combining Term frequency with an average of inverse document frequency factor. We denoted our scheme by TFmeanIDF. Our proposed method has high potential for imbalance text domain with high dimension. No feature selection or oversampling method is required. Extensive comparison results on 7 datasets validate the advantages of TFmeanIDF in terms of F_1 score obtained from widely used base classifiers, such as Logistic regression and Support Vector Machines. We found that F_1 score of minority class is higher than that of baseline term weighting schemes. Using TFmeanIDF as a term weighting shows promising result for logistics regression and support vector machines.

Povzetek: Avtorji so razvili novo shemo uteževanja termov pri neuravnoteženi klasifikaciji besedil.

1 Introduction

Learning from extreme imbalanced data is a challenging problem in real-world applications. We know that machine learning algorithms struggle with accuracy because of the unequal distribution of classes. This causes the performance of existing classifiers to get biased towards majority class. In this paper, we focus our attention on finding a new term weighting scheme that show promising result on imbalance text classification.

In text classification problem, we found that most well-known classifiers give high accuracy with balanced dataset. However, when apply to an imbalanced dataset, accuracy cannot be used to evaluate the model since the model often predicts as a majority class and rarely predict the minority class. Although the classifier predicts minority class incorrectly, the accuracy is still high because there are few instances of minority class.

Customer feedback on e-commerce platform such as Amazon is an example of imbalanced data set. The negative feedback is important for product owners, hence we consider these feedbacks as the positive class. Although the rating scales are provided for users, most customers do not rate their products. Therefore, the customer feedback in terms of textual data is our main focus. We found that the textual feedback of Amazon products is highly imbalanced. Most reviews are rated as positive feedback and is considered as a majority class whereas the negative feedback is considered as a minority class. The imbalance ratios of Amazon product reviews are quite high (vary from 5:1 to 15:1).

The rest of this paper is structured as follows: Section 2 reviews various studies on imbalance text classification problem; Section 3 presents our propose term weighting scheme; Section 4 presents the experimental results for the classification using various term weighting schemes and Section 5 provides the discussion and conclusions, as well as the directions for future work.

2 Related works

Given a document, a term weighting scheme is normally applied to represent the numerical text vector. Term frequency (TF) is a traditional term weighting technique for text vectorization that considers the frequency of the term found in the document [1]. Moreover, using TF may give a large weight to common terms that would weak the text discriminating ability and leads to adverse impact on the classification performance [2]. To alleviate this problem, a collection frequency factor IDF (inverse document frequency) is introduced into the TF scheme called TFIDF [1]. Considering IDF value, the more frequent documents containing the term, the lower the IDF score. This means that the word is less important, when they occur in several documents. The less a term occurs in different documents, the more weight it is given. We found that the assumption of TF is that multiple

* Corresponding author

occurrences of a term in a text are more important than single occurrences. Secondly, IDF assumes that rare terms are more significant than frequent terms. Thirdly, the normalization assumption which reduces the bias of the document length. We found that TFIDF is considered as a baseline term weighting method in text classification domain [3]

In order to reduce the effect of term with high local weighting factor TF, meanwhile the importance of term with low DF can be further enhanced. Tang, et al [4] replaced the IDF factor with a novel function (or new global weighting factor) that possesses the following two properties: (i) when the DF of the term changes, the value of the function has a larger attenuation rate, (ii) it should be a bounded function. Since the attenuation rate of exponential function is faster than the logarithmic function so inverse exponential frequency is chosen to construct the new function. Their term weighting scheme is called TFIEF. Another unsupervised term weighting method is called TFIHF that incorporate the nonlinear transformation functions. The hyperbolic function contributes as a global weighting factor namely inverse hyperbolic frequency [5].

To eliminate the influence of document length, Chen, et al. [2] proposed a new term frequency factor called term density function to normalize the TF. Their experiments showed promising result on 4 real world datasets which are Amazon (vocabulary size = 10,000 words), Movie Review (vocabulary size = 7,103 words), WebKB (vocabulary size = 8,791 words), and 20 Newsgroup (vocabulary size = 10,000 words). We found that the vocabulary size of these datasets is quite small (less than 20,746 words) and no imbalance problem found in these domains.

For imbalance text classification problem, the study of supervised term weighting schemes is also found in the literature. Lan et al [6] introduced the weighting scheme called TFRF. Their assumption is that if a high-frequency term is more concentrated in the positive class than in the negative class, it will make more contributions in selecting the positive samples from the negative samples. This idea of favouring positive terms can also be observed in [7], which tried to select such "positive" words using the square root of chi-square, as opposed to "negative" words indicative of negative documents.

We found that several term weighting schemes work well on balanced dataset, however, when applied to an imbalanced dataset, they gave high precision but low recall. Various methods to handle imbalance problem can be classified in one of the following categories: under sampling, oversampling, synthetic data generation and cost sensitive learning. Under sampling method randomly reduces the number of majority class instances to make the dataset balanced. This method is best to use when the data set is huge and reducing the number of training samples helps to improve run time and storage troubles.

Oversampling method randomly replicates the minority class instances to balance the data. An advantage of oversampling is that there is no information loss. However, the disadvantage of using this method is that it leads to overfitting problem.

Synthetic data generation method adds the minority class instances by generating artificial data. The well-known method namely SMOTE algorithm synthesis new instances based on feature space by using K nearest neighbours.

The concept of cost sensitive learning is to evaluate the cost associated with misclassifying instances [8]. This method does not create balanced data distribution. It deploys the cost matrices which represent the misclassification cost, cost of False Negative and cost of False Positive under the constraint that cost of False Negative is higher than that of False Positive. The goal of this method is to choose a classifier with lowest total cost.

3 Proposed term weight

To tackle the imbalanced text classification problem, we introduce a new supervised term weighting scheme namely *TFmeanIDF* that use class information of training documents. *TF* stands for term frequency whereas the *meanIDF* is calculated from the average of inverse document frequency of instances in majority and minority class. For each term, *i*, in Customer Feedback, *j*, *meanIDF* can be calculated using equation (1).

$$meanIDF_i = \frac{\left(\log\left(\frac{n_{major}}{df_{major,i}+1}\right)+1\right)+\left(\log\left(\frac{n_{minor}}{df_{minor,i}+1}\right)+1\right)}{2} \quad (1)$$

where

n_{major} = total number of customer feedbacks in majority class.

n_{minor} = total number of customer feedbacks in minority class.

$df_{major,i}$ = total number of customer feedbacks in majority class that contain term *i*.

$df_{minor,i}$ = total number of customer feedbacks in minority class that contain term *i*.

Finally, the feature of customer feedbacks, *j*, is obtained from the term frequency, tf_{ij} multiplied by the global weight, meanIDF, using the following equation.

$$TFmeanIDF = tf_{ij} * \frac{\left(\log\left(\frac{n_{major}}{df_{major,i}+1}\right)+1\right)+\left(\log\left(\frac{n_{minor}}{df_{minor,i}+1}\right)+1\right)}{2} \quad (2)$$

Where tf_{ij} = total number of occurrences of term *i* in customer feedbacks, *j*

Figure 1 illustrates the changing value of the traditional IDF and meanIDF when the number of documents in minority and majority class containing term *i* is varied. The X-axis represents the ratio of document frequency in minority to that of majority class. We found that the meanIDF reaches its minimum value when the document frequency ratio is equal to 1. That means the number of documents in minority class with the occurrence of term i is equal to the number of documents of majority class in which term *i* occur. It means that term, *i*, is less important.

From Figure 1, we found that the value of IDF is stable that means it has no feature selection power in text imbalance problem. The value of meanIDF is dynamic

when the document frequency ratio is changed. Therefore, incorporate the meanIDF as term weighting can increase the weight of important term in minority class and majority class as well. When the document frequency of term, i, likely occurs in both minority and majority class, that term becomes less important.

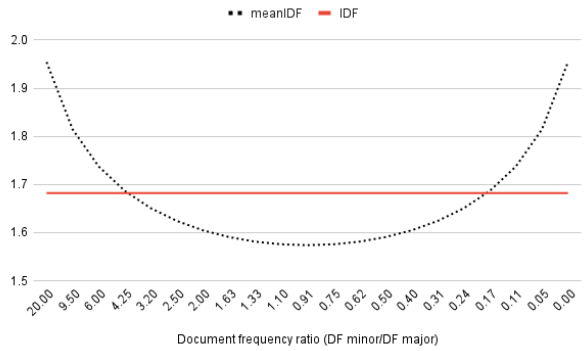


Figure 1: The behaviour of meanIDF and IDF under different document frequency ratio.

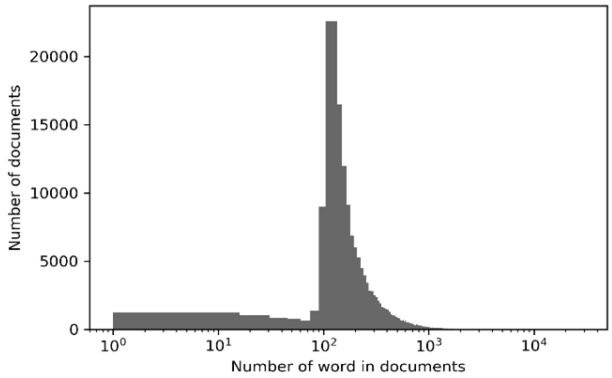


Figure 2: Frequency distribution of word occurrences found in customer feedback.

Product Category	Number of instances		Vocabulary Size	Imbalance Ratio (IR)
	Majority Class (Positive feedback)	Minority Class (Negative feedback)		
Instant Video	523,653	60,280	135,996	8.69
Musical Instruments	442,627	57,549	203,638	7.69
Digital Music	786,273	49,733	268,783	15.81
Baby	571,064	97,210	133,321	5.87
Patio Lawn and Garden	816,370	177,120	178,671	4.61
Automotive	1,187,496	186,272	214,752	6.38
Apps for Android	2,209,990	428,183	305,499	5.16

Table 1: Datasets used in our experiments.

4 Experiments and results

4.1 Dataset

Amazon customer feedback dataset obtained from seven product categories [9] are collected to validate the performance of our proposed method. The raw data consists of the product review text and the rating scales that represent the satisfaction levels. We consider the scales value 1-2 as the negative feedback whereas the scale 3-5 means positive feedback. Table 1 shows class distribution and vocabulary size of seven product categories which are Instant Video, Musical Instruments, Digital Music, Baby, Patio Lawn and Garden, Automotive and Apps for Android.

Note that the imbalance ratio (IR) is defined as follows:

$$\text{IR} = \frac{\#\text{instances in majority class}}{\#\text{instances in minority class}} \tag{3}$$

We found that all product categories are imbalanced. The highest imbalance ratio is found in Digital Music product with IR = 15.81. The lowest imbalance ratio is 4.61 found in Patio Lawn and Garden.

Amazon product feedback consists of different text lengths as shown in Figure 2. We visualize the histogram of text length in log scale (x-axis) It reveals that the minimum text length of product review is 1 (10^0) word and the maximum length is 30,000 ($3x10^4$) words. The average of text length is 360 ($3.6x10^2$) words and the median is 200 ($2x10^2$) words.

We do text preprocessing by removing HTML tags, converting text to lowercase, replacing punctuation with spaces, removing numbers, removing non English text and removing duplicate characters. Then tokenization is performed to separate the sentence into word tokens. Finally input vectors are prepared by text vectorization with different term weighting methods which are TFmeanIDF, TFIDF, TFRF, TFIEF and TFIHF using equation 2, 4, 5, 6 and 7, respectively.

$$TFIDF = tf_{ij} * \log\left(\frac{N}{df_i + 1}\right) + 1 \tag{4}$$

$$TFRF = tf_{ij} * \log_2\left(2 + \frac{A_{ij}}{\max(1, C_{ij})}\right) \tag{5}$$

$$TFIEF = tf_{ij} * e^{-\frac{df_i}{N}} \tag{6}$$

$$TFIHF = tf_{ij} * \frac{\frac{N}{df_i}}{1 + (\frac{N}{df_i})} \tag{7}$$

Where tf_{ij} = total number of occurrences of term i in j

df_i = total number of customer feedbacks containing term i.

A_i = total number of customer feedbacks containing term *i* in the positive class.

C_i = total number of customer feedbacks containing term *i* in the negative class.

4.2 Experimental results

We compare the performance of our term weighting method with several baseline approaches. In particular, supervised term weighting methods that considered the class label as a domain knowledge, and unsupervised term weighting methods. Three classifiers are experimented based on 5-fold cross validation. The selected classifiers are Logistic Regression, Support Vector Machine and multinomial Naïve Bayes. These classifiers are chosen since they are state of the art in sentiment analysis research [5, 10, 11, 12, 13, 14]. Therefore it is more challenging to see the potential of term weighting schemes in imbalanced text classification domain to see clearly whether *TFmeanIDF* can increase the performance or not.

Note that given the training data, the Logistic Regression algorithm calculates the event's probabilities and applies a logistic function to create the model [15]. Support Vector Machine (SVM) creates a decision boundary or hyperplane to predict the classes in high dimensional space [16]. Multinomial Naïve Bayes (MultinomialNB) creates the model using joint probabilistic distribution [17].

Product Category	TF mean IDF	TF Mean IDF+ over Sampling	TF IDF	TF IDF+ over Sampling	TF RF	TF IEF	TF IHF
Logistic Regression							
Instant Video	**0.69 (0.94)**	0.65 (0.90)	0.64 (0.93)	0.63 (0.89)	0.03 (0.89)	0.64 (0.93)	0.64 (0.93)
Musical Instruments	**0.69 (0.93)**	0.66 (0.89)	0.64 (0.93)	0.64 (0.89)	0.01 (0.88)	0.64 (0.92)	0.63 (0.92)
Digital Music	**0.61 (0.96)**	0.54 (0.91)	0.55 (0.95)	0.52 (0.90)	0.01 (0.94)	0.56 (0.95)	0.55 (0.95)
Baby	**0.72 (0.92)**	0.69 (0.88)	0.68 (0.91)	0.67 (0.87)	0.01 (0.85)	0.64 (0.92)	0.66 (0.91)
Patio Lawn and Garden	**0.74 (0.91)**	0.73 (0.88)	0.7 (0.90)	0.7 (0.87)	0.02 (0.82)	0.68 (0.9)	0.68 (0.9)
Automotive	**0.7 (0.92)**	0.67 (0.88)	0.65 (0.91)	0.65 (0.87)	0.01 (0.86)	0.64 (0.91)	0.64 (0.91)
Apps for Android	**0.74 (0.91)**	0.72 (0.88)	0.7 (0.91)	0.69 (0.87)	0.2 (0.85)	0.69 (0.91)	0.69 (0.91)
SVM							
Instant Video	**0.69 (0.941)**	0.65 (0.90)	0.65 (0.97)	0.62 (0.89)	0.07 (0.9)	0.62 (0.92)	0.62 (0.92)
Musical Instruments	**0.69 (0.93)**	0.65 (0.90)	0.65 (0.92)	0.63 (0.89)	0.02 (0.88)	0.62 (0.91)	0.62 (0.91)
Digital Music	**0.62 (0.96)**	0.54 (0.92)	0.57 (0.95)	0.51 (0.91)	0.01 (0.94)	0.54 (0.95)	0.54 (0.95)
Baby	**0.72 (0.92)**	0.69 (0.88)	0.68 (0.91)	0.67 (0.87)	0.01 (0.85)	0.62 (0.91)	0.65 (0.90)
Patio Lawn and Garden	**0.74 (0.91)**	0.72 (0.88)	0.7 (0.90)	0.69 (0.86)	0.05 (0.82)	0.67 (0.894)	0.67 (0.89)
Automotive	**0.7 (0.92)**	0.67 (0.88)	0.65 (0.91)	0.64 (0.83)	0.02 (0.86)	0.62 (0.91)	0.62 (0.91)
Apps for Android	**0.74 (0.91)**	0.71 (0.88)	0.69 (0.91)	0.68 (0.87)	0.39 (0.86)	0.69 (0.90)	0.69 (0.97)
Multinomial NB							
Instant Video	0.07 (0.9)	0.57 (0.86)	0.09 (0.90)	0.56 (0.86)	0.5 (0.92)	**0.62 (0.92)**	**0.62 (0.92)**
Musical Instruments	0.01 (0.88)	0.55 (0.84)	0.02 (0.88)	0.55 (0.84)	0.5 (0.91)	**0.61 (0.90)**	**0.61 (0.91)**
Digital Music	0.01 (0.94)	0.42 (0.86)	0.01 (0.94)	0.41 (0.86)	0.3 (0.94)	**0.5 (0.94)**	**0.5 (0.94)**
Baby	0.15 (0.86)	0.6 (0.83)	0.19 (0.86)	0.59 (0.83)	0.5 (0.88)	0.61 (0.90)	**0.62 (0.88)**
Patio Lawn and Garden	0.19 (0.83)	0.66 (0.84)	0.24 (0.84)	0.64 (0.83)	0.55 (0.87)	**0.67 (0.88)**	**0.67 (0.88)**
Automotive	0.07 (0.86)	0.61 (0.85)	0.12 (0.87)	0.6 (0.84)	0.43 (0.89)	**0.63 (0.90)**	**0.63 (0.90)**
Apps for Android	0.35 (0.86)	0.65 (0.84)	0.38 (0.87)	0.63 (0.84)	0.63 (0.88)	**0.67 (0.89)**	**0.67 (0.89)**

Table 2: F_1-score of minority class and (Micro-average) of various term weighting methods.

The performance of each model is evaluated using F_1-score as shown in equation 8.

$$F_1 = \frac{2*precision*recall}{precision+recall} \tag{8}$$

Where

$$precision = \frac{TP}{TP+FP} \tag{9}$$

$$recall = \frac{TP}{TP+FN} \tag{10}$$

Note that:

TP represents the number of customer feedbacks that are correctly classified as minority class.

FP represents the number of customer feedbacks that are incorrectly classified as minority class.

FN represents the number of customer feedbacks belonging to minority class but are incorrectly classified in the majority class.

Three experimental set ups are explored as follow:

A) Binary classification.

B) Multi-class classification.

C) Long text binary classification.

A) Binary classification

For binary classification, the class label is positive feedback and negative feedback.

The experimental results depicted in Table 2 confirm that the *TFmeanIDF* outperforms all selected supervised and unsupervised term weighting methods. For Logistic Regression and SVM, we found that *TFmeanIDF* outperforms other term weighting methods in all product categories. *TFmeanIDF* obtains the highest F_1 score at 0.74 on minority class of App for Android products.

Consider the highest imbalance ratio dataset which is Digital Music Product category (imbalance ratio =15.81), we found that using *TFmeanIDF* also gets the highest F_1 score at 0.61 and 0.69 on minority class for Logistic Regression and SVM respectively.

To validate whether the oversampling method can increase F_1 score of the baseline method (TFIDF), we found that random oversampling method has no contribution in this problem domain. Since the feature space (vocabulary size) is very high (vary from 133,321 to 305,499; see Table 1), no performance improvement obtained from the datasets.

The average of F1 score of minority classes in 7 product categories is illustrated in Figure 3. We found that Logistic Regression and SVM learned from TFmeanIDF gets the highest performance with the average F1 score at 0.70. Moreover, no performance loss on majority class since the micro-average of F1 score is 0.93 which is the highest among other term weighting schemes (See Figure 4 for details).

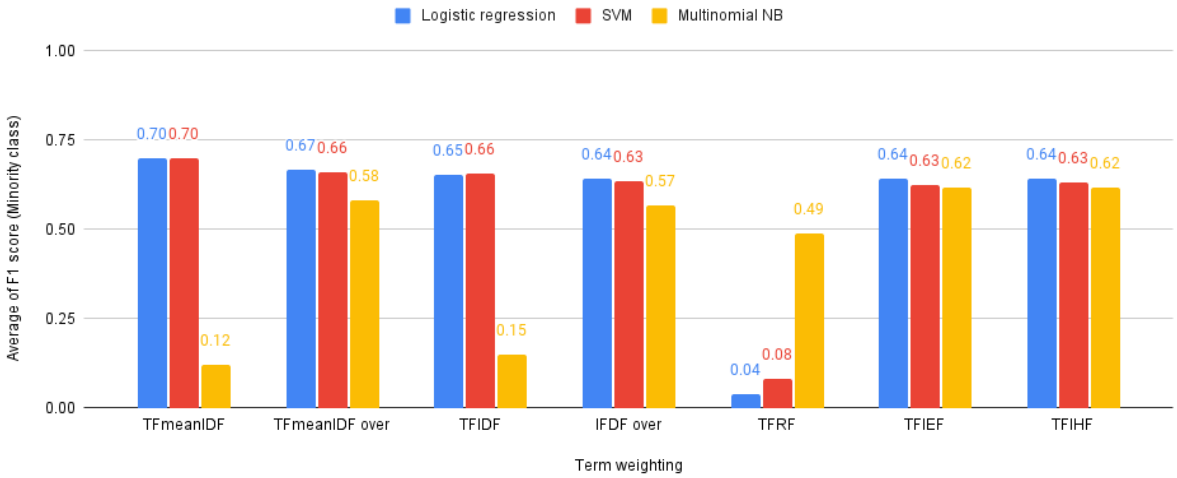


Figure 3: The average of F_1 score on Minority class from 7 products categories (Experiment A: binary classification).

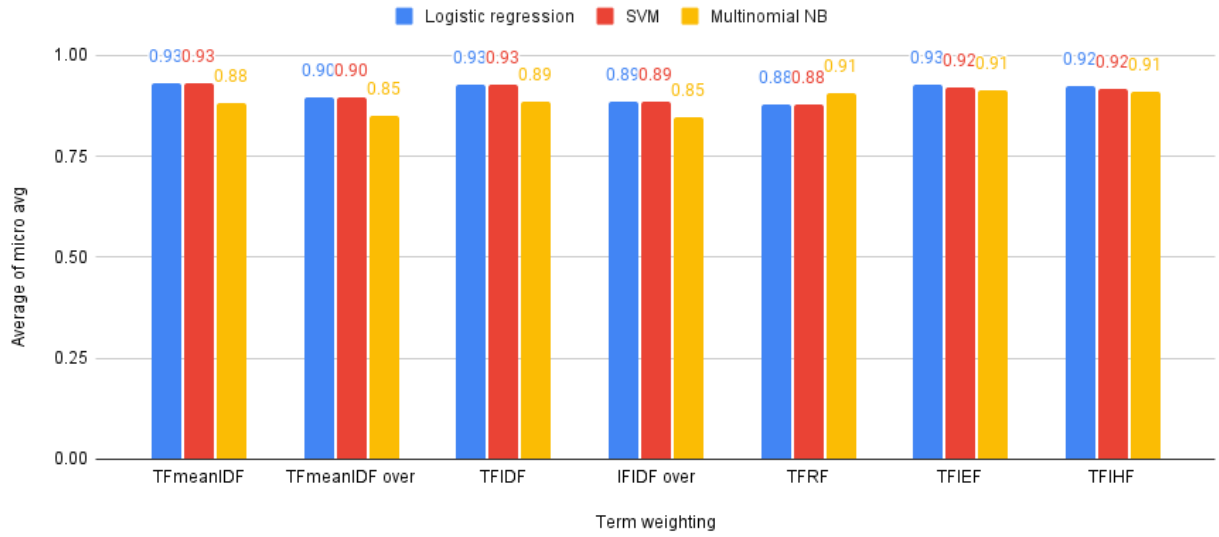


Figure 4: The average performance of 7 product categories measured in term of Micro average (F_1 score) (Experiment A: binary classification).

B) Multi-class classification

To see the potential of *TFmeanIDF* on multi-class classification, we consider the rating scales obtained from customers as the class labels. The range of satisfaction level starts from 1 to 5. (see Table 3).

The performance of multi-class classification with imbalance problem is normally measured in terms of the weighted average recall and weighted average F_1 as shown in equation (11).

$$Weight\_average\_recall = \frac{\sum_{i}^{n}(recall_i \times |c_i|)}{\sum_{i}^{n}|c_i|} \quad (11)$$

$$Weight\_average\_F1 = \frac{\sum_{i}^{n}(F1_i \times |c_i|)}{\sum_{i}^{n}|c_i|} \quad (12)$$

Where c_i is class i

n is number of classes

The result of multi-class classification shown in Table 4 confirms that the proposed term weighting method significantly outperforms all selected baselines supervised and unsupervised term weighting methods.

For Logistic Regression and SVM, we found that TFmeanIDF outperforms other term weighting methods in all product categories. TFmeanIDF obtains the highest

Product Category	Number of instances in C_i				
	1	2	3	4	5
Instant Video	36,785	23,495	79,349	137,840	306,465
Musical Instruments	34,931	22,618	38,537	93,306	310,784
Digital Music	29,988	19,745	41,344	122,479	622,450
Baby	58,913	38,297	12,557	94,749	463,758
Patio Lawn and Garden	119,633	57,487	80,891	174,356	561,123
Automotive	122,160	64,112	103,857	230,293	853,346
Apps for Android	294,284	133,899	253,549	561,831	1,394,611

Table 4: Data distribution for multi-class classification. problem.

weighted recall and weighted F1-score score at 0.78 and 0.73 respectively on the Digital Music.

In multi-class classification, the average of weight recall and weight F1-score illustrated in Figure 5 and 6 respectively. We found that Logistic Regression and SVM learned from TFmeanIDF gets the highest performance with the average weighted recall score at 0.71 and 0.70 respectively. Moreover, no performance loss on majority class since the weight-average of F_1score is 0.66 and 0.65 respectively which is the highest as well (See Figure 5).

C) Long text classification

To further explore the potential of *TFmeanIDF*, we set up another experiment for long text classification problem. This problem is the most difficult since we select only the long text as a test set by including customer feedbacks that contain more than 5000 words [18]. Note that imbalance

Product Category	TF mean IDF	TF Mean IDF+ over Sampling	TF IDF	TF IDF+ over Sampling	TF RF	TF IEF	TF IHF
Logistic Regression							
Instant Video	**0.73 (0.68)**	0.63 (0.66)	0.72 (0.67)	0.61 (0.64)	0.71 (0.66)	0.71 (0.66)	0.66 (0.52)
Musical Instruments	**0.70 (0.65)**	0.62 (0.64)	0.69 (0.63)	0.59 (0.62)	0.68 (0.63)	0.68 (0.63)	0.62 (0.48)
Digital Music	**0.78 (0.73)**	0.65 (0.68)	0.77 (0.72)	0.63 (0.67)	0.77 (0.72)	0.77 (0.72)	0.74 (0.64)
Baby	**0.70 (0.66)**	0.65 (0.66)	0.68 (0.64)	0.62 (0.64)	0.68 (0.63)	0.68 (0.63)	0.58 (0.43)
Patio Lawn and Garden	**0.68 (0.64)**	0.62 (0.63)	0.67 (0.61)	0.58 (0.61)	0.66 (0.61)	0.66 (0.61)	0.56 (0.41)
Automotive	**0.71 (0.66)**	0.66 (0.66)	0.70 (0.64)	0.68 (0.64)	0.69 (0.64)	0.69 (0.64)	0.62 (0.48)
Apps for Android	**0.64 (0.59)**	0.63 (0.58)	0.63 (0.58)	0.63 (0.58)	0.54 (0.39)	0.54 (0.39)	0.65 (0.61)
SVM							
Instant Video	**0.72 (0.68)**	0.62 (0.64)	0.71 (0.66)	0.6 (0.62)	0.69 (0.65)	0.69 (0.65)	0.65 (0.52)
Musical Instruments	**0.7 (0.65)**	0.6 (0.62)	0.68 (0.63)	0.57 (0.6)	0.66 (0.62)	0.66 (0.62)	0.62 (0.48)
Digital Music	**0.78 (0.73)**	0.61 (0.66)	0.77 (0.72)	0.59 (0.64)	0.75 (0.71)	0.75 (0.71)	0.74 (0.64)
Baby	**0.69 (0.66)**	0.63 (0.65)	0.68 (0.63)	0.6 (0.62)	0.66 (0.62)	0.66 (0.62)	0.58 (0.43)
Patio Lawn and Garden	**0.68 (0.63)**	0.6 (0.62)	0.67 (0.61)	0.56 (0.59)	0.65 (0.6)	0.65 (0.6)	0.56 (0.41)
Automotive	**0.71 (0.66)**	0.66 (0.66)	0.7 (0.64)	0.59 (0.6)	0.68 (0.63)	0.68 (0.63)	0.62 (0.48)
Apps for Android	**0.63 (0.56)**	0.63 (0.58)	0.62 (0.54)	0.62 (0.54)	0.53 (0.37)	0.53 (0.37)	0.64 (0.58)
Multinomial NB							
Instant Video	0.66 (0.53)	0.59 (0.63)	0.66 (0.53)	0.58 (0.61)	0.69 (0.65)	**0.69 (0.65)**	0.68 (0.58)
Musical Instruments	0.62 (0.48)	0.53 (0.57)	0.62 (0.48)	0.53 (0.56)	0.66 (0.61)	**0.66 (0.61)**	0.64 (0.52)
Digital Music	0.74 (0.64)	0.58 (0.63)	0.74 (0.64)	0.57 (0.62)	0.75 (0.71)	**0.75 (0.71)**	0.75 (0.66)
Baby	0.59 (0.46)	0.57 (0.6)	0.6 (0.46)	0.56 (0.58)	0.65 (0.62)	**0.65 (0.62)**	0.61 (0.5)
Patio Lawn and Garden	0.58 (0.45)	0.55 (0.57)	0.59 (0.46)	0.53 (0.56)	0.64 (0.6)	**0.64 (0.6)**	0.61 (0.5)
Automotive	0.63 (0.49)	0.63 (0.5)	0.63 (0.5)	0.62 (0.58)	0.67 (0.64)	**0.67 (0.64)**	0.64 (0.53)
Apps for Android	0.58 (0.46)	0.62 (0.58)	0.62 (0.58)	0.62 (0.58)	0.59 (0.5)	**0.59 (0.5)**	0.57 (0.45)

Table 3: Performance on Multi-class problem.

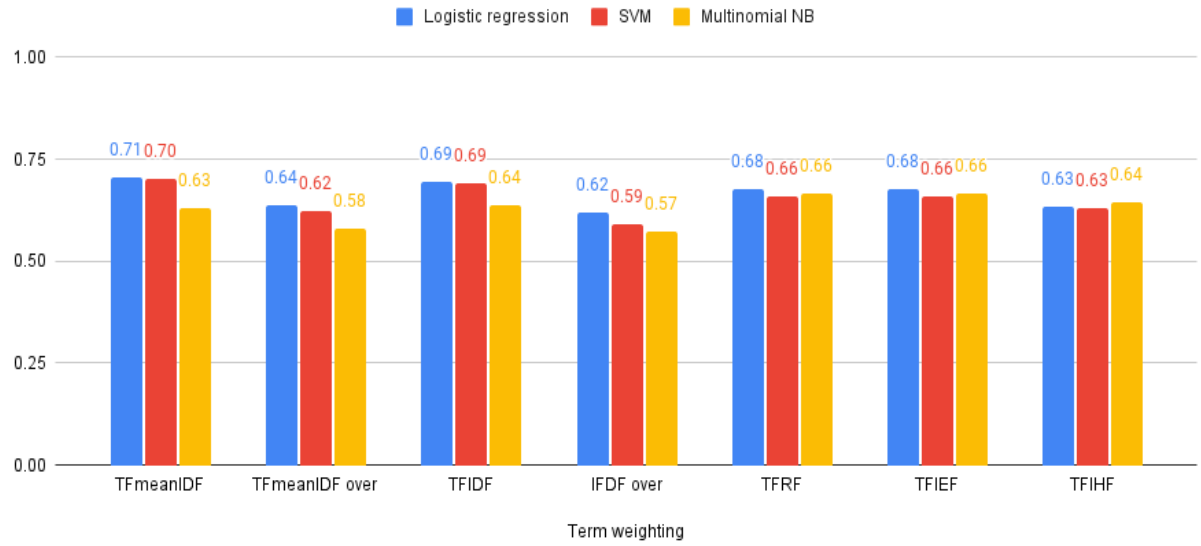


Figure 5: The average of weighted drecall obtained form 7 product categories (Experiment B: multi-class classification).

Product Category	Number of instances		Vocabulary Size	Imbalance Ratio (IR)
	Majority Class (Positive feedback)	Minority Class (Negative feedback)		
Instant Video	366,785	42,180	135,996	8.70
Musical Instruments	310,316	40,235	203,638	7.71
Digital Music	550,924	34,827	268,783	15.82
Baby	547,782	93,231	133,321	5.88
Patio Lawn and Garden	571,593	124,106	178,671	4.61
Automotive	831,179	130,595	214,752	6.36
Apps for Android	1,546,589	300,185	305,499	5.15

Table 6: The data distribution of long text feedbacks.

problem is presented in this experiment as well. The data distribution is shown in Table5. The performances of minority class for long text classification shown in Table 6 and Figure 8 reveal that SVM and Logistic Regression learning from *TFmeanIDF* are as good as those learning from TFIDF with the average F_1 score at 0.91. However, the average of F_1 score of minority class (Figure 7) obtain from 7 product categories is slightly dropped. SVM learning from *TFmeanIDF* gets the micro-average of F_1score at 0.60 where as SVM learning from TFIDF get the average of F_1 score at 0.62 on minority classes.

We found that TFRF cannot classify minority class instances in long text at all. This is the limitation of TFRF since it is supervised approach that needs the label information. As shown in equation [7], TFRF requires the value of number of customer feedbacks containing $term_i$ and the number of customer feedbacks containing $term_i$ (A_i, C_i). In this scenario, TFIDF is applied to create the test vectors [6]. Moreover this dataset is imbalance, the model learning from TFRF get the worst performance.

Product Category	TF mean IDF	TF Mean IDF+ over Sampling	TF IDF	TF IDF+ over Sampling	TF RF	TF IEF	TF IHF
Logistic Regression							
Instant Video	0.64 (0.89)	0.65 (0.82)	**0.7 (0.91)**	0.65 (0.83)	0 (0.81)	0.67 (0.89)	0.66 (0.88)
Musical Instruments	0.49 (0.96)	0.37 (0.88)	**0.53 (0.96)**	0.43 (0.91)	0 (0.94)	0.53 (0.94)	0.52 (0.93)
Digital Music	0.48 (0.97)	0.46 (0.93)	**0.53 (0.97)**	0.45 (0.94)	0 (0.97)	0.54 (0.97)	0.57 (0.97)
Baby	0.42 (0.88)	**0.51 (0.8)**	0.42 (0.88)	0.48 (0.79)	0 (0.87)	0.57 (0.87)	0.57 (0.87)
Patio Lawn and Garden	0.47 (0.87)	0.61 (0.85)	0.48 (0.87)	**0.63 (0.86)**	0 (0.85)	0.62 (0.89)	0.63 (0.89)
Automotive	**0.71 (0.93)**	0.57 (0.83)	0.69 (0.93)	0.51 (0.79)	0 (0.87)	0.75 (0.93)	0.73 (0.92)
Apps for Android	**0.77 (0.87)**	0.67 (0.72)	0.76 (0.87)	0.64 (0.7)	0.2 (0.7)	0.71 (0.83)	0.7 (0.83)
SVM							
Instant Video	0.65 (0.89)	0.65 (0.86)	0.68 (0.9)	**0.7 (0.89)**	0 (0.81)	0.56 (0.87)	0.56 (0.87)
Musical Instruments	**0.59 (0.96)**	0.42 (0.91)	0.56 (0.96)	0.43 (0.92)	0.14 (0.94)	0.41 (0.91)	0.37 (0.91)
Digital Music	0.5 (0.97)	0.45 (0.95)	0.57 (0.97)	0.49 (0.95)	0 (0.97)	**0.58 (0.97)**	0.56 (0.97)
Baby	0.47 (0.88)	0.49 (0.82)	0.52 (0.88)	0.52 (0.84)	0 (0.87)	0.55 (0.86)	**0.57 (0.86)**
Patio Lawn and Garden	0.45 (0.87)	**0.61 (0.87)**	0.51 (0.88)	0.59 (0.86)	0 (0.85)	0.59 (0.88)	0.58 (0.88)
Automotive	**0.81 (0.96)**	0.65 (0.88)	0.79 (0.95)	0.6 (0.86)	0 (0.87)	0.7 (0.91)	0.71 (0.91)
Apps for Android	**0.73 (0.85)**	0.65 (0.7)	0.71 (0.85)	0.59 (0.68)	0.6 (0.77)	0.67 (0.79)	0.68 (0.79)
Multinomial NB							
Instant Video	0 (0.8)	0.59 (0.76)	0 (0.8)	**0.82 (0.57)**	0.48 (0.64)	0.55 (0.75)	0.55 (0.75)
Musical Instruments	0 (0.94)	0.24 (0.86)	0.07 (0.94)	0.28 (0.88)	0.31 (0.89)	**0.34 (0.91)**	0.34 (0.91)
Digital Music	0 (0.97)	0.33 (0.89)	0 (0.97)	0.3 (0.88)	0.55 (0.96)	**0.55 (0.97)**	0.52 (0.96)
Baby	0.23 (0.87)	**0.43 (0.69)**	0.31 (0.87)	0.38 (0.68)	0.35 (0.69)	0.36 (0.7)	0.36 (0.7)
Patio Lawn and Garden	0.18 (0.86)	**0.5 (0.78)**	0.3 (0.87)	0.47 (0.78)	0.48 (0.8)	0.45 (0.79)	0.46 (0.79)
Automotive	0.29 (0.89)	0.39 (0.64)	0.43 (0.91)	0.37 (0.61)	**0.56 (0.82)**	0.5 (0.77)	0.49 (0.77)
Apps for Android	0.54 (0.77)	0.6 (0.62)	0.59 (0.79)	**0.61 (0.64)**	0.52 (0.45)	**0.61 (0.64)**	0.6 (0.64)

Table 5: Long text classification performance measured in terms of F_1-score of minority class and (micro-average of F_1 score).

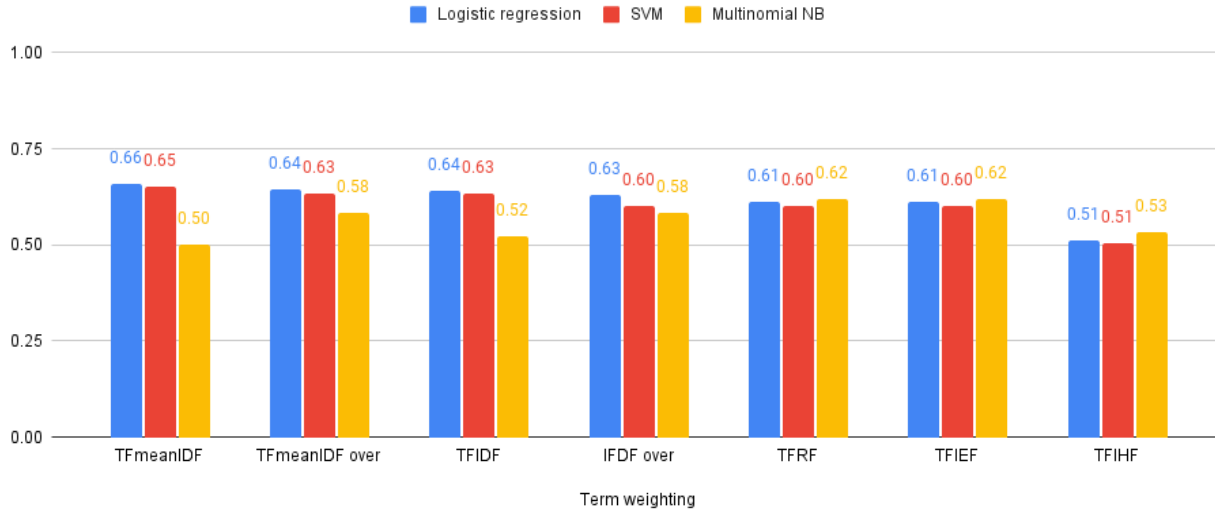


Figure 6: Weighted average of F1 score on multi-class classification problem (Experiment B: multi-class classification).

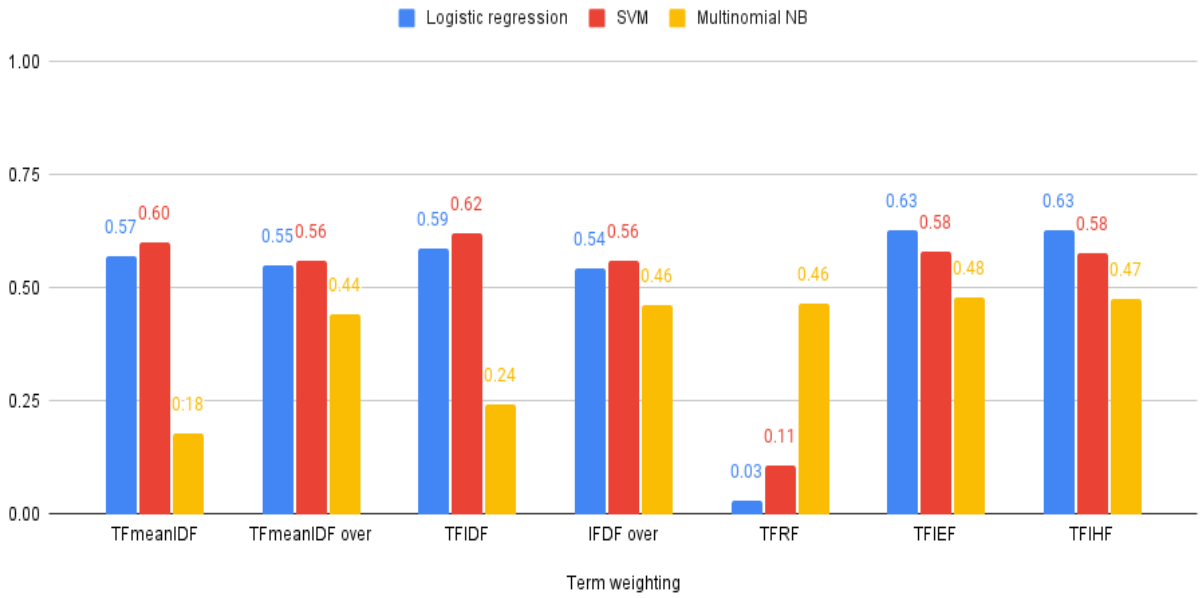


Figure 7: F1 score of minority class obtained from 7 product categories (Experiment C).

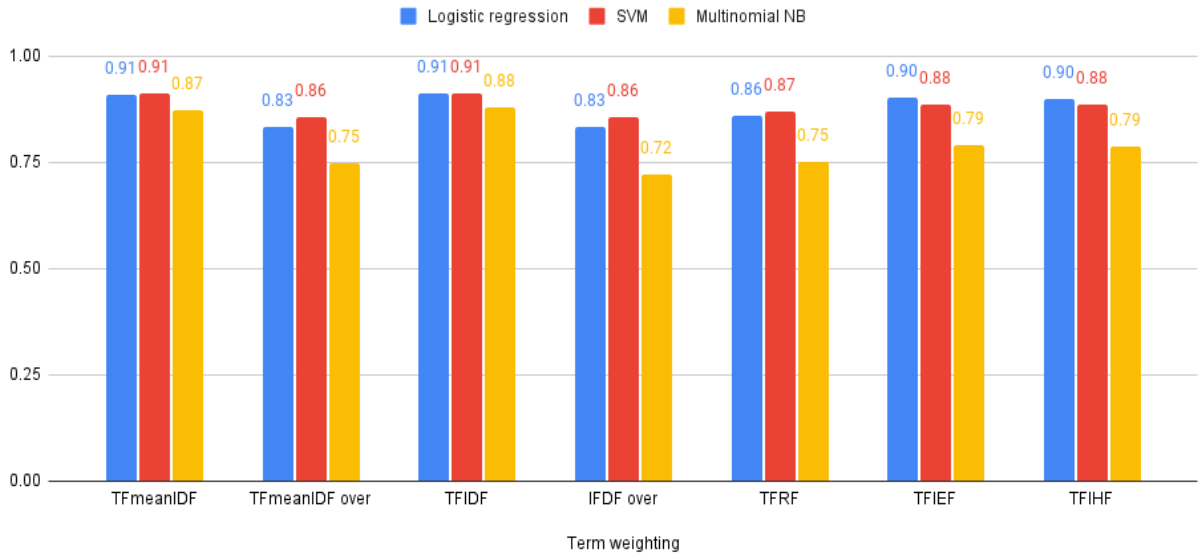


Figure 8: Comparision of Micro average of F1 s core obtained from various term weighting methods (Experiment C Long text classification problem).

We do statistical analysis to compare the performance of classifiers using different term weighting approaches on 3 classification problems. Table 7 represents the p-value obtained from Friedman test [19]. As shown in Table 7, all p-value are less than 0.05. Therefore, the classification performance of all classifiers are statistically significant.

Classification problem	Logistic Regression	SVM	NB	ALL
Experiment A	$2.98x10^{-06}$	$2.60x10^{-06}$	$3.58x10^{-21}$	$3.58x10^{-21}$
Experiment B	$5.10x10^{-06}$	$1.14x10^{-05}$	$3.20x10^{-05}$	$1.90x10^{-18}$
Experiment C	4.87x10-05	0.35x10-4	1.04x10-3	6.15x10-13

Table 7: The p-value from Friedman test.

5 Conclusion

This paper proposed a new term weighting scheme namely *TFmeanIDF* that can enhance the performance of SVM and Logistic Regression on imbalance text classification problem. The value of *TFmeanIDF* reflects the document frequency in minority class and majority class so that the word features that frequently found in minority class is more important than those found in majority class.

TFmeanIDF can be expected to increase the text classification performance for text documents with long length, since the proposed term weight method is computed by TF multiple by meanIDF. The term TF represents the frequency of words in a document. In addition, meanIDF illustrates the average of IDF in majority and minority class. When *TFmeanIDF* are computed, the length of the document is not biased to the output values, since meanIDF is normalized with the log scale as the same as the IDF term. Therefore, *TFmeanIDF* is able to represent the importance of each word in a document. The value of *TFmeanIDF* value is increase if the term frequently occur in the document. We found that *TFmeanIDF* can represent the proper term weighting when the term appears only in the majority or minority classes. However, the term weight will be small if the word occurs in both minority and majority class which is different from traditional TFIDF term weighting method that the weight is equal if the ratio is same.

To evaluate *TFmeanIDF*, three well-known algorithms which are Logistic Regression, SVM and multinomialNB are experimented on seven Amazon product datasets to compare the performance. The experimental result shows that *TFmeanIDF* works well on imbalance problem with high dimension. Our proposed term weighting scheme gets promising result for Logistic Regression and SVM compared to other term weighting schemes. Moreover, *TFmeanIDF* has low complexity. Future studies may consider to incorporate other domain knowledges, such as ontology, into the term weighting scheme to handle the imbalanced text problem.

Acknowledgement

This work was supported by Mahidol University, Research Fund under Grant No. A4/2563, Faculty of Science, Mahidol University, and Centre of Excellence in Mathematics, CHE, Thailand.

References

[1] Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. Vol. 24 (5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[2] Chen, L., Jiang, L. & Li, C. (2021). Using modified term frequency to improve term weighting for text classification, Engineering Applications of Artificial Intelligence. https://doi.org/10.1016/j.engappai.2021.104215

[3] Samant, S., Murthy, N., Malapati, A. (2021). Categorization of event clusters from twitter using term weighting schemes. Informatica, (45) 405–414. https://doi.org/10.31449/inf.v45i3.3063

[4] Tang, Z., Li, W. & Li, Y. (2020). An improved term weighing scheme for text classification. Concurrency Computation Practice and Expert, Vol. 32, 1-19. https://doi.org/10.1002/cpe.5604

[5] Tang, Z., Li, W., Li, Y., Zhao, W. & Li, S. (2020). Several alternative term weighting methods for text representation and classification. Knowledge-Based Systems, Vol. 207, 1-14. https://doi.org/10.1016/j.knosys.2020.106399

[6] Lan, M., Tan, C.L., Su, J. & Lu.Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31(4), 721-735. https://doi.org/10.1109/TPAMI.2008.110

[7] Ng, H.T., Goh, W.B. & Low, K.L. (1997). Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization, Proc. SIGIR '97, 67-73. https://doi.org/10.1145/278459.258537

[8] López, V., Fernández, A., Moreno-Torres, J.G.& Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Expert System with Applications, Vol. 39, 6585–6608. https://doi.org/10.1016/j.eswa.2011.12.043

[9] Ruining H. & Julian M. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. Proc of Int. World Wide Web Conference Committee (IW3C2), Canada. http://dx.doi.org/10.1145/2872427.2883037

[10] Ying L., Han T.L. & Aixin S. (2009). Imbalanced text classification: A term weighting approach. Expert System with Applications, 690-701. https://doi.org/10.1016/j.eswa.2007.10.042

[11] Abhilasha S.R., Amit A. & Preeti D. (2018). Comparative Study of Machine Learning Approaches for Amazon Reviews. International Conference on Computational Intelligence and Data Science (ICCIDS 2018), 1552-1561. https://doi.org/ 10.1016/j.procs.2018.05.119

[12] Kaur S. & Rajni Mohana, R. (2018). Prediction of Sentiment from Macaronic Reviews. Informatica. (42), 127–136.

[13] Tiwari, P., Pandey, H.M., Khamparia, A. & Kumar, S. (2019). Twitter-based opinion mining for flight service utilizing machine learning. Informatica 43 (2019) 381–386 https://doi.org/10.31449/inf.v43i3.2615

[14] Yayla, R. (2021). Determining of the User Attitudes on Mobile Security Programs with Machine Learning Methods. Informatica 393–403 https://doi.org/10.31449/inf.v45i3.3506

[15] Mark S., Nicolas L.R., & Francis B. (2017). Finite Sums with the Stochastic Average Gradient. Mathematical Programming, Springer Verlag, 83-112.

[16] John C. P. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized

Likelihood Methods. Advance in Large Margin Classifiers, 61-74.

[17] Manning, C.D., Raghavan, P., & Schuetze, H. (2008) Introduction to Information Retrieval. Cambridge University Press, 234-265.

[18] Wan, L., Papageorgiou, G., Seddon, M. & Bernardoni, M. (2019). Long-length Legal Document Classification. https://doi.org/10.13140/RG.2.2.36657.12646

[19] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc, 32, 675–701.

Investigating Energy Efficiency of Mobile Ad-hoc Network Routing Protocols

S.A. Ajagbe[1], M.O. Ayegboyin[2], I.R. Idowu[3], T.A. Adeleke[1] and Dang N. H. Thanh[4*]
E-mail: saajagbe@pgschool.lautech.edu.ng, mike.oludayo998@gmail.com, ifedotun.idowu18@kwasu.edu.ng, adetoyese4love@gmail.com, thanhdnh@ueh.edu.vn
[1] Computer Engineering Department, Ladoke Akintola University of Technology, Ogbomoso, Nigeria
[2] Computer Science Department, University of Ibadan, Ibadan, Nigeria
[3] Computer Science Department, Kwara State University, Ilorin, Nigeria
[4] Department of Information Technology, College of Technology and Design, University of Economics Ho Chi Minh City (UEH), Ho Chi Minh City, Vietnam

Keywords: AODV, DSDV, MANET, residual energy, routing protocol, security, wireless network

Received: June 2, 2021

Mobile Ad-hoc Network (MANET) is a wireless network that comes with a few routing protocols which have varied mechanisms. Studies show that routing operations consume energy while current research focuses more on MANET routing protocol operation and its performance evaluation, the required energy for successful routing operations equally demands quality attention of researchers. Hence, the need to expand the scope of study on MANET routing protocols to the neglected area of study. To bridge the research gap between MANET routing protocols and energy consumption, this paper used residual energy to analyze routing protocols' energy efficiency as a metric to analyze selected routing protocols; Destination Sequence Distance Vector (DSDV) and Ad-hoc On-demand Distance Vector (AODV) via simulation. It also compared the amount of energy required to transmit the data packets to their destinations in DSDV and AODV. Results, in terms of energy efficiency, indicated that AODV was better than DSDV because it consumed less energy for its successful routing operations.

Povzetek: Analizirana je energijska učinkovitost raznih protokolov mobilnih omrežij MANET.

1 Introduction

The Mobile ad-hoc network (MANET) is a wireless network. It can connect several mobile network interfaces to create a temporary network to facilitate the effortless transmission of data. Routing in MANET, like other network research, has been studied in recent years, and one of the major issues is the changing nature of its environment [1]. The changing environment should be addressed by opting for the best route to securely send source node data to a defined destination, routing protocol makes this task effortless. It requires no centralized infrastructure to run its functions. Being a self-configured and organized network, MANET aids efficient transmission and reception of nodes via its communication links without any barrier [2]. It's a good network arrangement where a wired network is practically impossible, and its setup process is effortless. At times, the wired network infrastructure may fail due to certain issues such as battlefield, hostile terrain operations, decision making, emergency search-and-rescue operations, and data acquisition [3]. Such issues could pose installation, financial, and security challenges for the creation of wired networks but communication is still possible through the MANET, or what is simply known as MANET. Each mobile node within the network runs both host and router functions by transporting packets to other mobile nodes in the network, especially those lying directly within each node's transmission range [4].

Continuous operations of routing protocols broadcasting in the network tend to increase energy consumption, in evaluating a routing protocol, many researchers have ignored the network's energy consumption. Those who look along that line are so few that no significant studies have been carried out on a protocol's energy consumption, and metrics parameters such as Dropped Packet Vs Time, and Packet Delivery Fraction (ratio of the packet received/packet sent), and Residual Energy Vs Time, being key performance indicators neglected in previous studies. There is a need to expand already conducted studies and pay keen attention to factors of energy consumption metrics ignored in previous studies. Hence, this paper shows an investigation of MANET routing protocols' energy efficiency. This study aims at bridging the research gap by investigating the efficiency of routing protocols mainly in the area of energy consumption. It will use simulations to compare and contrast the performance of DSDV and AODV

Our research contribution focus on:

* Corresponding author

(1) Investigating the energy efficiency of the DSDV and AODV MANET.

(2) Compare and contrast the performance of DSDV and AODV with important efficiency indicators such as Dropped Packet Vs Time, and Packet Delivery Fraction (ratio of the packet received/packet sent), and Residual Energy Vs Time.

The remaining part of this research is organized as follows: section 2 contained the review of related works, section 3 discusses the methodology employed, and section 4 presents experimental results and discussion on an investigation of MANET routing protocols' energy efficiency. Finally, section 5 concludes the paper

2 Review of related works

MANET has limits but is not volatile to the challenges of similar networks which include dynamic topology, security, computing, and communication resources such as bandwidth, time, memory, and energy or energy efficiency [5]. Designing a routing protocol could be negatively affected by the challenges raised above, as pointed out by a few studies. A major determinant of the efficiency of a routing protocol in a network is energy consumption. Time-tested routing protocols are known for their efficient energy capacity since a periodic update of routing protocols against malicious activities poses more energy burden on the performance of the network, especially the MANET wireless network. So, the rate at which energy is consumed on the MANET network tends to increase due to up-to-date processing of routing information across all nodes and broadcasting topology which is updated at regular time intervals. The updating operations, which could be termed Proactive and Reactive routing protocols, have a significant effect on the network's overall energy consumption [6].

Kanellopoulos, (2019) [7] while studying MANET and other emerging technologies that run as host and router in the network, proposed the AODV routing protocol. They concluded that the AODV showcased tremendous improvement, especially in the areas of stability and lifespan, as well as its many network paths. The focus was mainly directed to its lifespan because it was widely believed that residual energy poses little or no challenge to the AODV routing protocol, energy efficiency was not considered in this work.

AL-Dhief, et a, [8] and Al-khati & Hassan, [9] argued that MANET has several routing protocols. They confessed that these protocols have different mechanisms of make-up and performance. So, while studying the performance parameters of these protocols across varied environmental conditions, the researchers compared and contrasted AODV, DSDV, and the Dynamic Source Routing (DSR) protocols. The study was to ascertain the efficiency of the selected protocols. Simulation and evaluation of the selected protocols were carried out via the Network Simulator NS2.35 and the focus was on their average end-to-end delay, packet delivery ratio, packet loss ratio, and average throughput. While the variable number of nodes was considered in the course of an evaluation, the researchers have not paid attention to factors such as residual energy Vs time and dropped packet Vs Time, the key parameters of MANET evaluation that will be focused on in this research.

It is hard to point out the QoS and energy efficiency for MANET routing protocols. Studies show that different routing protocol features have been proposed recently and their scope of performance has been carefully evaluated. Evaluation focus has mostly focused on metrics like routing overhead, packet delivery ratio, and delay. Although, routing protocols were briefly examined alongside energy consumption and QoS. The energy consumption requirement of the network was analyzed with the aid of network simulator 2, also known as the NS2. A few mobility and traffic models were used in analyzing the energy consumption requirement of the network [10]. However, there is a need for a performance comparison of DSDV and AODV to arrive at a more correct evaluation.

Jamali, Rezaei, and Gudakahriz, (2013) [11] studied and compared the basic proactive, reactive, and hybrid features of the selected protocols, which were the AODV, DSDV, and the improved AODV protocols, and the study only addressed the selected protocols' end-to-end delay and packet delivery ratio. Though it was reported that the DSDV has a better result since it comes with enhanced functionality for optimum performance, the evaluation was not based on energy efficiency and its important parameters. An implementation of the above research was done with the NS 3 simulator to rate the performance of AODV, DSDV showed an improved performance than AODV routing protocols. The research did not study the energy efficiency of the network. Gopinath & Nagarajan, (2015) [12] worked on energy conservation on mobile nodes since they usually fall prey to battery issues. Transmission of packets and signal reception from interfering nodes require a bunch of energy. So, the researchers were inclined to study nodes' overall energy consumption to mitigate transmission and reception issues due to low energy and boost the lifespan of the network. Results, however, showed that simulating the network regularly could help save energy and limit undue interference from unauthorized protocols.

Researchers have shown through their studies in the last decade, the raw potential of MANET. One of these studies is [13] which studied the MANET and focused on the energy consumption and performance metrics without addressing it from the known examples of MANET which are AODV and DSDV. Also, the Energy Efficient Ad Hoc Distance Vector protocol (EE-AODV), could be a significant routing protocol to boost the available AODV routing protocol. The RREQ and RREP can be used to save energy in mobile devices and are also products of the algorithm that brought about the Energy Efficient Ad-hoc Distance Vector protocol (EEAODV). Any node that wants to act as an intermediate node will require minimum energy consumption and the energy requirement of the EE-AODV is fairly considerable. Simulation results indicated that the EEAODV increased the lifetime of the network, unlike the AODV [14]. However, the research did not look at the research from the known examples of MANET which are AODV and DSDV.

Gayathri, et al, 2013 [15] opined that DSDV performed better than AODV in a comparative study on AODV and DSDV energy consumption and QoS from the different network simulation results, and it was showed that the enhanced protocol performs more than the EQ-AODV (Energy and QoS supported AODV), while the regular AODV had significant energy dissipation. It can be argued that the energy issue affected some services within the network. Janani, et al, [16] like other studies, have evaluated both proactive and reactive performance of routing protocols through a few evaluation methods. Different results were arrived at since studies took place in different simulation environments.

Chaphekar, Sonkar, & Gupta, (2014) [17] proposed a method to increase data availability and reduce data traffic in the MANET. Each mobile node was provided a buffer for temporarily storing data for a particular moment. The overhead of the server and data traffic in the server zone was reduced according to [18]. The proposed approach reduces the time consumed by multiple nodes and data availability was increased. A similar method was the lifetime ratio (LR) of the active route for the intermediate node that was introduced to increase the number of unsuccessful packets delivery. The results focused on the improvement of the packet delivery in the routing protocol. In the course of evaluating the mobility of MANET nodes, parameters used include the packet delivery, average end-to-end, overhead, and energy consumption. Routing protocols' complexity is increased while the connection experienced some forms of flexibility due to the mobility of the nodes [19].

3 Methodology

An attempt was made to compare the reactive Ad-hoc On-Demand (AODV) routing protocol with the proactive DSDV. GloMoSim, the academics simulator for MANET, is used to simulate the two main environmental scenarios created. While the first is MANET scalability was represented by a few mobile nodes, the second had users' mobility and got represented by mobile nodes with pause time, minimum, and maximum speeds. Reactive versus proactive routing protocols' energy consumption in MANET is the major focus of this study, and it is carried out from MANET's different sizes and users' mobility conditions. The research procedure followed, and material used viz-a-viz other research tools for this study, simulation design, the simulator, and selection of routing protocols were discussed under research methodologies.

3.1 Routing protocols

Selected routing protocols are the DSDV and AODV. These protocols were selected because they cut across the divide between Proactive and Reactive routing protocols, being requisite demands of studies of this magnitude.

3.2 Study design

For this study, four basic approaches of Network Simulation namely; model design, configure parameters, run simulation, and result and analysis were employed in this research, the Network Simulation has four parts that can be deployed in the energy-related study. Obtained results will be remodeled if what was got turned out to be incorrect. Figure 1 shows the Block Diagram of the Simulation Model, which doubles as the basic working flow of the Network simulation.

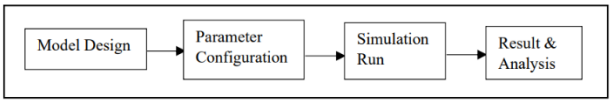


Figure 1: Block diagram of simulation model.

Figure 2: System design diagram.

3.3 Study design flow diagram

The tasks include writing or modification of TCL Scripts for the network simulator, as well as the integration of the routing algorithm. Hence, in this research work, the researcher used the NS scenario generator to write the scripts for the network, a move that is in line with some of the objectives of the study as presented in Figure 2. Layer 1 has the TCL generator, also known as NS2 Scenario Generator, while Layer 2 comes with the graph plotter (xGraph) and image file graph. Below is the design flow for the project and each block are further explained.

The Network Simulator 2(NS-2) framework aided the centered surface creation. Right parameters are set in the NS-2 to design a raw text file to house energy dissipation information on the network, while the XGraph refers to a plotting program used in creating graphic representations of simulation results. In other words, the XGraph houses the results of the NS-2 simulations. It is a tool used to plot the results. It is significant since it aids the basic animation of data sets, and also helps users view, understand, and appreciate the information in the NS-2 text file graphically. Overall simulation results will be hard to view if they do not come in graphical format. Image files used could be lines, curves, bars, or other symbols, but they show the measured relationships very clearly.

3.4 Performance of DSDV and AODV

The simulation of AODV and DSDV routing protocols was run on 10 nodes. The chosen nodes were made to

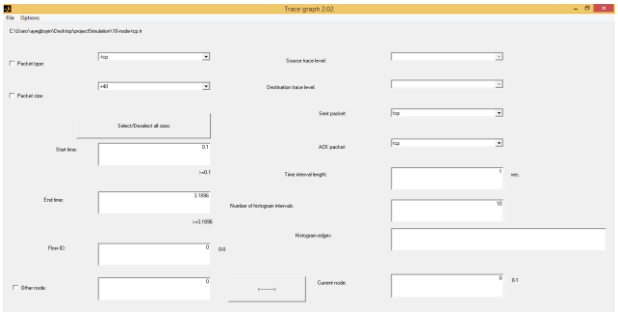


Figure 3: Trace graph graphical interface.

move under the area of 500m x 500m, while the transmission range was fixed at 250m. Each node in the network had initial energy of 1000 Joules, and the simulation results of the two routing protocols will be analyzed on the *.tr file, as well as these performance metrics: Residual Energy Vs Time, Dropped Packet Vs Time, Packet Deliver Fraction

Energy Efficiency. This study wants to measure and calculate the energy consumption of a node in the DSDV and AODV routing protocols, being one of the aims of the study. Nodes on these protocols will be compared and analyzed. In their study, Eiman, A., Biswanath, M., (2012) [20] proposed the Energy factor, an energy-efficient routing metric. The energy factor is used to calculate energy consumption in the current study, and it is defined thus:

$$E = E_{Remaining}/E_{Initial} \quad (1)$$

where remaining energy in (1) is defined as:

$$E_{Remaining} = E_{Initial} - E_{Consumed} \quad (2)$$

This metric was used to run the multipath concept, being the most energy-efficient concept, while the shortest path was used for data packets' transmission. Equation (2) of this metric is the proposed method to be used in the study to calculate the residual energy once the packets have gone through the transmission. To compare the two protocols, the evaluation was based on these simulation metrics:

- Packet delivery fraction refers to the packets data ratio, as delivered to desired destinations generated via the sources. Packet delivery fraction is calculated by dividing the total number of packets sent by the total number of packets received.
- Packet Drop points to the number of packets not received at the destination node.
- Total energy consumption in the network refers to the total energy used by nodes within the network for the transmission and reception of packets.
- Total Residual Energy of the network sums up the total residual energies of the nodes on the network. Residual Energy is calculated by subtracting Consumed Energy from the Initial Energy.

3.5 Software tool for implementation

The design and implementation of this study are rooted in the Network Simulator, also known as the NS-2. Being a network simulator, the NS-2 offers a significant virtual network communication environment that is compatible with MANET routing protocols. It is designed to provide an efficient power management system, unlike regular sensor networks and enhancement network platforms like MPLS and Ipv6. The NS-2 can be installed with Cygwin, a LINUX lookalike platform, on 64-bit versions of Windows 7 upwards. Host computers must have at least 7,500 GB Hard-disk and 4 GB Ram. Available packets that support the Simulator include Tool Command Language (TCL/TK), TCL with Classes (TCLCL), and Object-Oriented Extension of TCL (OTCL). The TK, a cross-platform and open-source widget toolkit, is a storehouse of a graphical interface for users. All the components complement one another, and they function in the orders that were built. NAM and Xgraph are two of the amazing tools that come with the NS-2 software packets.

Built on the NS 2.35 simulator software, Trace graph, NSG 2.1, Xgraph, and other easy-to-use supportive software components, the Network simulator aids the effortless generation of names for files, as well as their tracing. Already, NAM, as a Tool Command Language (TCL), uses an animation tool to view network simulation or monitor the packet trace. NAM animates the trace file being generated through the NS2 software. NAM, like some other tools, is used for packet-level, network animations, and other general-purpose operations.

3.6 Trace graph graphical interface

Trace files for this performance evaluation were all fed into the trace graph. While loading the files, TRACE GRAPH recognized and recorded the shown attributes and features in the trace graph window, as shown in Figure 3.

Packet Type: It enables calculations created for the packet types (+ to select and - to deselect) to either select or deselect a packet type. Users may select or deselect packet type by default, depending on what they want. Check the appropriate option or all the types will be used for calculations.

Packet Size: It enables calculations to go with the packet types. While the + option is used to select a packet size, the - option is used to deselect it. Feel free to either select or deselect the packet size by default. All the available types will be used for calculations when the user fails to select any of the options.

Start Time: Time interval begins with the start time.

End Time: Time interval ends with the end time.

Sent Packet: Right from the source, tap on this option will select the required sent packet type.

Ack Packet: Right there at the destination, users can tap this option to select ACK packet data.

Time interval length refers to a specified time interval, just for the .tr file. It takes a few seconds to load data via this option.

3.7 Add-on for NS-2

All the Trace files used to evaluate performance were lodged into the Trace graph. While loading these files, TRACE GRAPH recognized and recorded the shown attributes of the parameters in the trace graph window, as shown in Figure 3.

Packet Type: It enables calculations made for selected packet types (+ for selected and - for deselected) to either select or deselect a packet type. But, depending on what the user wants, select, or deselect operations have default options. The preferred option must be selected, or all types will fall under calculation operations.

Packet Size: It enables calculations to go with all the selected types (+ for selected and - for deselected) to either select or deselect a packet size. Also, the desired packet size can be selected or deselected by default, depending on what the user wants. All the types will be used in calculations if the user fails to select the appropriate option.

Start Time: This points to the start of the operations' time interval.
End Time: This refers to the end of the operations' time interval.
Sent Packet: Users tend to use this option to select desired sent packet type from the operations' source.
Ack Packet: The ACK packet type is selected at the destination when this option is used.
Time interval length refers to the operations' specified time, especially the .tr file. It takes a couple of seconds to load data from the operations' beginning or end.

4 Experimental results and discussion

The experimental result in this study was based on a comparison of the performance of the two protocols (AODV and DSDV) under distinct and varied environmental conditions.

4.1 Residual energy VS time

Certain energy of nodes dissipates after they have been used to send data to neighbors or receive data from them, with time, the residual energy of nodes began to decrease or reduce. Figure 4 presents the relationship between Time and Residual Energy for the DSDV Routing Protocol. While residual energy takes the y-axis, data on time occupies the x-axis. Analysis from the graph shows that the residual energy reduces gradually with time. Further analysis indicated that, since the DSDV runs a proactive routing protocol, each node could only use its routing tables to transmit a packet to its neighbors, and this has a significant effect on energy consumption. Figure 5 presents the relationship between Residual Energy and Time for AODV, the Ad-hoc Demand Distance Vector Routing Protocol. While the Residual Energy occupied the y-axis, Time took the x-axis part. One could see from the graph the rapid loss rate of the initial energy. The loss was due to the vast REQUEST/REPLY operations within the protocols that require a large volume of data packets for effortless transmission. Gradually, the energy residue began to decrease in the two protocols, as portrayed in Figures 4 and 5, but the rate of decrease of the residual energy in DSDV is lower than AODV. Further analysis showed that the DSDV consumed more energy than the AODV.

4.2 Dropped packet VS time

As the energy level reduces, destination nodes became unreachable and die prematurely since some packets had to drop off from the race. So, over time, a decrease in energy led to an increase in the rate of dropped packets, but this depends solely on used routing algorithms. The relationship between the dropped Packets and Time in the DSDV. Routing Protocol, as indicated by the y-axis and x-axis, respectively.

Figures 6 and 7 above showed the initial phase of some dropped packets. Then, gradually, the number of dropped packets began to increase in figure 7 than figure 6. At time 30secs, the number of drop packets is 350 in figure 7 while it is still 150 in figure 6 suggesting that lack of sufficient energy led to an increase in the number of dropped packets. So, the amount of energy used by routing operations in figure 7 is more than that of figure 6 and that is the reason why some nodes were less active in transmitting packet data thereby resulting in packet drop.

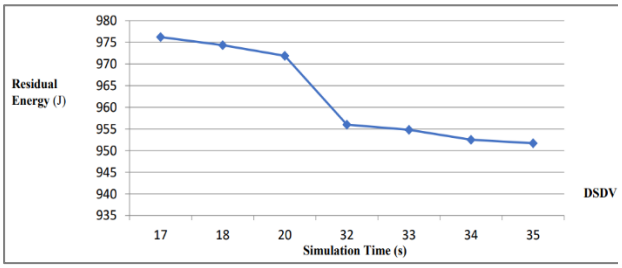


Figure 4:Residual energy Vs time in DSDV.

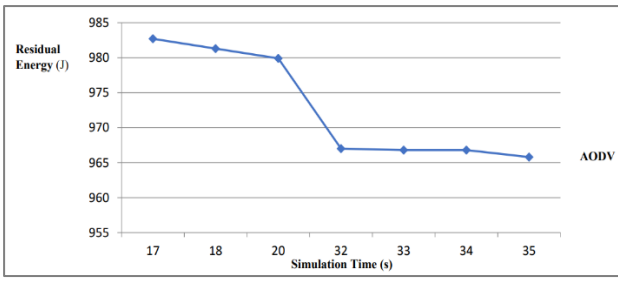


Figure 5: Residual energy Vs time in AODV.

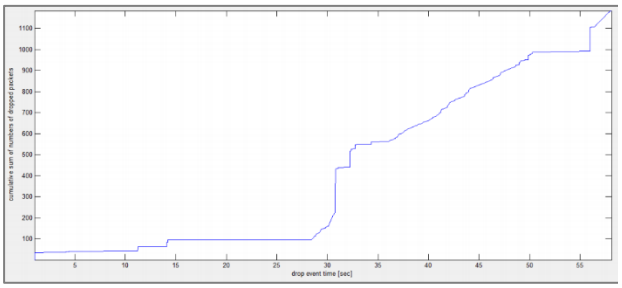


Figure 6: Dropped packet Vs time in DSDV.

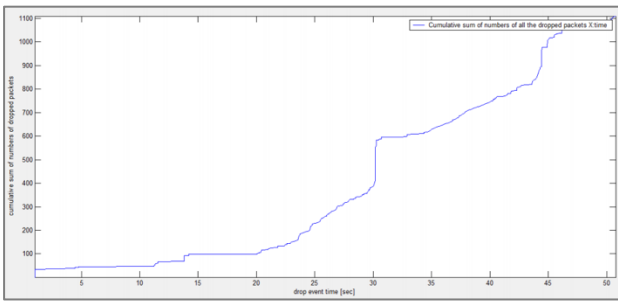


Figure 7: Dropped packet Vs time in AODV.

4.3 Packet delivery fraction

The ratio between sent and received packets is presented below. Sent packets were received at the destination generated in the traffic source.

In Figure 8, the Packet Delivery Fraction shows that AODV performs better than DSDV, and this result attested

Routing Protocol	№ of loss Packet	№ of Packet Send	№ of Packet Received
DSDV	2976	6832	3856
AODV	3314	6964	3650

Table 1: Packet delivery fraction of various routing protocols in MANET's network.

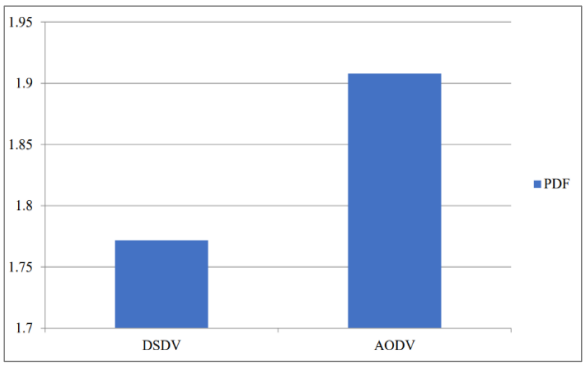


Figure 8: Packed delivery fraction of AODV against DSDV.

that DSDV does not always do well under all parameters. The packet delivery fraction of AODV was high than DSDV.

5 Conclusions

One notable challenge of the changing environment is its adverse effect on the energy consumption of the network. If care is not taken, it can drastically drain the energy and halt the network. To address this, the focus should be on energy capacity when one wants to choose a routing protocol, as this will aid efficient calls and transmissions of data within the network. This work sought to study the energy efficiency of the MANET routing protocol. To achieve its objectives, the study introduced and focused on two new evaluation parameters — residual energy Vs time and dropped packet Vs time — and the available methods. Two routing protocols — DSDV and AODV— were experimented and their performances were studied based on the raised parameters. A simulation was used to compare and contrast their performances and the results showed that AODV did well than DSDV, as presented in Figure 4-8. So, this study recommends the AODV routing protocol for use in a network design where energy for transmission of packet data is considered important. We hope to introduce the use of internet of things (IoT) applications and deep learning (DL) approach to mitigate the effects of energy consumption on MANET routing protocol at work station from the based station in our future work.

Acknowledgments

This research was funded by the University of Economics Ho Chi Minh City (UEH), Vietnam.

References

[1] Alslaim, M. N., Alaqel, H. A., & Zaghloul, S. S. (2014). A comparative study of MANET routing protocols. The Third International Conference on e-Technologies and Networks for Development (ICeND2014) IEEE (pp. 178-182). Beirut, Lebanon: IEEE. https://doi.org/10.1109/ICeND.2014.6991375

[2] Pande S., Khamparia A., Gupta D., Thanh D.N.H. (2021) DDOS Detection Using Machine Learning Technique. In: Khanna A., Singh A.K., Swaroop A. (eds) Recent Studies on Computational Intelligence. Studies in Computational Intelligence, vol 921. Springer, Singapore. https://doi.org/10.1007/978-981-15-8469-5_5

[3] Kaur, G. & P. Thakur. (2019). Routing Protocols in MANET: An Overview. *" 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) IEEE* (pp. 935-941). Kannur, India : IEEE. https://doi.org/10.1109/ICICICT46008.2019.8993294

[4] Bamimore, I., & Ajagbe, S. A. (2020). Design and implementation of smart home for security using Radio Frequency modules,. *International Journal of Digital Signals and Smart Systems, 4*(4), 286-303. https://doi.org/10.1504/IJDSSS.2020.111009

[5] Sandeep, J., & Kumar, J. S. (2015). Efficient Packet Transmission and Energy Optimization in Military Operation Scenarios of MANET. *Procedia Computer Science, 45*, 400-407. https://doi.org/10.1016/j.procs.2015.03.223

[6] Thanh, D. and Dvoenko, S. (2015). A Variational Method to Remove the Combination of Poisson and Gaussian Noises. In *Proceedings of the 5th International Workshop on Image Mining. Theory and Applications - IMTA-5, (VISIGRAPP 2015)*. https://doi.org/10.5220/000546090038004

[7] Kanellopoulos, D. (2019). Congestion control for MANETs: An overview. *ICT Express, 5*(2), 77-83. https://doi.org/10.1016/j.icte.2018.06.001

[8] AL-Dhief, F. T., Sabri, N., Salim, M. S., Fouad, S., & Aljunid, S. A. (2018). MANET Routing Protocols Evaluation: AODV, DSR and DSDV Perspective. *MATEC Web of Conferences*, (pp. 1-7). https://doi.org/10.1051/matecconf/201815006024

[9] Al-khati, A. A., & Hassan, R. (2018). Performance Evaluation of AODV, DSDV, and DSR Routing Protocols in MANET Using NS-2 Simulator. *Recent Trends in Information and Communication Technology* (pp. 276-284). Springer. https://doi.org/10.1007/978-3-319-59427-9_30

[10] Mathias, M., Mariki, M. E., & Jan, H. K. (2018). System usability scale evaluation of online banking services: A South African study. *South Africa*

Journal of Science, 114(3/4), 50-57. https://doi.org/10.17159/sajs.2018/20170065

[11] Jamali, S., Rezaei, L., & Gudakahriz, S. J. (2013). An Energy-efficient Routing Protocol for MANETs: a Particle Swarm Optimization Approach. *Journal of Applied Research and Technology. JART, 11*(6), 803-812. https://doi.org/10.1016/S1665-6423(13)71586-4

[12] Gopinath, S., & Nagarajan, N. (2015). Energy based reliable multicast routing protocol for packet forwarding in MANET. *Journal of Applied Research and Technology. JART, 13*(3), 374-381. https://doi.org/10.1016/j.jart.2015.07.003

[13] Taruna, S., & Purohit, G. N. (2011). Scenario Based Performance Analysis of AODV and DSDV in Mobile Adhoc Network. In *Advances in Networks and Communications* (Vol. 132, pp. 10-19). Springer-Verlag Berlin Heidelberg 2011. https://doi.org/10.1007/978-3-642-17878-8_2

[14] Malek, A., LI, C., Yang, Z., Naji, H. A., & Zhang, X. (2012). Improved the Energy of Ad Hoc On-Demand Distance Vector Routing Protocol. *IERI Procedia, 2* (355–361). https://doi.org/10.1016/j.ieri.2012.06.101

[15] Gayathri, V. M., Raichal, S., Soniya, & Nedunchelian, R. (2013). Comparison of AODV and DSDV Protocols Based on Energy Consumption and QoS performance. *International Journal of Engineering and Technology (IJET), 5*(3), 2837-2841

[16] Janani, A. P., Sakthivel, M., & Baskaran, K. (2013). A competitive performance analysis of reactive and proactive routing protocols of MANET under short time communication scenario. *International Journal of Wireless and Mobile Computing*, 253-260. https://doi.org/10.1504/IJWMC.2013.055763

[17] Chaphekar, D., Sonkar, B., & Gupta, G. (2014). Study of Framework of Mobile IP and MANET Integration. *Intelligent Computing, Networking, and Informatics. , 243*, 985-992. https://doi.org/10.1007/978-81-322-1665-0_100

[18] Aluvala, S., Sekhar, K. R., & Vodnala, D. (2016). A novel technique for node authentication in mobile Ad-hoc networks. *Perspectives in Science, 8*, 680-682. https://doi.org/10.1016/j.pisc.2016.06.057

[19] Ghaffari, A. (2014). An Energy Efficient Routing Protocol for Wireless Sensor Networks using A-star Algorithm. *Journal of Applied Research and Technology. JART, 12*(4), 815-822. https://doi.org/10.1016/S1665-6423(14)70097-5

[20] Eiman, A., Biswanath, M., (2012) A survey on routing algorithms for wireless Ad-Hoc and mesh networks. *Computer Networks, 53(2) pp.* 940-965. https://doi.org/10.1016/j.comnet.2011.10.011

An Empirical Study to Demonstrate that EdDSA can be used as a Performance Improvement Alternative to ECDSA in Blockchain and IoT

Guruprakash J and Srinivas Koppu
E-mail: guruprakash.j2019@vitstudent.ac.in, srinukoppu@vit.ac.in and www.vit.ac.in
School of Information Technology and Engineering, Vellore Institute of Technology, India

Student paper

Keywords: digital signature, ECDSA, EdDSA, blockchain, IoT

Received: November 1, 2021

Digital signatures are a vital part of the digital world. The trust factor in the digital world is ensured with a digital signature. Over the evolution, the purpose remained constant, but the applicability and frontier continued to evolve, thus raising the demand for continuous performance, security level and computational improvement. Especially with emerging IoT, blockchain and cryptocurrency, the digital signature security level and performance improvement demand continue to rise. A digital signature scheme (DSS) is used to generate signatures. This paper investigates the widely used elliptic curve digital signature algorithm (ECDSA) and its application to blockchain and IoT. Then, we performed an empirical comparison of ECDSA with the Edwards curve digital signature algorithm (EdDSA). The study concludes by showing that EdDSA is superior to ECDSA and can be applied in blockchain and IoT domains to reap immediate benefits.

Povzetek: Avtorja primerjata dve metodi generiranja varnih digitalnih podpisov in pokažeta, da je EdDSA boljša kot ECDSA.

1 Introduction

Since the evolution of digital technology, digital signatures have played a vital role in providing integrity and security to the system. The rapid advancement and technological requirements did not leave the digital signature dormant, and they kindled the advancement needed for its coexistence. The transition could be observed in the rapid transformation from RSA to elliptic curve cryptography (ECC) and toward the advanced elliptic curve digital signature algorithm (ECDSA). These evolutions captured the researcher's attention, and ECDSA was mainly adopted due to its superiority in providing enhanced security with smaller keys and less operational space. This study aims to find a better performing alternative to ECDSA that can cater to IoT and blockchain based applications. The remaining part of this section introduces the digital signature, digital signature algorithm, and elliptic curve cryptography.

1.1 Digital signature

The digital signature (DS) is generated from a standard algorithm called the digital signature algorithm (DSA), which has to follow a specified standard scheme as put forth in the digital signature scheme (DSS). National Institute of Standards and Technology (NIST) and security council groups reviewed, tested, and described these standards. A digital signature authenticates identity, detects unauthorised data manipulation, guarantees against data tamper and is the only way for nonrepudiation in the digital world. Nonrepudiation assures evidence to the third party by the signatory. Later, the signatory cannot deny the activity with the third party or repudiate the sign [1].

The Figure 1 on page 277 shows that digital signatures provide authenticity, integrity and nonrepudiation.

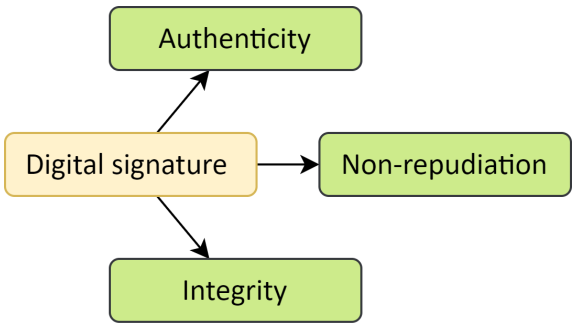


Figure 1: Application of Digital signature.

1.2 Digital signature algorithm

To achieve digital signatures, several forms of secure cryptographic standards of the digital signature algorithm (DSA) can be deployed. Figure 2 on page 278 shows digitally signed node interactions in an IoT domain. These

standards vary on the basic arithmetic operation involved, modular exponentiation and discrete logarithmic problem. DSA has a generic systematic flow for key generation, message signing, and verification; the standard provides deterministic and nondeterministic output. Generally, deterministic factors are considered to be safer and more secure.Evolution and demand brought in the need for modern cryptography, ECC. ECC was considered a successor to conventional DSA because it provided a shorter key, shorter signature, higher security and better performance. DSA, based on ECC, worked on cyclic groups of an elliptic curve over a finite field and difficult was based on the elliptic curve discrete logarithmic problem (ECDLP). The widely used variants are the ECDSA and Edwards curve digital signature algorithm (EdDSA).

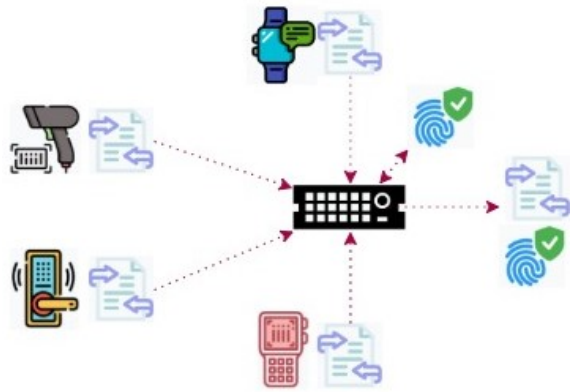


Figure 2: Digitally signed IoT node interaction in a connected domain.

1.3 Elliptic curve cryptography

ECDSA and EdDSA are among the variants of ECC with different curves as shown in Figure 3 on page 279. ECDSA, is based on the ElGamal signature and works on the group elements. The ECDSA computes a hash based on random keys, these random keys open a potential vulnerability, and the attacker can use it to gain an advantage. To overcome this, a non-deterministic variant based on HMAC (Hash message authentication code)became an alternative. EdDSA is the successor to ECDSA with fast DSA using Edwards curves Ed25519 and Ed448. EdDSA a variant of deterministic Schnorr's signature; solves the inherent problem of ECDSA. EdDSA is simple, secure and fast; unlike ECDSA, it relies on discrete logarithmic problems. Table 1 on page 279 shows various curve forms, their representation, plot and summary.

The article is structured as follows; Section 2 presents the objective and contribution. Section 3 illustrates the existing literature and application of digital signatures. Section 4 compares the elliptic and Edwards curve basic operations and metric of functional operations. Section 5 presents discussion and Section 6 concludes this article.

2 Objective & contributions

This work aims to perform an empirical study on ECDSA vs EdDSA and provide evidence-based results to prove that EdDSA has performance advantages. We collected evidence-based on the existing state-of-the-art current research in the fields of 1) the application of digital signatures, 2) the application of elliptic curve in blockchain, 3) the application of elliptic and Edwards curve in the IoT, 4) Schnorr's signature application, and 5) The application of Schnorr's-based aggregate signature in blockchain.In addition 6) Attribute to attribute comparison studies on Edwards curve DSA vs ECDSA help us demonstrate our stands on the superiority of the Edwards curve.

3 Literature review

In this section, we review the literature based on five broad questions. 1.How researchers have used digital signatures for blockchain and IoT based applications? 2.How does ECC find its application in blockchain, and what benefits they provide? 3.How elliptic and Edwards curves keep IoT devices secure and resources optimal? 4.How does Schnorr's signature help generate multiple signs and its advantage in blockchain-based applications? 5.How aggregate signatures have added value in blockchain applications? The review of the existing research proved the cumulative advantage and provided guidance for better alternative choices when designing future systems.

3.1 Digital signature

In this sub-section research works on various applications of Digital signatures are reviewed and presented.

[6] conducted a review of the existing DSA, RSA, ECDSA, and EdDSA and highlighted that DSA operations based on algebraic properties forming public-key cryptosystems can mutually authenticate. [7] implemented a self-sufficient library X64ECC for ECC that supported basic cryptographic functions such as key exchange, Zero Knowledge Proofs, and digital signature. The library was able to accelerates mission-critical arithmetic operations by leveraging the compiler intrinsic. [8] proposed a new public key scheme by implementing a twisted Edwards curve model. The secure message transmission in this scheme was ensured by using the property of one way, indistinguishability (IND) under chosen-plaintext attack (CPA), chosen-ciphertext attack (CCA) and the variant of other digital signature algorithms. [9] developed an ECC processor based on the Edwards25519 curve implemented using FPGA for increased speed, reduced area, and simple arithmetic with efficient hardware using projective coordinates. [10] proposed using EdDSA with SHA-512 for Bitcoin as it would help in better security and efficiency compared with existing secp256k1 with hash SHA-265.[11] proposed a new method for counting the order of Edwards Curve (Ed) and elliptic curves over a finite field that can be used

Curve form	Representation	Plot	Summary
Edwards	$ax^2 + y^2 = 1 + dx^2y^2$	a)x^2 + y^2 = 1-299 x^2 y^2	fast and complete
Weierstrass	$y^2 = x^3 + ax + b$	b) y^2 = x^3 - 0.5 x + 0.8	form is slow,trusted and incomplete
Jacobi-quartic	$y^2 = x^4 + 2ax^2 + 1$	c) x^2 = y^4 - 1.9 y^2 + 1	has the capacity for extensions
Hessian	$ax^3 + y^3 + 1 = dxy$	d) x^3 - y^3 + 1 = 0.3 x y	has a uniform and week representation

Table 1: Curves and represents [2] [3] [4] [5].

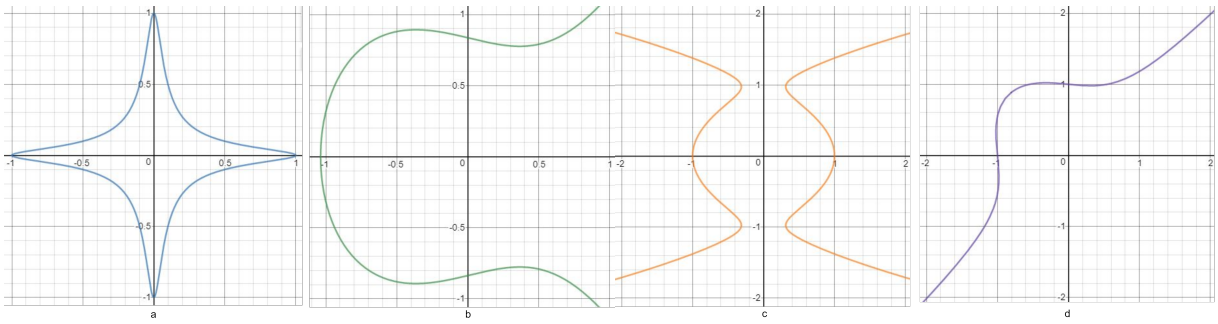


Figure 3: Plot of Curves a) Edwards, b) Weierstrass, c) Jacobi-quartic and d) Hessian.

for fast group operations with less complexity. The proposed method determines whether the curve is supersingular over a finite field. [12] cryptographically implement a simple Maps connecting Kummer, Montgomery curves and twisted Edwards lines. It was also evident that we can propose a low-power, low-area FPGA implementation of the Edwards curve and Montgomery curve [13]. In the method, the EdDSA provides faster digital signatures than existing schemes. [14] suggested secure and efficient software implementation of EdCDH, and EdDSA using SIMD parallel processing and obtained high performance.[15] focused on the Edwards curve to increase the performance and security over binary fields to overcome side-channel attacks. [16] implemented ECC on an FPGA with more speed, less area, and side-channel attack resistance. The processor supports point multiplication on different derivatives and achieves accelerated point multiplication with minimal hardware utilisation.[17] evaluated Montgomery-Twisted-Edwards and ECC implementations on IoT devices based on three different factors - ROM, RAM, and execution time. Their study provides reference results for the transition from legacy ECC to MoTE-ECC. [18] proposed a flexible Pedersen commitment implementation based on elliptic curves in twisted Edwards form, which helped to improve security, adaptive data size, and data point flexibility.[19] investigated optimal prime fields for lightweight ECC implementation, focusing on performance and security.[20] performed a comparative study of algorithms for batch verification of Edwards curve digital signatures and showed that small batch size algorithms S2' and SP yield better speedup results than the default algorithm N'.

Table 2 on page 280 shows various applications of Digital signature.

3.2 Application of elliptic curve cryptography in blockchain

In this sub-section research works on various applications of Elliptic curve cryptography in blockchain are reviewed and presented.

[22] proposed a four-layer framework for decentralised privacy-preserving management for electronic medical records using blockchain with Elliptic curve based digital signatures and content extraction signature(CES) and achieved access control and data privacy.[23] deployed an EMR on a blockchain-based infrastructure and mitigated a single point of failure on electronic medical records. ECC and ECDSA provide the security backbone for all operations. The major problems in the Bitcoin system were solved by employing ECDSA to circumvent security traps and generate new keys for each transaction, thereby improving security against attacks [24]. Increasing the number of electric vehicles and the internet of electric vehicles bring trust issues into the environment [25] exploited the use of blockchain and smart contracts to bring trust and tackle disputes in energy trading. [26] used blockchain-based mobile crowd sensing application for collective intelligence, and blockchain aided in keeping the process decentralised, secure, fast, optimised storage and privacy preserved. [27] proposed a modified ECC based on identity and derived an Elliptic curve access control mechanism and a lightweight digital signature algorithm to ensure data privacy and security. [28] employed ECC and a certificateless aggregate system(CAS) to achieve traceability, integrity and secure storage of electronic health records. The use of ECC and CAS helped to safeguard from unauthorised access when utilising the cloud storage. [29] build a

EC / DSS / DSA	Domain	Solution	#
		Security enhancement	
	Cryptocurrency	Efficiency improvement	[10]
		Security enhancement	[8]
		Security enhancement	[11]
		Security enhancement	
		Cost reduction	[12]
Edwards		Optimal arithmetic	
Curve		Performance improvement	
		Optimal arithmetic	[14]
		Reduced execution time	
		Performance improvement	
	General	Security improvement	[15]
		Speed improvement	
		Reduced space	
		Enhanced security	[16]
		Optimal arithmetic	
		Security enhancement	
		Runtime optimisation	[18]
		Enhance computation speed	[20]
Generic		Review	[6]

Table 2: Digital Signature applications in various domain.

designated verifier proof of assert (DV-POA) for currency exchange entirely based on ECC making it provable, secure and efficient. Voting requires great privacy, and the real identity of the voter should not be exposed. Nevertheless, the voter should be a verifiable identity. [30] built an E2E verifiable voting system based on blockchain. The cryptography and signatures are supported by BLS over a well-known elliptic curve providing a short signature and voter anonymity. User privacy is the most difficult challenge to overcome in regard to data mining and sociological mining. [31] created a blockchain-based privacy protection scheme with a ring signature and an elliptic curve that adds privacy to data storage and performs anonymous mining in a secure manner. Edge computing is always resource-starved, and enhancement needs to be in place for continuous improvement. [32] used an elliptic curve cryptosystem to preserve privacy and eliminated attacks. Their scheme was resilient with a computing environment based on the random public key. Medical data integrity is vital, while utilising cloud-based storage sufficient mechanism to protect the data needs to be employed. [33] employed ECDSA and build a lightweight auditing scheme for medical data privacy. [34] built a security and privacy scheme for coin mixing using an elliptic curve digital signature scheme with standard ring signature. They were able to achieve unforgeability with appreciable transaction efficiency. Unmanned aerial vehicles(UAVs) have captured substantial market share, and privacy-preserving and authentication have become growing issues in UAVs. [35] proposed a solution based on ECC and DS to achieve all the cryptography services required for UAV certification. [36] presents signcryption using the advanced Elliptic curve to protect the stringent legal and privacy required for the E-prescription system. The use of ECC helped them achieve the required protection in a low resource and computing constraint environment. [37] used elliptic curve cryptography to sign and encrypt data uploaded through patient authorisation via third party proxies. The use of ECC made their scheme light and suitable for authentication on cloud-based medical systems that provide computing and storage service to the healthcare domain. [38] survey demonstrated that most blockchain and cryptocurrency used structured based on the elliptic curve digital signature algorithm. Bitcoin mainly uses secp256k1 and assumed to be a platform with high encryption and security.In [37] they built a system based on an elliptic curve digital signature and practical Byzantine fault-tolerance, to achieve security demanded by China's electricity market.[39] used signatures based on bilinear pairing and elliptic curves to ensure transmission integrity and reliability for data security of shared storage system based on a smart contract. Compared with the traditional method, the application of the elliptic curve reduced block confirmation and improved transmission.[40] provided privacy enhancement to a Bitcoin transaction by mixing an elliptic curve digital signature and ring signature scheme. The outcome resulted in the users being able to identify the customer associated with the address.[41] employed elliptic curve discrete logarithm and bilinear for aggregate signature to shorten and compress a single signature. Their work demonstrated that the signature size remained constant irrespective of multiple inputs and outputs in the transaction. [75] the authors applied and enhanced ECC based encryption and decryption of IoT data transmission to its core ecosystem. They showed that application of ECC significantly help to improve the overall performance.

Table 3 on page 281 shows various applications of ECC in blockchain

EC / DSS / DSA	Domain	Solution	#
ECC	Computation	Key generation, Privacy-preserving, Security enhancement	[32]
	Egovernance	Enhance security, Digital signature, Identity management	[30]
	Healthcare	Storage optimisation, Efficiency improvement, Access control, Authentication, Security enhancement, Privacy, Identity management	[22],[23], [28],[36], [37]
	IoT Mobile	Privacy, Authentication, Key generation, Privacy-preserving, Security enhancement	[35],[26]
	Security Enhancement	Enhance security, Runtime optimisation	[42],[72]
ECDSA	Cloud Cloud storage	Performance improvement, Security enhancement, Storage, Signature , Privacy, Data management	[27],[33], [41],[31]
	Cryptocurrency	Signature generation verification, Provability, Privacy, Enhanced security, Identity Key management	[29],[34], [38],[40], [24]
	Electric vehicle	Storage, Security enhancement, Privacy, protection	[37],[25], [39]

Table 3: Application of ECC in blockchain.

3.3 Edwards curve application in IoT

In this sub-section research works on various Edwards curve applications in IoT are reviewed and presented.

[9] designed a public key generation with unified point addition on twisted Edwards curve for IoT security. The choice of Edwards curve for the digital signature was its fast grouping operation and resistance to side-channel attack, which was the drawback of elliptic curves. [43] presented the need for public-key cryptography for IoT based applications to provide exceptional efficiency, reduce resources and increase security levels in a small setup. Using Edwards curve cryptography for resource and power constrained IoT applications was an optimal choice. [13] presented EdDSA implementation using ED25519 and achieved reduced hardware implementation complexity for IoT applications. The application of ECC is catching up recently for IoT and allied application. [44] proposed the Edwards curve to optimise the power and memory consumption in a physical device. [45] presented a fast, low power and highly secure cryptography for IoT by using a binary Edwards curve. They were able to achieve optimised curve arithmetic and provide the performance benefit of intrinsic security for IoT devices against physical attacks. ECC is widely used for keys, encryption, decryption and digital signatures. [21] used a twisted Edwards curve variant and demonstrated performance improvement on target platform. Verification of digital signature from ECDSA requires double scalar multiplication. These issues result in speed and size issues in IoT applications. [76] enhanced Edwards curve to achieve nonrepudiation in IoT blockchain ecosystem. [19] Their work demonstrates the use of the Edwards curve and showcases how they can reduce implementation space, and operational cost and perform fast verification.

Table 4 on page 282 summarises on how the Edwards curve is used in IoT for security enhancement.

3.4 Schnorr's signature for multi-signature

In this sub-section research works on various applications of Schnorr's signature for Multi-signature are reviewed and presented.

A digital signature is the building block of a transaction in the blockchain. [46] addressed the problem of time consumption in multiple signature endorsement transactions. Their new schemes helped to achieve secure, transaction efficiency and low storage utilisation.Partial random signature generation attacks are the most common type of attack in a multi-signature environment. [47] proposed a Schnorr based multi-signature with a verifiable and deterministic nonce that can work non-interactively with a zero-knowledge proof. [48] proposed a multi-party computation protocol that is computationally cheap than the similar multi-signature model. They achieved a considerable contribution to privacy protection in blockchain. Their main idea was to merge and sign transactions under anonymous conditions using Pedersen commit with the Schnorr signature.[49] used a combination of identity and Schnorr to authenticate the mobile system. Their method was proposed as an alternative to certificate-based proxy methods and are secure against possible attacks. The multi-party ElGamal and Schnorr based signature for authentication proposed by [50] achieved multi-party computation across the unauthentic channel.[51] proposed a new Secret Handshake scheme with a Multi-Symptom Intersection derived from a Schnorr signature. They authorised Private Set Intersection only if their target authentication policies are satisfied to execute. [52] proposed a new single and multi blind signature scheme that combines the Schnorr signature schemes and RSA based. MuSig is a new Schnorr-based multi-signature scheme proposed by [53].The use of Schnorr signatures makes it simple, efficient and support key aggregation. In [54] the authors introduced a security model for general aggregate signature schemes based on multi-user and thereby achieved a significant reduction

EC/DSS/DSA	Domain	Solution	#
Edwards Curve ECC	IoT	Security enhancement	[9], [13], [44], [45], [17] , [76] [19], [43]

Table 4: Elliptic and Edwards curve application in IoT.

in key size. [55] proposed a novel trust management system based on ECC for the MANET and classified different trust levels and types of attackers. [56] focused their work on developing path-checking protocols to ensure that supply chains have valid paths. Their method achieved multi-signatures and verification based on a modified Schnorr signature scheme. [57] the authors used developers self-signature and centre's signature on Schnorr's based signature scheme and created an Android self-signature policy. [58] created a scheme for signatures with multivariate linear polynomials using Schnorr's signature scheme and ElGamal public key cryptography for verification based on a threshold. [59] presented a comprehensive resilient security framework for multipath routing wireless ad hoc networks. They were using a self-certified public key and an integrated multi-signature scheme to ensure secure data transfer. Table 5 on page 283 summarize multi-signature implementation using Schnorr signature.

3.5 Aggregate signature using Schnorr's signature

In this sub-section research works on various application of Aggregate signature using Schnorr's signature are reviewed and presented.

[60] proposed Key Aggregable Interactive Aggregate Signatures (KAIAS), which has a verification function that uses only a single aggregated public key, dynamic signature aggregation, and only allows messages to be signed only by the message initiator. These benefits benefited in reducing the size of signatures for Bitcoin implementation. [61] proposed an alternative to the current Bitcoin signature scheme with the Schnorr signature scheme, based aggregation scheme.Their protocol allowed participants to run a decentralised mixer on the Bitcoin blockchain to exchange coins. [62] demonstrated that the aggregate signature from groups actually works against the ordinary implementation of DSA based aggregate signatures with Schnorr's variants. Their proposed scheme had maximum performance and compatibility. [63] compared a non-interactive aggregate signature cryptographic scheme to the ECDSA cryptographic schemes with Schnorr. Their work showcased the need for enhancement with a non-interactive model. [64] analysed Mimblewimble's provable security and formally demonstrated that under standard assumptions, the inflation and coin theft can be provably secured using Pedersen commitments with Schnorr or Pedersen commitments with BLS signatures. [54] introduced a multi-user secure model for conventional aggregate signature schemes, in disparity to BGLS's original. They achieved a reduction of key-prefixed BLS security in a multi-user model. They also used Katz and Wang technique to demonstrate a reduction from a variant of multi-user key-prefixed [65] used lightweight identity-based Schnorr signature scheme and proposed an identity-based aggregate signature scheme variant where signer need not agree on common randomness. The scheme reduced time-consuming bilinear pairing operation, making it computationally effective. [76] proposed an aggregate signature scheme, which can be used in [77] to enhance the data privacy and has the potential to be applied in [78] an pipelined cryptography verification to provide a novel hybrid method.

Table 6 on page 283 shows the application of Schnorr's signature for Aggregate signature

4 Elliptic curve cryptography

The last decade showed a slow transition of the Digital signature from the RSA signature to DS, and towards ECC with modern inventions, the shift focused on optimised performance. Modern cryptography based on the ECDSA is the adoption trend due to its key length, signature length, security level and performance [66]. The Elliptic curve has taken a broad, proven space in cryptography, replacing its ancestor DSA. The elliptic curve has a successor called the edwards curve, and they are taking up the elliptic curve in cryptography space. edwards curve have a superior advantage and doubling and tripling than the Weierstrass form of the elliptic curve. Edwards addition laws do not have exceptions as in the Weierstrass curve. Elliptic curves concepts are widely used for ECC. Harold Edwards, in 2007 explores and studied the Elliptic curve family and presented a new addition named Edwards curves. The Edwards curve became the core of the Edwards curve digital signature algorithm (EdDSA). EdDSA offers standard performance and overcomes most of the security problems faced by conventional digital signature schemes (DSSs).

4.1 Elliptic Curve DSA (ECDSA) and Edwards Curve DSA (EdDSA)

Figure 4 represents the flow of EdDSA and ECDSA.

4.2 Basic operations

4.2.1 Addition

The Edwards curve handles addition and doubling using the same formula compared with another version of the curve that uses different operation formulas. Addition law is a century-old but captured cryptography attention in the recent decade. Unlike other elliptic curves that use chords

EC / DSS / DSA	Domain	Solution	#
Schnorr's signature	Blockchain		[46], [48]
	General	Security enhancement	[58]
	Healthcare		[51]
	Network		[59]
	Mobile	Security enhancement, Authentication	[49], [57]
	Supply chain		[56]
	Cryptocurrency	Security enhancement, Performance optimisation	[47]
ECC	Mobile	Security enhancement, Trust Management	[55]
Boneh-Gentry-Lynn	Security	Security enhancement	[54]

Table 5: Schnorr's signature for Multi-signature.

EC/DSS/DSA	Domain	Solution	#
Aggregate and Schnorr Signature	Blockchain	Verification speed	
		Security enhancement	[62]
		Storage optimisation	
	Cryptocurrency	Security enhancement	
		Computation improvement	
		Signature size	[60]
		Security enhancement	[61]
		Speed improvement	[63]
		Signature size-reduction	[64]
	Security Enhancement	Security enhancement	
		Security enhancement	
		Security enhancement	[54],[75]
		Identity authentication	[65],[73]

Table 6: Aggregate signature using Schnorr's signature.

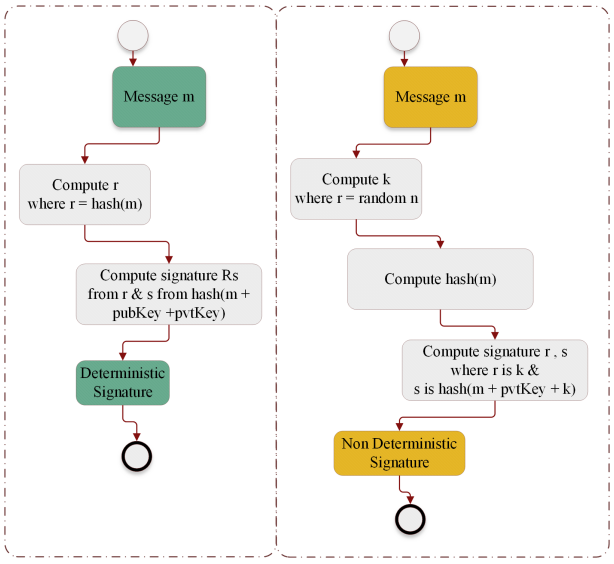


Figure 4: Flowchart of EdDSA and ECDSA.

and tangent to construct a point, the Edwards curve uses its method as a form of unit circle addition law. This says that if there are (x1,y1) and (x2,y2) in the Edwards curve the following (x3,y3) are known to be derived from the same curve such that x3 = (x1.y2 + x2.y1)/(a.(1+x1.y1x.2.y2)) and y3 = (y1.y2 — x1.x2)/(a.(1 — x1.y1x2.y2))

4.2.2 Doubling

Similarly, the doubling property can be applied by replacing (x2, y2) with (x1, y1) in the addition formula to obtain the doubling formula (x1, y1) + (x1, y1) = (x3, y3) such that x3 = (x1.y1 + x1.y1)/(a.(1+x1.y1.x1.y1)) and y3 = (y1.y1 — x1.x1)/(a.(1 — x1.y1.x1.y1)) This makes Edwards curves calculate a target point quickly [2].

4.2.3 Domain parameter and Key generation

Key generation starts with a self-generated private key and initializes the domain parameters. The generated private key is not accessible from outside by any third party. The public key is generated using the private key and the initialised parameter. This public key is accessed and readable from the outside. The private and public key pair ensure transmission protection.

4.2.4 Signature computation

The Edwards curve uses a hash of message instead of a random number, making this system collision-free and a different key for every message. Signature computation is based on the SHA algorithm are fixed length. The signature and the private key are used to generate a signature and sign the message. This digital signature allows the receiver to determine authenticity and offer non-repudiation.

4.2.5 Signature verification

Signature verification is a process by which the system can determine the authenticity of the message using the public key. The operation required the public key, digital signature and message.

Table 7 on page 285 represents the arithmetic operation cost comparison of the Edwards curve with other curves used in ECC.

4.3 Elliptic vs Edwards curve

EdDSA is a deterministic elliptic curve signature scheme with two curves Ed25519 and Ed448 [68]. ECDSA relies on cyclic groups over the finite field of the curve and is of discrete logarithm problem and a variant of the ElGamal signature scheme. The more improved version EdDSA is a variant based on Schnorr's signature scheme. EdDSA is simple, secure and fast compare with ECDSA.

4.3.1 Comparison of EdDSA vs ECDSA

The comparison of basic arithmetic, group law, and prime order are optimal for EdDSA. Curve safety is high in EdDSA. The performance of EdDSA is high in the segment and prevents security flaws. In the case of key loss or stolen its impossible to recovery in EdDSA. All of this is presented in 8 on page 285. Table 8 on page 285 tabulates the Comparison of EdDSA vs ECDSA as derived from [68] and [69]. Table 9 on page 285 presents the parameter comparison as presented in [70].

4.3.2 Security comparison

As presented in Figure 5 on page 284, the Edwards curve security level is far stronger at all security levels of 128,192 and 256 compared with the Weierstrass curve. Security level comparison of rho complexity is derived from [71] and [72].

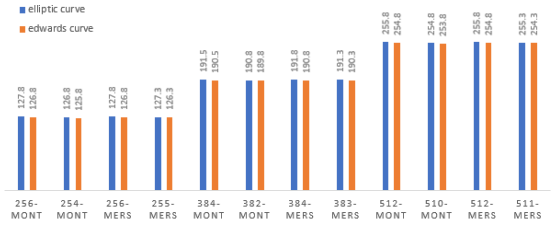


Figure 5: Security level[71] and complexity comparison[72].

4.3.3 Cost comparison

Figure 6 on page 284 shows the comparison of the total cost of TLS handshake both in compressed (C) and uncompressed (U) formats, illustrating that the Edwards curve has

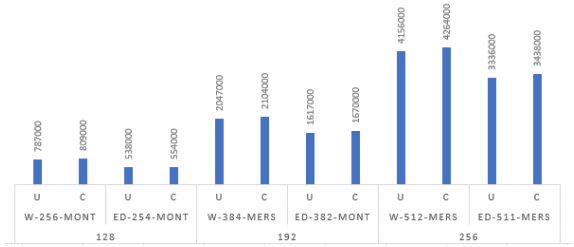


Figure 6: Cost estimate comparison[72].

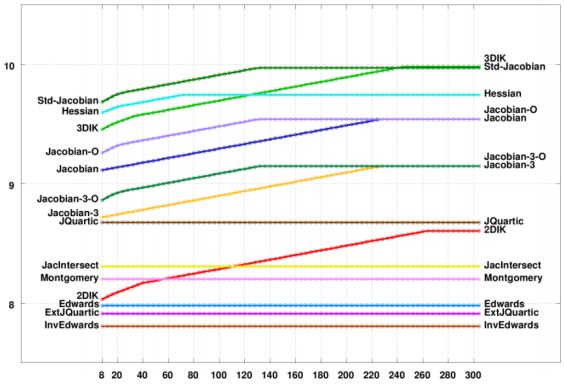


Figure 7: Speed comparison of curves [73].

the best values compared to the elliptic curve. Total cost estimate derived from [72]

4.3.4 Computation comparison

Figure 7 on page 284 shows the computation comparison on various curves. The Edwards curve has the least requirement, benefiting its implementation across various domains where the computation resource limits and requires speedy computation.

5 Discussion

Evidently, as noted by Bill gates in [74], the computing society and researcher should ensure a continuous disaster recovery plan and immediately switch to an alternative method when an existing security method proves fallible. This brings in the responsibility of the research community to explore and keep a standby successor digital signature in the event of a compromising situation or as an improvement over the existing digital signature that supports the security of digital technology. Elliptic curves and their variants have been re-purposed and used widely since 1980; Our empirical study demonstrates that the Edwards variant of the curve can be considered a performance improvement alternative for application in blockchain and IoT domains. Our review of the existing works shows the following 1).Application of ECC in Blockchain and IoT 2).Application

Curve	ADD	reADD	mADD	DBL	UNI
Edwards	10M +1S+1D	10M+1S+1D	9M+1S+1D	3M+4S	10M+1S+1D
Hessian	12M	12M	10M	7M+1S	12M
Jacobi intersection	13M+2S+1D	11M+2S+1D	11M+2S+1D	3M+4S	13M+2S+1D
Jacobi quartic	10M+3S+1D	9M+3S+1D	8M+3S+1D	2M+6S+2D	10M+3S+1D
Projective	12M+2S	12M+2S	9M+2S	5M+6S+1D	11N+6S+1D

Table 7: Comparison of arithmetic operation cost [67].

Attribute	Edwards Curve DSA	Elliptic Curve DSA
Curves	ax^2 + yx^2 = 1 + dx^2y^2	y^2=x^3+ax+b
Signature scheme	Schnorr signature scheme	ElGamal signature scheme
Performance	Faster	Slower
Order	Not prime order	Prime order possible
Key recover	Not possible	Possible
Curve safety	More	Less
Curve form	General	Subset
Curve arithmetic	Faster addition	Slower
Group law	Complete	Exception

Table 8: Comparison of EdDSA vs ECDSA [68] [69].

Parameters	ECDSA	EdDSA
Length of Key	384b	10b
Time take for -		
key generation(sec)	0.799s	0.0006s
sign generation(sec)	0.0016s	0.0002s
sign verification(sec)	0.0082s	0.0007s

Table 9: Parameter comparison [70].

of Schnorr's signature for Multi and Aggregate signature demonstrating. The application of ECC and Schnorr's signature shown in all these work supports how the digital signature is part of almost all digital applications. Our quest to find a performance improved alternative was solved with superior primary basic operation properties on the Edwards curve over the elliptic curve. 1).Addition used unit circle addition law 2).Doubling had a fast target point calculation than the elliptic curve 3).The private key and domain parameter are not accessible only the public key is exposed for external use 4).The Signature computation is based on hash, and 5).The verification is few expensive than other curve forms as tabulated in Table 7 on page 285 , Table 8 on page 285 and Table 9 on page 285. Direct secondary level comparison on the Edwards vs elliptic curve was strong, sound and clear on how Edwards outstood the predecessor. Edwards curve implementation was able to achieve 1).A higher security level as illustrated in Figure 5 on page 284 2).competitive cost values as illustrated in Figure 6 on page 284 3).speedy computation with fewer resource requirements than elliptic curve based implementation as illustrated in Figure 7 on page 284. Using Edwards curve in building Cross-domain Applications in Internet of Things such as [77] would provide performance improvement.

6 Conclusion

In this empirical study, we have provided a comparison of EdDSA vs ECDSA and concluded that EdDSA has advantages over similar DSAs. The Edwards curve performs simple and faster arithmetic and has high performance on various applications. Signature generation does not mandate the use of unique random numbers. An attack on the system built using the Edwards curve is not catastrophic. The key size and signature have small footprints; moreover, they are complete and hash collision-resistant. EdDSA is better than ECDSA and is recommended as a better replacement but depends on the use case. ECDSA is still in use on Bitcoin and Ethereum, as signature recovery is easy compared to EdDSA.

References

[1] Kerry C.F. and Gallagher P.D., "Digital signature standard (DSS)," FIPS PUB, pp. 186-4, 2013. https://doi.org/10.6028/nist.fips.186-4

[2] Edwards H.M., "A normal form for elliptic curves," Bulletin of the American Mathematical Society, vol. 44, no. 03, pp. 393- 423, 2007. https://doi.org/10.1090/s0273-0979-07-01153-6

[3] Laska M., "An algorithm for finding a minimal Weierstrass equation for an elliptic curve," Mathematics of Computation, vol. 38, no. 157, pp. 257-257, 1982. https://doi.org/10.1090/s0025-5718-1982-0637305-2

[1] Peretti C., Leoncini A., Gastaldo P., and Zunino R., "Edwards Curves and Extended Jacobi Quartic-Curves for Efficient Support of Elliptic-Curve Cryptosystems in Embedded Systems,"

International Journal for Information Security Research, vol. 4, no. 3, pp. 449-458, 2014. https://doi.org/10.20533/ijisr.2042.4639. 2014.0052

[2] Smart N.P., "The Hessian Form of an Elliptic Curve," in Cryptographic Hardware and Embedded Systems - CHES 2001, pp. 118-125, Springer, 2001. https://doi.org/10.1007/3-540-44709-1_11

[3] Aggarwal S. and Kumar N., "Digital signatures," Advances in Computers, The Blockchain Technology for Secure and Smart Applications across Industry Verticals, pp. 95-107, 2021. https://doi.org/10.1016/bs.adcom.2020.08.004

[4] DuPont B., Franck C., and Großschädl J., "Fast and Flexible Elliptic Curve Cryptography for Dining Cryptographers Networks," Mobile, Secure, and Programmable Networking, pp. 89-109, 2021. https://doi.org/10.1007/978-3-030-67550-9_7

[5] Kirlar B.B., "Efficient message transmission via twisted Edwards curves," Mathematica Slovaca, vol. 70, no. 6, pp. 1511- 1520, 2020. https://doi.org/10.1515/ms-2017-0444

[6] Islam M.M., Hossain M.S., Hasan M.K., Shahjalal M., and Jang Y.M., "Design and Implementation of High-Performance ECC Processor with Unified Point Addition on Twisted Edwards Curve," Sensors, vol. 20, no. 18, 2020. https://doi.org/10.3390/s20185148

[7] Semmouni M.C., Nitaj A., and Belkasmi M., "Bitcoin security with a twisted Edwards curve," Journal of Discrete Mathematical Sciences and Cryptography, pp. 1-19, 2020. https://doi.org/10.1080/09720529.2019. 1681673

[8] Skuratovskii R. and Osadchyy V., "The Order of Edwards and Montgomery Curves," WSEAS TRANSACTIONS ON MATHEMATICS, vol. 19, pp. 253-264, 2020. https://doi.org/10.37394/23206.2020.19.25

[9] Hisil H. and Renes J., "On Kummer Lines with Full Rational 2-torsion and Their Usage in Cryptography," ACM Transactions on Mathematical Software, vol. 45, no. 4, pp. 1-17, 2019. https://doi.org/10.1145/3361680

[10] Mehrabi M.A. and Doche C., "Low-Cost, Low-Power FPGA Implementation of ED25519 and CURVE25519 Point Multiplication," 2019. https://doi.org/10.3390/info10090285

[11] Faz-Hernández A., López J., and Dahab R., "High-performance Implementation of Elliptic Curve Cryptography Using Vector Instructions," ACM Transactions on Mathematical Software, vol. 45, no. 3, pp. 1-35, 2019. https://doi.org/10.1145/3309759

[12] Hu Z., Gnatyuk S., Kovtun M., and Seilova N., "Method of Searching Birationally Equivalent Edwards Curves Over Binary Fields," Advances in Intelligent Systems and Computing, pp. 309-319, 2018. https://doi.org/10.1007/978-3-319-91008-6_31

[13] Islam M.M., Hossain M.S., Hasan M.K., Shahjalal M., and Jang Y.M., "FPGA Implementation of High-Speed Area- Efficient Processor for Elliptic Curve Point Multiplication Over Prime Field," IEEE Access, vol. 7, pp. 178811-178826, 2019. https://doi.org/10.1109/access.2019.2958491

[14] Seo H. and Kim H., "MoTE-ECC based encryption on MSP430," Journal of Information and Communication Convergence Engineering, vol. 15, no. 10, pp. 160-164, 2017. https://doi.org/10.6109/jicce.2017.15.3.160

[15] Franck C. and Großschädl J., "Efficient Implementation of Pedersen Commitments Using Twisted Edwards Curves," Mobile, Secure, and Programmable Networking, pp. 1-17, 2017. https://doi.org/10.1007/978-3-319-67807-8_1

[16] Liu Z., Großschädl J., Hu Z., Jarvinen K., Wang H., and Verbauwhede I., "Elliptic Curve Cryptography with Efficiently Computable Endomorphisms and Its Hardware Implementations for the Internet of Things," IEEE Transactions on Computers, vol. 66, no. 5, pp. 773-785, 2017. https://doi.org/10.1109/tc.2016.2623609

[17] Karati S. and Das A., "Batch Verification of EdDSA Signatures," Security, Privacy, and Applied Cryptography Engineering, pp. 256-271, 2014. https://doi.org/10.1007/978-3-319-12060-7_17

[18] Liu Z., Weng J., Hu Z., and Seo H., "Efficient Elliptic Curve Cryptography for Embedded Devices," ACM Transactions on Embedded Computing Systems, vol. 16, no. 2, pp. 1-18, 2017. https://doi.org/10.1145/2967103

[19] Naresh V.S., Reddi S., and Allavarpu V.D., "Blockchain-based patient centric health care communication system," International Journal of Communication Systems, vol. 34, no. 7, pp. 34-34, 2021. https://doi.org/10.1002/dac.4749

[20] Saini A., Zhu Q., Singh N., Xiang Y., Gao L., and Zhang Y., "A Smart-Contract-Based Access Control Framework for Cloud Smart Healthcare System," IEEE Internet of Things Journal, vol. 8, no. 7, pp. 5914-5925, 2021. https://doi.org/10.1109/jiot.2020.3032997

[21] Jasem F.M., Sagheer A.M., and Awad A.M., "Enhancement of digital signature algorithm in bitcoin

wallet," Bulletin of Electrical Engineering and Informatics, vol. 10, no. 1, pp. 449-457, 2021. https://doi.org/10.11591/eei.v10i1.2339

[22] Sadiq A., Javed M.U., Khalid R., Almogren A., Shafiq M., and Javaid N., "Blockchain Based Data and Energy Trading in Internet of Electric Vehicles," IEEE Access, vol. 9, pp. 7000-7020, 2021. https://doi.org/10.1109/access.2020.3048169

[23] Arulprakash M. and Jebakumar R., "People-centric collective intelligence: decentralised and enhanced privacy mobile crowd sensing based on blockchain," The Journal of Supercomputing, 2021. https://doi.org/10.1007/s11227-021-03756-x

[24] Kavin B.P., Ganapathy S., Kanimozhi U., and Kannan A., "An Enhanced Security Framework for Secured Data Storage and Communications in Cloud Using ECC, Access Control and LDSA," 2020. https://doi.org/10.1007/s11277-020-07613-7

[25] Benil T. and Jasper J., "Cloud based security on outsourcing using blockchain in E-health systems," Computer Networks, vol. 178, pp. 107344-107344, 2020. https://doi.org/10.1016/j.comnet.2020.107344

[26] Wang H., He D., and Ji Y., "Designated-verifier proof of assets for bitcoin exchange using elliptic curve cryptography, "Future Generation Computer Systems, vol. 107, pp. 854-862, 2020. https://doi.org/10.1016/j.future.2017.06.028

[27] Kumar M., Chand S., and Katti C.P., "A Secure End-to-End Verifiable Internet-Voting System Using Identity-Based Blind Signature," IEEE Systems Journal, vol. 14, no. 2, pp. 2032-2041, 2020. https://doi.org/10.1109/jsyst.2019.2940474

[28] Li X., Mei Y., Gong J., Xiang F., and Sun Z., "A Blockchain Privacy Protection Scheme Based on Ring Signature," IEEE Access, vol. 8, pp. 76765-76772, 2020. https://doi.org/10.1109/access.2020.2987831

[29] Ernest B. and Shiguang J., "Privacy Enhancement Scheme (PES) in a Blockchain-Edge Computing Environment," IEEE Access, vol. 8, pp. 25863-25876, 2020. https://doi.org/10.1109/access.2020.2968621

[30] Zhang X., Zhou Z., Zhang J., Xu C., and Zhang X., "Efficient lightweight private auditing scheme for cloud-based wireless body area networks," International Journal of Electronic Security and Digital Forensics, vol. 12, no. 2, pp. 139-139, 2020. https://doi.org/10.1504/ijesdf.2020.10027592

[31] Ansah A.K.K. and Gyamfi D.A., "Enhancing user and transaction privacy in bitcoin with unlinkable coin mixing scheme," International Journal of Computational Science and Engineering, vol. 23, no. 4, 2020. https://doi.org/10.1504/ijcse.2020.10035561

[32] Chen C.L., Deng Y.Y., Weng W., Chen C.H., Chiu Y.J., and Wu C.M., "A Traceable and Privacy-Preserving Authentication for UAV Communication Control System," Electronics, vol. 9, no. 1, 2020. https://doi.org/10.3390/electronics9010062

[33] Ullah I., Amin N.U., Almogren A., Khan M.A., Uddin M.I., and Hua Q., "A Lightweight and Secured Certificate-Based Proxy Signcryption (CB-PS) Scheme for E-Prescription Systems," IEEE Access, vol. 8, pp. 199197-199212, 2020. https://doi.org/10.1109/access.2020.3033758

[34] Zhang X., Zhao J., Mu L., Tang Y., and Xu C., "Identity-based proxy-oriented outsourcing with public auditing in cloud- based medical cyber-physical systems," Pervasive and Mobile Computing, vol. 56, pp. 18-28, 2019. https://doi.org/10.1016/j.pmcj.2019.03.004

[35] Taleb N., "Prospective applications of blockchain and bitcoin cryptocurrency technology," TEM Journal, vol. 8, no. 03, pp. 48-55, 2019. https://dx.doi.org/10.18421/TEM81-06

[36] Chen X. and Zhang X., "Secure Electricity Trading and Incentive Contract Model for Electric Vehicle Based on Energy Blockchain," IEEE Access, vol. 7, pp. 178763-178778, 2019. https://doi.org/10.1109/access.2019.2958122

[37] Liu Y., Liu X., Tang C., Wang J., and Zhang L., "Unlinkable Coin Mixing Scheme for Transaction Privacy Enhancement of Bitcoin," IEEE Access, vol. 6, pp. 23261-23270, 2018. https://doi.org/10.1109/access.2018.2827163

[38] Yuan C., xue Xu M., and ming Si X., "Research on a new signature scheme on blockchain," Security and Communication Networks, vol. 2017, 2017. https://doi.org/10.1155/2017/4746586

[39] Sajjad A., Afzal M., Iqbal M.M.W., Abbas H., Latif R., and Raza R.A., "Kleptographic Attack on Elliptic Curve Based Cryptographic Protocols," IEEE Access, vol. 8, pp. 139903-139917, 2020. https://doi.org/10.1109/access.2020.3012823

[40] Lara-Nino C.A., Diaz-Perez A., and Morales-Sandoval M., "Lightweight elliptic curve cryptography accelerator for internet of things applications," 2020. https://doi.org/10.1016/j.adhoc.2020.102159

[41] Lara-Nino C., Diaz-Perez A., and Morales-Sandoval M., "Energy/Area-Efficient Scalar Multiplication with Binary Edwards Curves for the IoT," Sensors, vol. 19, no. 3, pp. 720-720, 2019. https://doi.org/10.3390/s19030720

[42] Loiseau A., Fournier J.A., and J, "Binary Edwards Curves for Intrinsically Secure ECC Implementations for the IoT," Proceedings of the 15th International Joint Conference on e-Business and Telecommunications. International Conference on Security and Cryptography, 2018. https://doi.org/10.5220/0006831506250631

[43] Xiao Y., Zhang P., and Liu Y., "Secure and Efficient Multi-Signature Schemes for Fabric: An Enterprise Blockchain Platform," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1782-1794, 2021. https://doi.org/10.1109/tifs.2020.3042070

[44] Nick J., Ruffing T., Seurin Y., and Wuille P., "MuSig-DN: Schnorr Multi-Signatures with Verifiably Deterministic Nonces," Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020. https://doi.org/10.1145/3372297.3417236

[45] Feng L., Jie Y., Deli K., and Jiayin Q., "A Secure Multiparty Computation Protocol Combines Pederson Commitment with Schnorr Signature for Blockchain," 2020 IEEE 20th International Conference on Communication Technology (ICCT). IEEE, 2020. https://doi.org/10.1109/icct50939.2020.9295819

[46] Sanae H., Laassiri J., and Berguig Y., "MULTI-AGENT identity combined key Signature authentication PROTOCOL based schnorr signature with provable security under AVISPA," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 5, pp. 7628-7635, 2020. https://doi.org/10.30534/ijatcse/ 2020/102952020

[47] Vu D.H., Luong T.D., and Ho T.B., "An efficient approach for secure multiparty computation without authenticated channel," Information Sciences, vol. 527, pp. 356-368, 2020. https://doi.org/10.1016/j.ins.2019.07.031

[48] Wen Y., Zhang F., Wang H., Gong Z., Miao Y., and Deng Y., "A new secret handshake scheme with multisymptom intersection for mobile healthcare social networks," Information Sciences, vol. 520, pp. 142-154, 2020. https://doi.org/10.1016/j.ins.2020.02.007

[49] Tan D.N., Nam H.N., Hieu M.N., and Van H.N., "New Blind Multisignature Schemes based on ECDLP," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 2, pp. 1074-1074, 2018. https://doi.org/10.11591/ijece.v8i2.pp1074-1083

[50] Maxwell G., Poelstra A., Seurin Y., and Wuille P., "Simple Schnorr multisignatures with applications to Bitcoin," 2019. https://doi.org/10.1007/s10623-019-00608-x

[51] Lacharité M.S., "Security of BLS and BGLS signatures in a multiuser setting," Cryptography and Communications, vol. 10, no. 1, pp. 41-58, 2018. https://doi.org/10.1007/s12095-017-0253-6

[52] Singh O., Singh J., and Singh R., "Multilevel trust based intelligence intrusion detection system to detect the malicious nodes using elliptic curve cryptography in MANET," Cluster Computing, vol. 21, pp. 51-63, 2018. https://doi.org/10.1007/s10586-017-0927-z

[53] Xin W., Wang M., Shao S., Wang Z., and Zhang, "A variant of schnorr signature scheme for path-checking in RFID- based supply chains," 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). IEEE, 2015. https://doi.org/10.1109/fskd.2015.7382368

[54] Lee H.C., Jung J.H., and Yi J.H., "Multi-Signature Based Tamper Detection Scheme of Android Applications," Sensor Letters, vol. 11, no. 9, pp. 1820-1827, 2013. https://doi.org/10.1166/sl.2013.3004

[55] Shen Z. and Yu X., "Threshold signature scheme with threshold verification based on multivariate linear polynomial," Journal of Shanghai Jiaotong University (Science), vol. 16, no. 5, pp. 551-556, 2011. https://doi.org/10.1007/s12204-011-1186-4

[56] Vaidya B., Makrakis D., Park J.H., and Yeo S.S., "Resilient Security Mechanism for Wireless Ad hoc Network," Wireless Personal Communications, vol. 56, no. 3, pp. 385-401, 2011. https://doi.org/10.1007/s11277-010-9978-7

[57] Kojima R., Yamamoto D., Shimoyama T., Yasaki K., and Nimura K., "A Novel Scheme of Schnorr Multisignatures for Multiple Messages with Key Aggregation," Lecture Notes in Networks and Systems,Advances on Broad-Band Wireless Computing, Communication and Applications, pp. 284-295, 2020. https://doi.org/10.1007/978-3-030-61108-8_28

[58] Barbara F. and Schifanella C., "DMix: decentralised mixer for unlinkability," 2020 2nd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS). IEEE, 2020. https://doi.org/10.1109/brains49436.2020. 9223282

[59] Zhao Y., "Practical Aggregate Signature from General Elliptic Curves, and Applications to Blockchain," ACM Asia Conference on Computer and Communications Security, 2019. https://doi.org/10.1145/3321705.3329826

[60] Pedrosa A.R., Potop-Butucaru M., and Tucci-Piergiovanni S., "Scalable lightning factories for Bitcoin," The 34th ACM/SIGAPP Symposium on Applied Computing. ACM, 2019. https://doi.org/10.1145/3297280.3297312

[61] Fuchsbauer G., Orrù M., and Seurin Y., "Aggregate Cash Systems: A Cryptographic Investigation of Mimblewimble," Advances in Cryptology - EUROCRYPT 2019, pp. 657-689, 2019. https://doi.org/10.1007/978-3-030-17653-2_22

[62] Selvi S.S.D., Vivek S.S., Shriram J., and Rangan C.P., "Identity based partial aggregate signature scheme without pairing," 35th IEEE Sarnoff Symposium. 2012 35th IEEE Sarnoff Symposium, 2012. https://doi.org/10.1109/sarnof.2012.6222731

[63] Sury O., "Use of the SHA-256 Algorithm with RSA, Digital Signature Algorithm (DSA), and Elliptic Curve DSA (ECDSA) in SSHFP Resource Records," Request for Comments, vol. 6594, 2012. https://doi.org/10.17487/rfc6594

[64] Bernstein D.J. and Lange T., "Faster Addition and Doubling on Elliptic Curves," in Advances in Cryptology - ASIACRYPT 2007, pp. 29-50, Springer. https://doi.org/10.1007/978-3-540-76900-2_3

[65] Josefsson S. and Liusvaara I., "Edwards-Curve Digital Signature Algorithm (EdDSA)," Internet Research Task Force, Crypto Forum Research Group, RFC, vol. 8032, pp. 257-260, 2017. https://doi.org/10.17487/rfc8032

[66] Pornin T., "Deterministic usage of the digital signature algorithm (DSA) and elliptic curve digital signature algorithm (ECDSA)," Internet Engineering Task Force RFC, vol. 6979, pp. 1-79, 2013. https://doi.org/10.17487/rfc6979

[67] Shivani Y.N., Srinivas A., Thanmayi B.K., Vignesh V., and Srividya B.V., "EdDSA Over Galois Field GF(pm̂) for Multimedia Data," Journal of Engineering Research and Reports, pp. 1-7, 2019. https://doi.org/10.9734/jerr/2019/v4i416911

[68] B. Black, J. W. Bos, C. Costello, P. Longa, and M. Naehrig, Elliptic curve cryptography (ECC) nothing up my sleeve (NUMS) curves and curve generation. 2014.

[69] Bos J.W., Costello C., Longa P., and Naehrig M., "Selecting elliptic curves for cryptography: an efficiency and security analysis," Journal of Cryptographic Engineering, vol. 6, no. 4, pp. 259-286, 2016. https://doi.org/10.1007/s13389-015-0097-y

[70] Bernstein D.J. and Lange T., "Analysis and optimisation of elliptic-curve single-scalar multiplication Data set," Finite Fields and Applications,Contemporary Mathematics, pp. 1-19, 2008. https://doi.org/10.1090/conm/461/08979

[71] Gates, B., Myhrvold, N., Rinearson, P. and Domonkos, D., The road ahead. London, England: Viking, 1995.

[72] Guruprakash J, Koppu S. EC-ElGamal and genetic algorithm-based enhancement for lightweight scalable blockchain in IoT domain. IEEE access: practical innovations, open solutions. 2020;8:141269–141281. http://dx.doi.org/10.1109/access. 2020.3013282

[73] Abouelkheir, Eman, and Jolanda G. Tromp. "A pairing free secure identity-based aggregate signature scheme under random oracle." Informatica 42.2 (2017).

[74] Gu K, Yang L, Liu Y, Yin B. Trajectory data privacy protection based on differential privacy mechanism. Informatica. An International Journal of Computing and Informatics. 2018;42(3). http://dx.doi.org/10.31449/inf.v42i3.1638

[75] Bitat A, Merniz S. Formal verification of pipelined cryptographic circuits: A functional approach. Informatica. An International Journal of Computing and Informatics. 2021;45(4). http://dx.doi.org/10.31449/inf.v45i4.3176

[76] Jayabalasamy G, Koppu S. High-performance Edwards curve aggregate signature (HECAS) for non-repudiation in IoT-based applications built on the blockchain ecosystem. Journal of King Saud University - Computer and Information Sciences. 2021. http://dx.doi.org/10.1016/j.jksuci.2021.12.001

[77] Benkhaled S, Hemam M, Djezzar M, Maimour M. An ontology – based contextual approach for cross-domain applications in internet of things. Informatica. An International Journal of Computing and Informatics. 2022;46(5). http://dx.doi.org/10.31449/inf.v46i5.3627

Deep Reinforcement Learning-based Anomaly Detection for Video Surveillance

Sabrina Aberkane and Mohamed Elarbi-Boudihir
E-mail: s.aberkane@esi-sba.dz, m.elarbiboudihir@esi-sba.dz
ESI-SBA, High school of Computer Sciences, Sidi Bel-Abbès, Algeria

Student paper

Keywords: deep reinforcement learning, anomaly detection, video surveillance

Received: June 22, 2021

The anomaly detection in automated video surveillance is considered as one of the most critical tasks to be solved, in which we aim to detect a variety of real-world abnormalities. This paper introduces a novel approach for anomaly detection based on deep reinforcement learning. In recent years, deep reinforcement learning has been achieving a significant success in various applications with data of a high degree of complexity such as robotics and games, by mimicking the way humans learn from experiences. Generally, the state-of-the-art methods classify a video as normal or abnormal without pinpointing the exact location of the anomaly in the input video due to the unlabeled clip-level data in training videos. We focus on adapting the prioritized Dueling deep Q-networks to the anomaly detection problem. This model learns to evaluate the anomaly in video clips by exploiting the video-level label to obtain a better detection accuracy. Extensive experiments on 13 cases class of real-word anomaly show that our DRL agent achieved a near optimal performance with a high accuracy in the real world video surveillance system compared to the state-of-the-art approaches.

Povzetek: Razvita je nova metoda globokega spodbujevalnega učenja za prepoznavanje anomalij pri videonadzoru.

1 Introduction

In video surveillance systems, the ability to recognize actions can be used to detect and prevent abnormal or suspicious events. Such intelligent systems would be greatly helpful for providing security to people. Indeed, the surveillance cameras also make some people feel more safe, knowing that the culprits are being watched. Generally, these kinds of systems are powered by different algorithms[1, 2, 3], which are action recognition, object tracking and object classification. The conception of such algorithms is typically addressed in computer vision research which works on how to make machines gain some human understanding of data from digital images and videos.

In this work, we focus on designing an intelligent visual surveillance system, which aims to detect abnormalities in urban places. The anomaly detection task has been one of the most talked about issues for decades, and is still a very hot topic due to the broad real-world applications including visual surveillance. To address the abnormality detection problem, some researchers attempted to give a general definition, which covers all existing normal/abnormal motions in daily life. Otherwise, many workers considered the task as an activity classification problem. All these researches share one main purpose which is to build an intelligent machine imitating the human capability of interpreting complex human behaviors in a cluttered environment. Is it possible that a machine could perform the recognition task at the same level as humans?

In our paper, we try to answer the question above by demonstrating that a machine can be as efficient as a human as long as it succeeds in reproducing the human's native learning mechanism. Indeed, we consider building an agent able to learn from the environment through a sequence of trial/error. The video analytics framework takes a video clip as input, then the pre-trained agent will provide two principal elements separately: the first is an estimation of the existence of an abnormal content in the video. The second, indicates the anomaly score for each segment in the video. The system architecture is inspired by a trending approach called deep reinforcement learning, which is a branch of machine learning based on the concept that an agent learns from interacting with an environment. The agent training was done through a new large-scale dataset of 1900 videos, 128 hours long, untrimmed real-world surveillance footage, with 13 cases of realistic abnormalities.

The organization of this paper is given as follows: After the introduction in this section, we will present the state-of-the-art anomaly detection approach in section 2. Subsequently, we introduce how to implement the system using the Dueling DQN as well as the anomaly localization in section 3. Section 4 will present the results and conclu-

System	Techniques	Scene	Localization	Dataset(s)	Accuracy
Schuldt et al [4]	Patterns Represtation SVM classification	Outdoor Indoor Uncrowded	Disable	Action database (available on request)	71,70 %
Hu et al [6]	Modeling trajectories Cluster-based	Outdoor crowded traffic	Disable	Action database (available on request)	80%
Qiao et al [7]	Modeling Optical flow Deep autoencorder	Outdoor Indoor	Disable	Lawn indoor plaza	98.33%
Khaleghi et al [8]	deep learning	Outdoor Indoor	Enable	UCSD dataset	88.1%
Shean Chong et al [9]	Spatiotemporal architecture Convolutionnel network Autoencoder	Outdoor Crowded	Enable	UCSD dataset	89.9%
Hasan et al [10]	Learning pattern model Autoencoder	Outdoor Indoor Crowded	Enable	CUHK Avenue UCSD Ped1 UCSD Ped2 Subway Entrance Subway Exit	70,2% 81.0% 90.0% 94.3% 80.7%
Sultani et al [11]	Multiple instance learning deep learning	Outdoor Crowded	Enable	UCF Dataset	75.41%
Oh et al [12]	Reinforcement learning	-	Disable	GeoLife GPS TST	35% 93%

Table 1: The comparison of properties between state-of-the-art approaches.

sions are finally made in Section 5.

2 Related work

The initial studies on anomaly detection have been reported in [4, 5, 6, 7], where the systems model the normal motion of individuals as trajectory, the anomaly is detected as a deviation from that normal trajectory. More recently, the following works used deep learning, which achieves competitive performances in video data. In this paper [8] a deep learning-based technique is used on both features extraction phase and rare events detection phase. The authors in [9] employ a spatiotemporal autoencoder to design a framework for events detection, which is composed of both spatial feature representation and the learning temporal evolution of the spatial features. Hasan et al [10] also used deep learning with autoencoders to present a fully connected autoencoder to learn the model of anomaly detection. To learn anomalous events, Waqas et al [11] constructed a new framework based on a deep multiple instance learning which leverages weakly labeled training videos. The authors in [12] applied the inverse reinforcement learning (IRL) for sequential anomaly detection, the system captures the sequence of actions of a target agent as input data, to return observation and evaluate whether it follows a normal pattern or not. The proposed approach works with a reward function which is inferred via IRL.

Table 1 compares the properties between previous systems. 'Scene' indicates where the anomaly occurs and the number of individuals on-site(crowded, uncrowded); 'Localization' specifies the property to locate where the anomaly is occurring. The 'scene' column of [12] is indicate as (-), because the used dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude and altitude.

3 System modelisation

In this paper, we formulate the anomaly detection as a sequential decision-making process, then we propose the concept of a deep anomaly detection network to estimate the probability of covering an abnormality for each video segment. We assume that for a given video only a small number of segments contain the anomaly. Hence, we employ the reinforcement learning approach to train our detection network, which encourages high scores anomalous video segments as compared to normal segments. The latter is the equivalent of going through the process of finding the N segments with the highest abnormality scores from an input video.

Figure 2: A sample of a distribution of anomaly segments in an abnormal video(red).

Deep reinforcement learning offers two different struc-

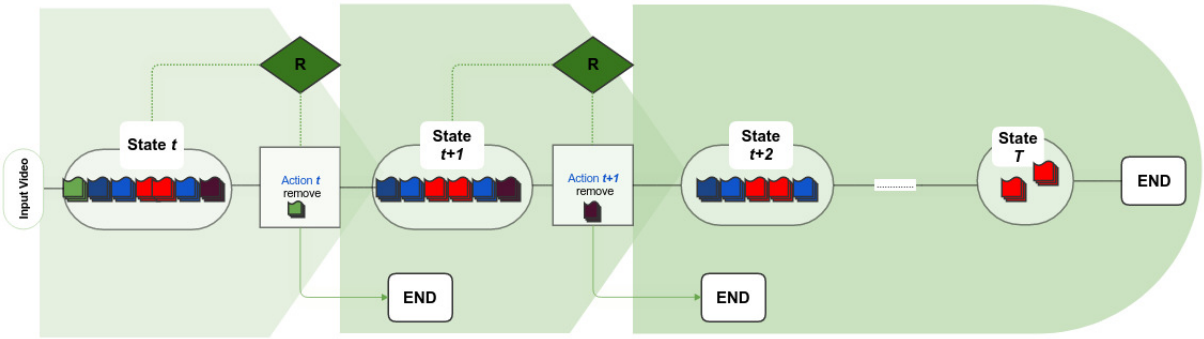


Figure 1: The system selection process.

tures where a machine can teach itself based on the results of its actions. One is the Deep Q-network DQN structure which only relies on the evaluation of actions to make decisions, while the other is the Deep Dueling Q-network DDQN that unlike the first, it takes advantage of both action value and environment information.

We adopted the dueling structure that was introduced by Wang et al [13], which explicitly sets apart the representation of state value and the state-dependent actions advantage via two separate streams.

$$Q(s,a) = V(s) + A(s,a) \tag{1}$$

$A(s,a)$ denotes the advantage stream that outputs a vector having dimensionality equal to the number of actions, representing the value of selecting an action a_i at state s_t. $V(s)$ denotes the value stream which outputs a scalar to represent the value of state s_t . The Value of a state is independent of actions. Both streams are combined at the end to produce the Q-function estimate through the combining module that can simply aggregate the value and the advantage streams as in [14]. The final output is a set of Q values $Q(s,a)$, one for each action.

As known, The markov decision process MDP is the underlying basis of any Reinforcement Learning model. Thus, to train the reinforcement learning agent for detecting abnormal events or behaviors on a video, it's very crucial to structure the environment of the processed video to the agent in some way to satisfy the Markov property, which is defined by a tuple (S, A, P, R, γ) such as states S, actions A, etc.

State: we consider that an input video V contains N segment x_i where $V = \sum_{i=1}^{N} x_i$. The state of the environment is a set of video segments $S = [x_1^V, x_2^V, x_3^V, ..., x_N^V]$. Where the initial state S_0 is composed of $N = 32$ segments, and for the next states, the initial number of segments decreases until it reaches the minimum number allowed $N = 5$, which is defined as the terminal state.

Action: An action $a_i \in A$ is every action executed by the RL agent to achieve its final objective, which is finding the set of segments x_i that are covering the anomaly to trigger an alarm.

Reward: In the literature, we find reward/penalization functions that are quite simple, as used in games [15], others much less simple like the one defined in [16]. Generally, in more complex environments, the reward function must be designed in such a way to suit the agent's environment to reflect what the agent has actually learned from the last episodes.

We can describe our environment as real, complex and unpredictable. In this context we based the rewarding/penalizing scheme on two axes:

- **Actions A related to segments X from normal videos V_n:** The reward value $r(V_n, a_i)$ reached the maximum when $Q(s_t, a_i) = 0$. The agent is penalized when $Q(s_t, a_i) = 1$, this means that the action is judged corresponding to the database annotations only.

- **Actions A related to segments X from abnormal videos V_{ab}:** The reward value $r(V_{ab}, a_i)$ reached the maximum when $Q(s_t, a_i)$ is similar to Expert E_c evaluation $Q(s_t, a_i) = E_c(s_t, a_i)$, and the minimum when $|Q(s_t, a_i) - E_c(s_t, a_i)| = 1$.

3.1 Selection process

Two different schemes are assumed to isolate the accurate anomaly representative segments. One is to directly score each segment and then consider the most representative one that is judged by the highest anomaly value from each video. The other is to remove the worst segment that is judged by the lowest anomaly value gradually, and the remaining segments are the most representative ones. still, due to the shortage of data (annotated videos) for the learning process, its preferable to maximize the iterations within one video, to increase the learning results even with using less resources. Additionally, it is not easy to select the right segments from the first run. It is obvious that finding the worst segment in a video is less complex and more rewarding than directly finding the segments of interest. Thus, we

adopt the second plan, where an agent performs an action a_i by removing a segment x_i at state s_t. Therefore the action space is limited, and not the same at each step t.

The state s_t is represented by the remaining segments after t moves, the action a_i is represented by excluding a segment x_i at move t. Excluding a segment may lead to two states: s_{t+1} and termination, where termination means that we already found a set of segments that contain an anomaly.

The estimation feedback from environment r_i for (s_t, a_i) is not only provided by video level annotations, but also by video segment level annotations provided by a nominated expert E_c in anomaly recognition. The expert teaches our network as its recognition performance indicates the qualities of input segments x_i. To force the RL agent to learn the environment dynamics by itself. The environment does not provide any other feedback to the agent apart from the state and the reward.

3.2 Deep dueling-based anomaly detection

The system is mainly based on the dueling structure, which relies on two different sub-networks. They share the same features extractor layer [17]. The inputs is a video that is subsequently fragmented to 32 segments x_i, and considered as the initial state s_0 at t_0.

$$V = S_0 = \{x_i \mid 1 \leq i \leq 32\} \tag{2}$$

To transform the raw video segment data into an understandable format to the artificial agent, for each video segment x_i, we extract the visual features using the C3D Feature Extractor [17], then we obtain the corresponding final state format as follows:

$$V = S_0 = \{f_i^{x_i} \mid 1 \leq i \leq 32\} \tag{3}$$

Technically, each state s_t is a set of visual features representation $f_i^{x_i}$, which encapsulates all the data about the 3D convolutional features of the video segments ${X_i}^{Video}$.

To estimate the anomaly value of states s_t/video, all the extracted segment features $f_i^{x_i}$ are combined together, only to be used by the advantage stream as an input, which gives a single output value $V(s_t)$, the latter is given by equation 4. The stream's output indicates the probability of a video containing an anomaly.

$$V(s_t, \theta, \beta) = V(\{f_i^{x_i} \mid 1 \leq i \leq N\}, \theta, \beta) \tag{4}$$

The action-dependent advantage function $A(s, a)$ computes the value of the advantage of selecting a particular action(or segment) over the base value of being in the current state (or video).

We estimate the advantage stream by using the C3D features of the N remaining segments separately as input where $N = 32$ at step t_0/state s_0 presented by the following formula:

$$A(s_t, a_t, \theta, \alpha) = A(\{f_i^{x_i} \mid 1 \leq i \leq N\}, f_i^{x_i}, \theta, \alpha) \tag{5}$$

After the state value $V(s_t)$ and the video segments' value $A(s_t, a_i)$ are calculated, the output values of these two streams will be combined by an aggregation layer to evaluate each video segment x_i, in accordance with the following equation:

$$Q(s_t, a_i, \theta, \alpha, \beta) = V(s_t, \theta, \beta) + A(s_t, a_t, \theta, \alpha) - |\tfrac{1}{A}| \sum_{a_{t+1}} A(s_t, a_t, \theta, \alpha) \tag{6}$$

$Q(s_t, a_i)$ corresponds to the conditional probability of executing action a_i, which represents the deletion of a segment x_i from a state s_{t+1} at step t, the criteria of deletion is defined as a minimum Q value among values.

$$a_i \mid s_t = \begin{cases} \text{not}\operatorname*{argmin}_a(s_t, a) & x_i \text{ is a part of } s_{t+1} \\ \operatorname*{argmin}_a Q(s_t, a) & x_i \text{ is not a part of } s_{t+1} \end{cases} \tag{7}$$

The features f_i of the selected segments x_i will in turn be extracted at state s_t to return the next state s_{t+1} as following:

$$S_{t+1} = \begin{cases} \sum_{i=1}^{N} f_i^{x_i} - f_b^{x_b} & \text{if } b = \operatorname*{argmin}_a Q(s_t, a) \\ Terminal & \text{if } N = 5 \end{cases} \tag{8}$$

Then, the system judges the agent's decision by reward r_t provided by two different functions. For the videos annotated as normal, and abnormal defined by equations 9 and 10 respectively. The reward functions are as follows:

$$r_n^V(s_t, a_i, s_{t+1}) = \begin{cases} +1 & \text{IF } Q(s_t, a_i) = 0 \\ -Q(s_t, a) & \text{ELSE} \end{cases} \tag{9}$$

$$r_{ab}^V(s_t, a_i, s_{t+1}) = \begin{cases} +1 & \text{IF } Q(s_t, a_i) = E_c(s_t, a_i) \\ +|1 - ((s_t, a_i) - E_c(s_t, a_i))| & \text{IF } Q, E_c < T_h \vee Q, E_c > T_h \\ -|1 - ((s_t, a_i) - E_c(s_t, a_i))| & \text{IF } Q|E_c < T_h \vee Q|E_c > T_h \end{cases} \tag{10}$$

T_h is a predefined value $2 < T_h < 5$ representing the threshold that is used to signal an anomaly. The goal is reached once the anomaly is located in abnormal videos, or all segments of a normal video are well-judged $Q = 0$.

3.3 Prioritized experience replay

It is an improvement [18] to the Experience Replay mechanism used in the DQN algorithm that outperformed humans in Atari games [15]. The basic Experience Replay samples the batch uniformly (selecting the experiences randomly for training) these relevant experiences that occur rarely have practically no chance of being selected. As the name suggests, in Prioritized Experience Replay, a buffer is created to store the transition tuples by changing the sampling distribution based on a criterion to define the priority of each tuple of experience. The replay buffer is a cache D of

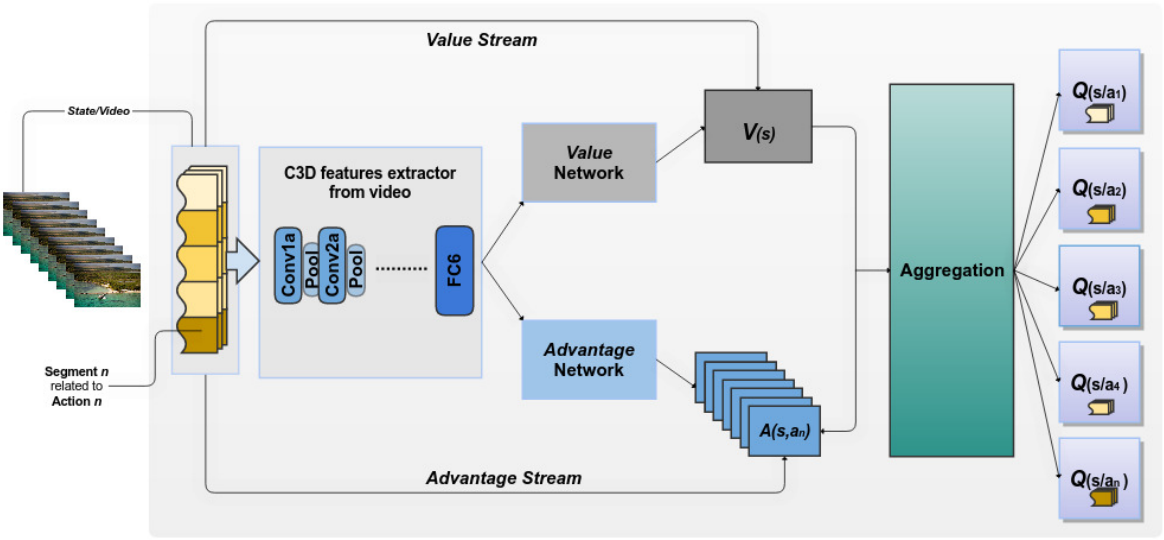


Figure 3: Deep dueling system for training a video surveillance agent.

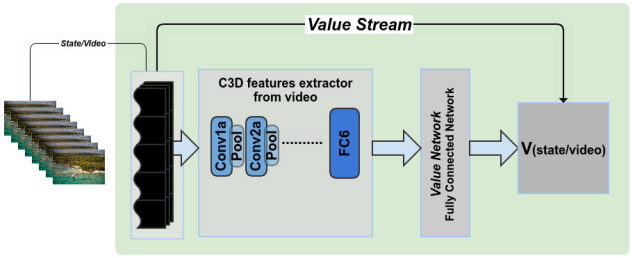


Figure 4: The value stream.

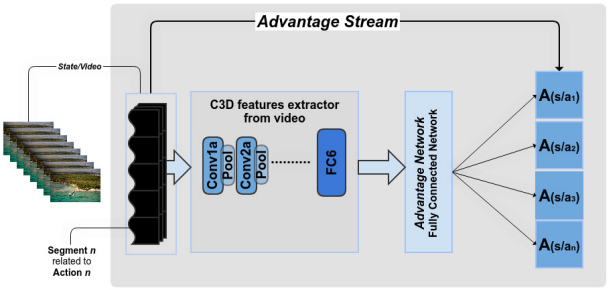


Figure 5: The advantage stream.

finite size to feed the neural network. Each (s_t, a_t, r_t, s_{t+1}) transition relative to the tempral-difference (TD) error. The highest priority is given to samples that produced a larger TD error, plus some constant to avoid zero probability for an experience being chosen.

4 Experimental results

Unlike the dataset used in [12], which is based on GPS trajectory, we choose to train our system on a dataset including multiple anomaly classes that are similar to real-world anomalies, in order to get as close as possible to the context of surveillance videos. So, we perform experiments on a large-scale dataset named UCF-Anomaly-Detection-Dataset[11] to evaluate the performance of our DRL anomaly event detector agent. The dataset is composed of long untrimmed surveillance videos which cover 13 real-world anomalies, including Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. The UCF-Anomaly-Detection-Dataset collected a 950 unedited real-world surveillance videos with clear anomalies as well as a 950 normal videos. The UCF dataset provides only video-level annotations, however, to train the system, we need segment-video level labels. For that purpose, we used an external expert system for video segment evaluation. In remainder of this section, we will describe the methods and the setup used for configuring and evaluating the learning video surveillance agent, and expose the details of the experiment results. Additionally, we compared our approach with state-of-the-art video anomaly detection.

4.1 Hyper-parameters

In our experiments, the preprocessing of the video data is made through the extraction of the visual features from the fully connected (FC) layer FC6 of the C3D network provided by authors of [17]. Before computing features, we resize each video frame to 240x320 pixels and fix the frame rate to 30 fps. We compute C3D features for every 16-frame video clip followed by l2 normalization. The agent's network is implemented with a fully connected Feedfor-

ward neural network setup by the network configuration described in Table 2. The network includes 3 layers. The ReLU function [19] is used for the two first layers, the output layer takes the sigmoid function for activation to build the output.

Layer N	1^{st} layer	2^{nd} layer	Output layer
Type	Dense	Dense	Dense
Unit size	512	256	64
Activation	ReLU	ReLU	Sigmoid
Weight Regularizer	decay $l2(0.001)$	decay $l2(0.001)$	decay $l2(0.001)$

Table 2: The network configuration.

After splitting every video into 32 non-overlapping segments, the agent starts learning by playing one episode per video. During an episode, the agent is allowed to play a number of steps to find the segments that cover the anomaly to be reported. The number value of steps depends on the annotation of the processed video, if it is annotated as normal, the agent has 32 steps t to remove all segments with a low abnormality score, otherwise, the segments are removed as there is no anomaly. In the case of an episode of a video annotated as abnormal, the number 5 has been set as a minimum of segments per state s_t, so the agent has 27 chances to remove the segments with the weakest anomaly score, in other words, it keeps the segment with high predicted anomaly score. We built our agent on 1600 episodes. Training was performed on a 4GB NVIDIA GeForce RTX 2070 SUPER GPU. The methods are implemented using Python with the help of Keras. The training process works under Adagrad optimizer algorithm [20] with MSE loss function and a learning rate with a value of 0.01, it's remaining parameters are set as default.

4.2 Results and analysis

Firstly, we study the sample results shown in Figures 6 and 7. Figure 6 represent a successful abnormality detection on videos containing anomalous events. The localization of anomalies is highlighted by the red frames, corresponding to the highest predicted anomaly score, and green frames highlighted segments corresponding to anomaly score approaching zero. The figure 7 represents a successful abnormality detection in videos with normal events, all frames in the video are shown in green. Meaning that all segments in the video have a low anomaly score.

The false alarms are considered as the weak point of an artificial video surveillance system. Based on the observations, we have made several attempts to reduce the false anomaly detection in the system. During the first evaluation, we noticed a large number of false negative cases, then we deduced that this was due to the default value of the number of minimum final segments (equal to 5). So, we rebuilt our model based on a new criterion, which is the final minimum number of segments will be decide by the expert E_c. In others words, for a given video that contains an anomaly, the stop criterion of the episode is the number of segments whose anomaly value judged by the expert is higher than a given threshold. Across many evaluations, the threshold value with the best results are 3.2.

We also observed a high score of false alarms in case of sudden people assembly as it happened with our principal chosen expert E_c, to reduce this phenomenon, we defined this case as a very important experience through prioritized experience replay mechanism. We managed to reduce the error score by up to 60.1%. However, the system failed in many cases of very crowded scenes.

The goal was to completely automate the video surveillance system. However, there were some false alarms, so we decided to set criteria to trigger the alarm (such as calling the police, or locking all the doors, etc). Therefore, the predicted anomaly score should be greater than a threshold to trigger the alarm automatically. Otherwise, we propose to send the video segment to a human assistant to take the final decision.

4.3 Comparison with SOTA methods

Table 3 summarizes the comparison of the proposed approach with the existing state-of-the-art methods using two different datasets. For the case of UCSD-dataset, our approach demonstrates inferior performance compared to the methods including Qiao et al [7], Khaleghi et al[8], Shean Chong et al[9], Hasan et al[10]. On the other side,

Approach	References	UCSD Dataset	UCF Dataset
Machine Learning	[21]	63.8%	54.3%
Deep Learning	[7]	98.33%	-
	[8]	88.1%	-
	[9]	89.9%	-
	[10]	90.0%	65.5%
	[11]	-	75.41%
Proposed system	#	87.44%	83.12%

Table 3: AUC comparison of the proposed system with SOTA baseline models on both UCSD dataset and UCF-Anomaly-Detection-Dataset using machine learning and deep learning methods.

the proposed system produces superior performance compared to the machine learning-based methods as Lu et al[21].

As far as we know, deep learning is a dominant source nowadays due to achieving high performance in many fields. the table shows that deep learning -based methods achieved better results than our approach. The obtained performance is due to the volume of data which is considered insufficient as it includes only 50 video samples for

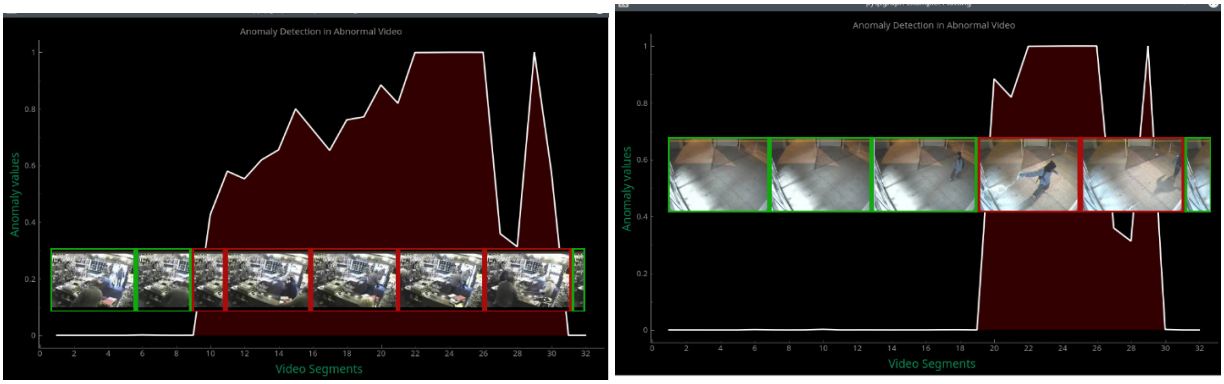


Figure 6: Examples of anomaly detection in Abnormal Videos.

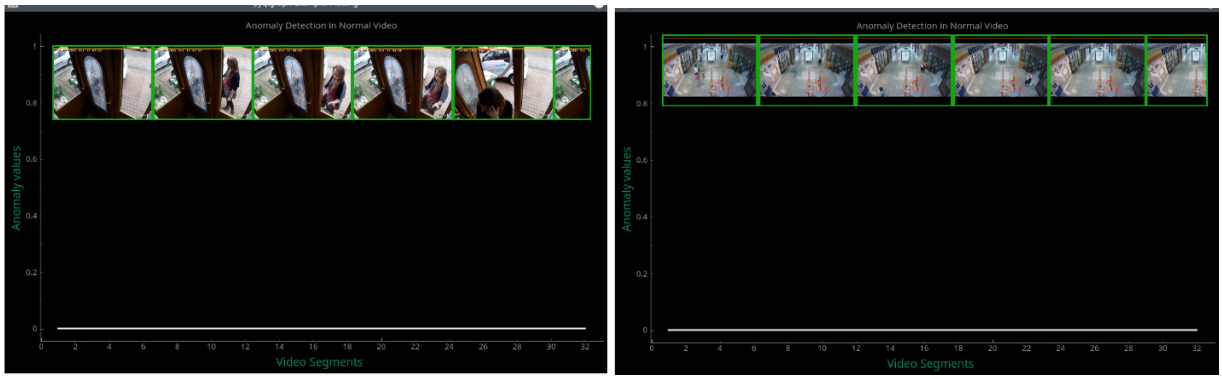


Figure 7: Example of anomaly detection in Normal Videos.

training, we assume that it surely affects the agent's learning process.

For UCF-Anomaly-Detection-Dataset that is considerably larger than the previous one, our system outperforms the deep learning and machine learning techniques. We conclude that the proposed system requires a big amount of training data for optimal performance.

[11] provides a framework to detect suspicious events in video surveillance by combining the two techniques of multiple instance learning with deep learning, resulting in an anomaly scores for each video.

Expert system E_c	Basic results	Proposed model results
Deep MIL[11]	75.41 %	83.12 %
Dictionary[21]	54.3 %	65.2 %
Deep auto-encoder[10]	65.5 %	71.09 %

Table 4: AUC comparison of multiples methods used as Expert system.

Our agent learned faster since in [11], the system started to predict the right anomalous score after 3000 iterations, but in our system, it was just after 1450 iterations. it is due to the fact that we have strengthened the learning phase with segment-level labels and the increasing of the exploration time. We were also able to surpass the compared method [11] in the reduction of false alarms as it generates a score of 1.9 for normal videos versus a score of 1.02 generated by our system. Additionally, by comparing our anomaly detection approach to two anomaly detection models: dictionary based approach [21] and deep auto-encoder based approach [10], the settings used for the comparison of both models are exactly the same ones set by [11].

Table 4 shows the comparison results for [11, 21, 10] frameworks, while simultaneously using the said approaches as the anomaly detection expert E_c, as well as a provider of the video-segments level annotations.

5 Conclusion

In this paper, an automatic video surveillance system including an anomaly detection based on deep reinforcement learning technique is proposed. In order to accelerate the agent's learning process and achieve a higher accuracy, this approach is relying not only on the annotations of the videos-level, but also on a video segment-level score provided by an expert system.

The system is trained on a variety of real-world anomalies to make it as efficient as possible in real life situations. The described method has achieved a very competitive performance that has surpassed some expert performances. Based on those results, we concluded that segment level annotations would greatly increase the system's performance if the annotations were done by a humans.

We employed many techniques of reinforcement learning such as prioritized replay and dueling architecture, though, there are still more recent improvements such as Rainbow model or NROWAN-DQN for network noise reduction.

References

[1] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans*, Pattern Analysis and Machine Intelligence, 2013. https://doi.org/10.1109/tpami.2012.59.

[2] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *NIPS*, 2014.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. *CVPR*, 2014. https://doi.org/10.1109/cvpr.2014.223.

[4] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A ¨local svm approach. in IEEE ICPR, 2004. https://doi.org/10.1109/icpr.2004.1334462.

[5] M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *ICCV*, pp. 1593–1600, 2009. https://doi.org/10.1109/iccv.2009.5459361.

[6] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval Image Processing. *IEEE Transactions*, vol. 16, no. 4, pp. 1168–1181, 2007. https://doi.org/10.1109/tip.2006.891352.

[7] Meina Qiao, Tian Wang, Jiakun Li, Ce Li, Zhiwei Lin, and Hichem Snoussi. Abnormal event detection based on deep autoencoder fusing optical flow. *Control Conference (CCC) 36th*, IEEE, Chinese, pages 11098-11103, 2017. https://doi.org/10.23919/chicc.2017.8029129.

[8] Ali Khaleghi and Mohammad Shahram Moin. Improved anomaly detection in surveillance videos based on a deep learning method. *2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN)*, IEEE, pages 73–81, 2018. https://doi.org/10.1109/rios.2018.8406634.

[9] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. *International Symposium on Neural Networks*, Springer, pages 189–196, 2017. https://doi.org/10.1109/ascc.2015.7244871.

[10] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. *CVPR*, 2016. https://doi.org/10.1109/cvpr.2016.86.

[11] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *Center for Research in Computer Vision (CRCV)*, 2018. https://doi.org/10.1109/cvpr.2018.00678.

[12] Min-hwan Oh, Garud Iyengar. Sequential Anomaly Detection using Inverse Reinforcement Learning. *arXiv:2004.10398v1 [cs.LG]*, 2020.

[13] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot and Nando de Freitas. Dueling Network Architectures for Deep Reinforcement Learning, *arXiv:1511.06581v3[cs.LG]*, 2016.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning, *Nature*, 518(7540):529–533, 2015. https://doi.org/10.1038/nature14236.

[15] Mnih, V. Kavukcuoglu, K. Silver, D. Graves, A. Antonoglou, I. Wierstra, D. Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[16] X. Lan, H. Wang, S. Gong, and X Zhu. Deep reinforcement learning attention selection for person re-identification. *BMVC*, 2017. https://doi.org/10.5244/c.31.121.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *ICCV*, 2015. https://doi.org/10.1109/iccv.2015.510.

[18] Tom Schaul, John Quan, Ioannis Antonoglou and David Silver. Prioritizied Experience Replay. *arXiv:1511.05952v4 [cs.LG]*, 2016.

[19] Xavier Glorot, Antoine Bordes, Yoshua Bengio.Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. PMLR 15:315-323, 2011.

[20] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res*, 2011.

[21] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. *ICCV*, 2013. https://doi.org/10.1109/iccv.2013.338.

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please register as an author and submit a manuscript at: http://www.informatica.si. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

SUBSCRIPTION

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twentyeight years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica web edition is free of charge and accessible at http://www.informatica.si.
Informatica print edition is free of charge for major scientific, educational and governmental institutions. Others should subscribe.

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2022 (Volume 46) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut, Peter and Jaša Žnidar; borut.znidar@gmail.com.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Slovene Society for Pattern Recognition (Vitomir Štruc)
Slovenian Artificial Intelligence Society (Sašo Džeroski)
Cognitive Science Society (Olga Markič)
Slovenian Society of Mathematicians, Physicists and Astronomers (Dragan Mihailović)
Automatic Control Society of Slovenia (Giovanni Godena)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Mark Pleško)
ACM Slovenia (Nikolaj Zimic)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 46 Number 2 June 2022 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics