Volume 46 Number 4 December 2022 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

Electronic and mobile health and applications

Guest Editors:

Sergio Crovella, Erik Dovgan, Flavio Rizzolo, Ivana Truccolo

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.

Executive Editor – Editor in Chief
Matjaž Gams
Jamova 39, 1000 Ljubljana, Slovenia
Phone: +386 1 4773 900, Fax: +386 1 251 93 85
matjaz.gams@ijs.si
http://dis.ijs.si/mezi/matjaz.html

Editor Emeritus
Anton P. Železnikar
Volaričeva 8, Ljubljana, Slovenia s51em@lea.hamradio.si
http://lea.hamradio.si/~s51em/

Executive Associate Editor - Deputy Managing Editor Mitja Luštrek, Jožef Stefan Institute
mitja.lustrek@ijs.si

Executive Associate Editor - Technical Editor
Drago Torkar, Jožef Stefan Institute Jamova 39, 1000 Ljubljana, Slovenia
Phone: +386 1 4773 900, Fax: +386 1 251 93 85
drago.torkar@ijs.si

Executive Associate Editor - Deputy Technical Editor Tine Kolenik, Jožef Stefan Institute
journal.informatica.si@gmail.com

Editorial Board
Juan Carlos Augusto (Argentina)
Vladimir Batagelj (Slovenia)
Francesco Bergadano (Italy) Marco Botta (Italy)
Pavel Brazdil (Portugal)
Andrej Brodnik (Slovenia)
Ivan Bruha (Canada) Wray Buntine (Finland)
Zhihua Cui (China)
Aleksander Denisiuk (Poland)
Hubert L. Dreyfus (USA) Jozo Dujmovic´ (USA)
Johann Eder (Austria) George Eleftherakis (Greece)
Ling Feng (China)
Vladimir A. Fomichov (Russia)
Maria Ganzha (Poland)
Sumit Goyal (India) Marjan Gušev (Macedonia)
N. Jaisankar (India)
Dariusz Jacek Jakóbczak (Poland)
Dimitris Kanellopoulos (Greece)
Samee Ullah Khan (USA)
Hiroaki Kitano (Japan)
Igor Kononenko (Slovenia)
Miroslav Kubat (USA) Ante Lauc (Croatia)
Jadran Lenarčič (Slovenia)
Shiguo Lian (China)
Suzana Loskovska (Macedonia)
Ramon L. de Mantaras (Spain)
Natividad Martínez Madrid (Germany)
Sando Martinčić-Ipišić (Croatia)
Angelo Montanari (Italy)
Pavol Návrat (Slovakia)
Jerzy R. Nawrocki (Poland)
Nadia Nedjah (Brasil)
Franc Novak (Slovenia)
Marcin Paprzycki (USA/Poland)
Wiesław Pawłowski (Poland)
Ivana Podnar Žarko (Croatia)
Karl H. Pribram (USA)
Luc De Raedt (Belgium)
Shahram Rahimi (USA)
Dejan Rakovic´ (Serbia)
Jean Ramaekers (Belgium)
Wilhelm Rossak (Germany)
Ivan Rozman (Slovenia)
Sugata Sanyal (India)
Walter Schempp (Germany)
Johannes Schwinn (Germany)
Zhongzhi Shi (China) Oliviero Stock (Italy)
Robert Trappl (Austria)
Terry Winograd (USA)
Stefan Wrobel (Germany)
Konrad Wrona (France)
Xindong Wu (USA)
Yudong Zhang (China)
Rushan Ziatdinov (Russia & Turkey)

Honorary Editors
Hubert L. Dreyfus (United States)

https://doi.org/10.31449/inf.v46i4.4587 Informatica **46** (2022) 441–442 **441**

AI and Games at IJCAI - ECAI 2022

An interview with Prof. Jonathan Schaeffer
Editorial by Matjaž Gams

At IJCAI we had an opportunity to discuss with Jonathan Schaeffer, a Canadian artificial intelligence researcher, professor in computing science, and former dean of science at the University of Alberta, Edmonton, Alberta. He is best known as the primary author of the World Man-Machine Checkers Champion Chinook, solving Checkers in 2007. In August 2019, Jonathan Schaeffer superseded David Levy as president of the ICGA, International Computer Games Association.

Figure 1: Prof. Jonathan Schaeffer, president of ICGA.

Question: Congratulations on all achievements – which ones would you highlight?

Reply: The wistful highlight of my career was my program Chinook winning the World Checkers Championship in 1994 – the first time a computer won a human world championship in any game. I call it wistful because, sadly, our success happened at the tail end of the career of the remarkable human champion, Dr. Marion Tinsley. Soon after Chinook won the championship, Tinsley passed away from cancer.

Another highlight was solving checkers. I started computations to solve checkers running in 1989 and in 2007 they were completed. It took 18 years and hundreds of computers to announce that perfect play leads to a draw.

Question: Any comment on the IJCAI computer championship?

Reply: Man versus machine competitions in chess started in 1970. In 1974 the first World Computer Chess Championship was held. Forty-eight years later, we are still holding this competition, including here at the 2022 IJCAI conference. In the early 2000s, chess programs went superhuman. The last time a human grandmaster defeated a strong chess program was in 2005. Whereas the human World Chess Champion Magnus Carlsen has a roughly 2,850 ELO rating, the top chess computers have ratings over 3,500! The programs still get stronger every year!

Computer chess performance benchmarking is the longest-running experiment in computing science history.

Question: Are chess programs approaching their limit – playing optimally? ELO ratings stalling might indicate so.

Reply: I don't think so, although we are seeing diminishing returns for the effort expended. The limit, of course, will come when chess is "solved". However, given the roughly 10^{45} possible positions in the game, solving chess is not going to happen for a very long time (and without major hardware and software technology breakthroughs).

Question: In which games are computers optimal/dominant/comparable/worse than humans?

Reply: If we limit the discussion to the classic board and card games: Solved games include 8x8 checkers, 2-person limit Texas Hold'em poker, Awari, and so on. Superhuman games include chess, Go, shogi, backgammon, and so on. Worse include bridge.

Question: Do humans play better due to computer chess?

Reply: Yes. Computers can help humans train (available to play 24 hours a day), study the openings (checking analysis, and uncovering new lines of play), and reveal new ideas.

Question: AI and games used to be one of the main topics of research. What changed?

Reply: Building superhuman game-playing programs – especially chess – was one of the early grand-challenge problems of AI research. But 60 years of creating innovative algorithms and using ever-faster hardware has meant that this AI goal has been achieved. Time to move on to more challenging problems, such as video games.

Question: Elon Musk says that AI progress is so much faster than that of humans that it is only a matter of time when AI will supersede humans. Agreed?

Reply: Yes and no. AI has already exceeded humans in some domains, with many more to come. But there are areas where right now it is hard to see AI exceeding human abilities (do you think an AI could write like Shakespeare or paint like da Vinci?). One thing I have learned the hard way about AI is never to predict the future. There is so much innovative research happening today that tasks that seem hard now may be easy tomorrow. The game of Go was such an example. Within a year the problem of beating the world champion went from seemingly impossible to mission accomplished (2016). Never count out human ingenuity!

Question: When will superintelligence appear?

Reply: The definition of super-intelligence is not clear. AI already has some abilities that exceed those of any human in some areas.

Question: Many AI researchers imply that the progress of human civilization is in recent decades mainly due to the progress of AI. Is this an overstatement?

Reply: Yes! AI tools are already improving the quality of human life (and perhaps some that are not). AI's role will likely become much more important (and sooner, rather than later). But today we already have other technology tools that are aiding in the progress of human civilization. Smartphones. Discoveries stemming from an understanding of the human genome (including vaccines). Electric cars…

Question: AI and superintelligence should help humans as cars or robots for example. When and why appeared this freak horror about robots or AI killing humans?

Reply: I have given many public AI talks. Inevitably, someone asks "When is AI going to destroy civilization." This question angers me. Every AI scientist that I know is working on developing AI technology for the benefit of humankind. There is an urgent need to educate the public about what AI can do – and especially what AI cannot do (or is unlikely to do).

Figure 2: Computer world championship at IJCAI 2022 in Vienna, Austria

https://doi.org/10.31449/inf.v46i4.4161 Informatica **46** (2022) 443–448 **443**

Recommending Relevant Services in Electronic and Mobile Health Platforms

Gjorgji Noveski[1] and Jakob Valič[1]
E-mail: gjorgji.noveski@ijs.si, jakob.valic@ijs.si
[1]Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia

Keywords: Electronic and mobile health, recommendation systems, embeddings

Received: May 6, 2022

Electronic and mobile health (EMH) is becoming an integrated part of healthcare as we move in the future. The opportunity in bringing closer healthcare services with the advent of the internet is growing larger. This is why it is important to adequately provide those services to the people that need them and to also further improve them. Regarding electronic and mobile healthcare systems, it is fairly easy for users to get lost while searching for some information due to the vast amount of data that is present for different illnesses, healthcare institutions and healthcare services. In this paper we present a platform that provides various healthcare services to people, namely the Insieme platform (ISE-EMH). Knowing the difficulty of finding relevant information on platforms and that user preferences vary to a great extent, we additionally give an overview of an implementation of the recommendation system that is part of the Insieme platform which helps users pick services that might be relevant to them.

Povzetek: Razvita ja pratforma ISE-EMH za Italijo in Slovenijo kot jedrnata verzija "dr. Googla".

1 Introduction

With the increasing popularity of digital interventions and digital solutions, electronic and mobile health applications are becoming more popular throughout many sectors of the public healthcare. The idea started since the beginning of the era of smartphones, where the convenience of accessing the internet from the comfort of our homes, moved to conveniently accessing the internet from the palm of our hands. This shift provided fertile ground for a vast amount of new solutions and applications to emerge. Naturally, the healthcare domain was also affected by this and with the collaboration of healthcare professional and people from the domain of information and communication technologies, solutions are built that benefit the people.

In the digital era, more things are becoming available online and with this, general quality of life of people is increasing. Many companies, institutions and various stakeholders shift towards online solutions regarding their data availability. This is beneficial to both the companies and the people using their services for various reasons. Firstly, it allows remote access to employees regardless of location, giving them the option to continue working even when they are not on-site. Furthermore, it can allow commonly used services to be available to the general public. Such examples are online booking for doctor's appointment or online overview of waiting queues. Advances and increase in usage of wearable devices give opportunities for tracking many aspects of a patient's vital signs, heart rate, breathing, blood pressure, etc. This information can be integrated into a system that builds up an electronic health record from the user and provides health checks, monitoring and suggestions. Clearly we can see that there are many opportunities to be explored related to using digital technologies and data on person's health.

During design and implementation of healthcare platforms, ease-of-use and user experience need to be considered. This is because the platforms are going to be used by a variety of individuals, the healthy and the ill. Because of this, the process in which they obtain the required information from the healthcare platform must be as easy as possible. Many different digital solutions have been developed which provide healthcare assistance like embedded [1], and decision support systems [3]. In recent time advances in artificial intelligence pave way to more complex solutions that bring a bigger array of services and benefits in the field of healthcare. These benefits are from areas such as predictive analysis of diseases, patient care, patient monitoring, etc. [6].

Like previously mentioned, the time needed to reach desired information on a healthcare platform should be as low as possible. In order to achieve this, a certain system must be able to learn the patterns of user's interaction with the platform. This way, the system provides information to the user while also building a model representation for that user. This learned context gives the system an ability to tailor which information is shown to which user, thus providing more specialized experience. Such functionalities are typically implemented using recommendation systems.

In our paper, we present the methodology for recommending services on the Insieme platform, i.e., an Electronic and Mobile Health platform that we developed within the ISE-EMH project [4]. We give an overview of the platform as well as the recommendation system used to provide helpful service suggestions. Additionally, analysis is per-

formed on gathered user-service interaction data in order to get some insight on site usage.

The rest of the paper is organized as follows. Section 2 provides examples of related work in the field of Electronic and Mobile Health as well as similar applications. In Section 3 the implementation of the recommendation system is explained while Section 4 gives brief explanation on the EMH platform that the system is used on. Interaction analysis from user data is given in Section 5 and lastly a conclusion is presented in Section 6.

2 Related work

Taking a look at related systems, we can find examples of different platforms and applications that work towards bringing ease of use to the user. The actors in the healthcare system are the patients and the clinicians. Consequently, solutions that benefit either the patients or the clinicians benefit the whole healthcare system. We will present some solutions that are aimed at both clinicians and patients.

Showcasing the importance and benefits of using recommendation systems in the healthcare domain is meticulously presented in [9] where an overview of the existing research in the EMH domain is provided. The analysed scientific papers were also carefully selected and filtered by their criteria which included: selecting papers published no earlier than year 2000, papers referenced with 15 sources and more, containing a detailed discussion on recommendation techniques, etc. After discarding papers that did not meet these criteria, 98 papers remained to be further studied. Examples of the reviewed systems are systems that can recommend food, drugs, healthcare professionals and, etc. The authors listed three main aspects that need to be considered when designing a recommendation system: users, items, and usage context. In our case, users are the patients that will use the system and items are the entities which will be recommended to them. Usage context refers to the set of factors that might influence the selection of the items to be recommended to the user. For example, if the user's health record is integrated into the recommendation system then different users can get different items recommended to them just because of their dietary preference, allergies, etc.

De Croon et al. [2] reviewed a total of 73 published studies that had reported the implementation and evaluation of recommendation systems in the healthcare domain. The authors showed that the most prominent categories of recommended items are about lifestyle, nutrition, general health information, and specific health condition related information. In our work, the recommendation system recommends relevant services to the users, which are related to the specific health condition related information. In literature the most frequently used type of recommendation system is a hybrid one which is a combination of collaborative filtering and content-based filtering (see Section 3. for details).

Diaz Ochoa et al. [7] applied neural networks to create a recommendation system that achieves recall of 98 % and accuracy of 64 %. It takes into account patient data alongside with specific treatments encoded into treatment keys. This information is further processed by clustering in order to lower the dimensionlity of the problem and finally a deep learning model is trained. The advantages of the developed solution are faster training time due to its fewer number of parameters and transparency due to its use of multi-criteria decision operators. The Logic-Operator neural network they use simulates cognitive processes with fuzzy logic. The logical operators are or and and gates which represent the last layers of the network. Furthermore mean squared error is used as a loss function, ReLU as an activation function and the ADAM method as optimization. The recommendation system consists of two separate systems. One of them is trained on the entire patient dataset, while the other is trained only on patients which had positive outcome of their treatments. This provides two predictions and by comparing the two predictions a recommendation is made with low or high confidence. If the predicted treatments match then high confidence is assumed. Likewise if the predictions of the two systems do not match, a treatment is still recommended but with low confidence.

Punn et al. [8] developed a recommendation system in the healthcare domain, which is based on collaborative filtering and recommends remedies by taking as input the patient's symptoms. Since there is a limited amount of data linking remedies to various diseases that is suitable for creating recommendation systems, the authors also provided a dataset for this purpose. This dataset consists of around 1,100 diseases and close to 300 symptoms. Their recommendation system uses singular value decomposition and cosine similarity in order to assess which symptoms give more rise to a certain disease. The system was evaluated with symptoms that were related to one or more diseases and in both cases it recommended one or more remedies respectively.

3 Recommendation systems

Recommendation systems aim at providing reliable recommendations for the domains they are built for, e.g., songs, movies, products at an online shop, etc. With the term "items" we associate all the possible types of recommendations. Because of the ability of the recommendation systems to adapt suggestions based on specific users, several online services that include recommendation systems have greatly improved user experience.

The recommendation system takes as input a search query vector and computes which items are most similar to it. There are several options to compute the similarity between different vectors. Several recommendation systems use one or more of the following metrics:

1. Cosine
2. Dot product
3. Euclidean distance

These metrics calculates a score of how much the current choices of the user are similar to all the other items. The items which relate the highest are taken as recommendation to the users. Our system uses the cosine similarity metric.

There exist three types of recommendation systems: collaborative filtering, content-based filtering, and hybrid ones which are a combination of the first two. In our platform, we use the LightFM recommendation system [5], which is a hybrid one. It works as a hybrid matrix factorization model where it represents the users u and items i as embeddings (latent vectors). These are defined as linear combination of the features that describe each user or item. When these features are present in the model, it defines the recommendation by calculating the dot product of u and i representations, adjusted by the respective biases b_u and b_i. When no item or user features are available, the LightFM model falls back to pure collaborative filtering. These types of recommendation systems are further explained in the sections below.

3.1 Content-based filtering

This is the simplest type of recommendation system. In essence it uses the similarity between items to recommend items similar to what the user has chosen in the past. From the constructed user-item interaction matrix, item features are extracted and a similarity metric is calculated. This approach has several advantages and disadvantages, for instance it is scalable to a large number of users because it considers only one user at a time. In addition, it can provide recommendations to very niche items, by learning from the user's previous interactions with the whole system. Downside of such a system is that it needs domain knowledge to some extent for the creation of item features. These item features for example can represent the categories of a certain movie (horror, comedy, action, etc.). Further downside is that it can not expand the user's taste by suggesting items which are not closely similar to what the user has chosen in the past.

3.2 Collaborative filtering

To address some of the concerns and drawbacks of the content-based filtering method, collaborative filtering was developed. Collaborative filtering takes into account the similarities between users and items simultaneously in order to provide recommendations. The benefit of this is that while the system looks for similar items, it also looks for similar users to the target user, which means new items can be suggested to him/her. Consequently, the users are able to discover new interests.

The recommendation system can operate in two modes, depending on how it interprets the user-item interaction matrix. These two modes are:

1. Implicit feedback
2. Explicit feedback

In the implicit feedback mode, the model regards the absence of data in some fields as negative feedback. This means that for the items the user did not supply a rating, it considers them as if the user gave a negative rating. The idea behind this is that the user consciously made a decision about which items he/she will rate, so the system regards the items that aren't rated, as items that the user does not like. On the other hand, an explicit feedback model is the opposite. A rating for an item has to be supplied in order for the system to make any judgements. The items which do not have a rating are considered as data that is unknown. Which one to use is typically dependant on a use-case basis.

3.3 Embeddings

An embedding represents a relatively low dimensional space which captures some semantic meaning from a higher dimensional input. When working with categorical data, as is in our case, having many features quickly brings the size of the input to a large number. This results in longer training times of the models and more data that needs to be processed. Embeddings can transform the data from a categorical one, to a real-valued representation in fewer dimensions. On each of these dimensions the system learns to represent some concept, for example the N-th dimension might represent the fact whether the item belongs to a class of horror movies or action movies.

In our recommendation system, embeddings are learned internally and used for predicting services. We also use them in order to acquire a list of most similar tags to be suggested to the user. To create meaningfull embeddings, LightFM gives the option to add additional features that describe the items and features that describe the users. Consequently, better embeddings are learned by the system and used for generating better recommendations.

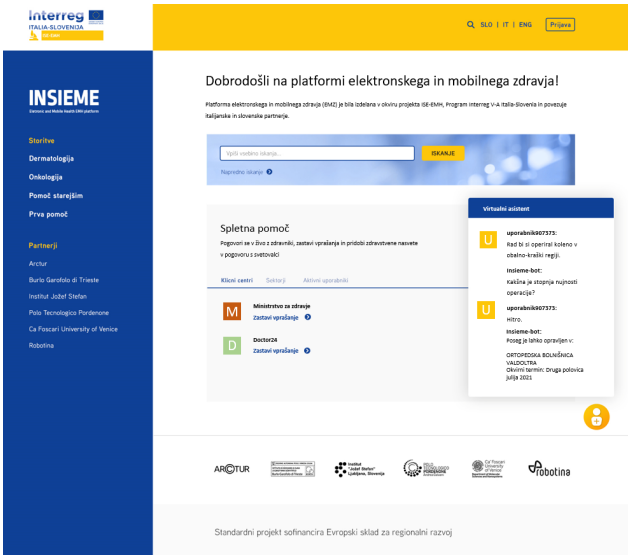


Figure 1: The Insieme platform.

4 Insieme platform

The ISE-EMH platform connects various partners, medical institutions, and patients from both Italy and Slovenia. The medical institutions and partners provide information about their services through the platform so the users can have one location where that information is easily accessible. The platform also facilitates communication between a specialist and a patient. Furthermore, a chatbot is developed that integrates various services and question answering. For example, it can list the waiting queues for different medical institutions and provide helpful information about diseases. When getting information about various diseases, the user is also provided with information about medical institutions that deal with them alongside the symptoms associated with the disease. In Figure 1 we can see the graphical user interface of the platform.

5 Interaction analysis

An integrated part of the platform is logging of the user interactions on every click. When a new user visits the site and he/she click on a service, that interaction is saved in a database. The system keeps track of which users interact with which services by means of internet cookies. We consider the time between the user visiting the platform and leaving it as a "session". The information from the sessions is crucial if we want to train our recommendation system. Each user has his/her own unique session identification number for as long as he/she uses the platform. It is possible for the same person to use the platform multiple times but on different devices. This way the system will assign a different identification number to the different sessions. During design we presume that each identification number belongs to a different person even though it may not always be the case. This choice was taken since we assume that the number of occurrences of this event will be insignificant. The goal of this analysis is to acquire and present empirical data of the platform usage to acquire a better understanding of user patterns in order to further refine the recommendation system.

The interaction analysis was performed as follows. We analysed the number of times services were visited. As seen in Figure 2, majority of services have been visited between one and thirty times so far. A popularity trend can be also observed, e.g., the top three services which were the most chosen are:

1. Psoriasis
2. Ambient assisted living
3. Breast cancer

Figure 3 shows how many services were visited in each session. This figure shows that the majority of users (more than half at around 500) have only clicked on a single service. Consequently, we can assume that most of the users come to the platform to obtain specific information and know what they are searching for. It should be noted that sessions in which more than 50 services were visited are filtered out from this data. We consider them as outliers and rare occurrences. Nevertheless, a high number of services visited per session might indicate users which do not come to the platform with a specific goal in mind, but just casually educate themselves in resources the Insieme platform provides.

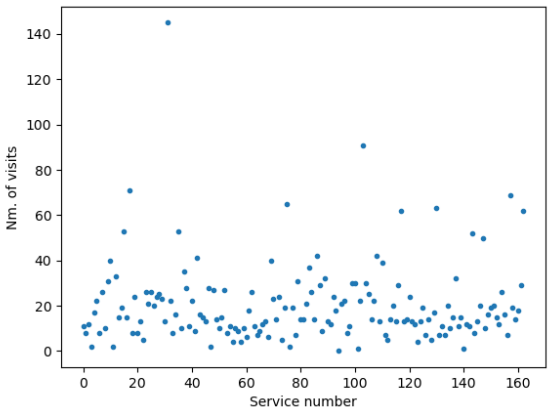


Figure 2: Number of times a certain service was clicked.

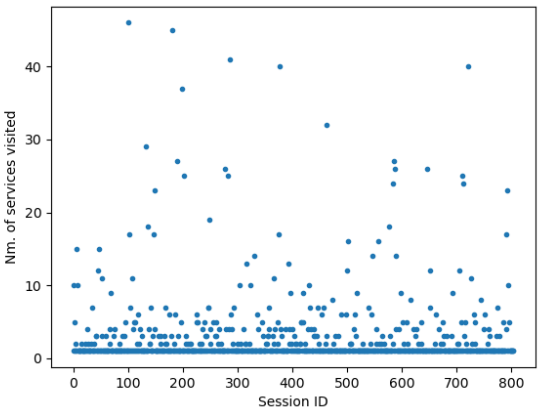


Figure 3: Number of services visited in each session.

The implementation of the recommendation system on the Insieme platform is regularly trained with new data. The time period between each training is one day. This ensures enough time for new data to be generated by users throughout the day whilst at night when site traffic is lower the system is trained and improved. The presented interaction data was collected over the course of several months.

Table 1 shows an example of interaction matrix between (subsets of) user sessions and services. The "+" sign indicates that a service was visited during a session.

Each of the services on the Insieme platform has a set of tags associated with it, for example, organ-liver, duration-

Table 1: The interaction matrix with subsets of user sessions and services.

Services / Sessions	Acne	Dermatology	Heart Diseases	Dementia	Epilepsy	Elderly support	Mental illness
4relftnjnvvzjogwtu3ap2pnx	+	+					
7dmar25i0cqm0a64fajznt4fm			+		+	+	
3sgxsbskwqszpbg2dom2y6k7f				+			
0n5c4ms9jk27khzcbnfvou7qt				+			+
clainomcth543lm7ss6lpisv0	+	+					
8b9k48ced5pbij88nk1ommvkx						+	+
0b8w50rwyel1bxutiz4m2ujg0	+				+		
4di23cq5t5637es2eoa43at4n			+	+	+		

chronic, disease course-progressive, etc. By providing recommendations to users based on these tags, the users are able to easily search for and identify the needed services. In addition, the usage of the tags enables us to learn better embeddings and provide improved suggestions. An additional procedure for suggestion making is to obtain the most similar services embedding-wise. The system can provide recommendations by searching the proximity of the queried service in its lower dimensional embedding space and returning services which are in that proximity. Since embeddings can capture semantics, the obtained services will be those that are closely related to the searched service.

We also trained a secondary model that provides recommendations for tags instead of services. This is a more general model and not as specific as the model used for recommending services. The main advantage of this model is that it requires less training data because we have a smaller number of tags compared to the number of services. Both the service suggestion model and the tag suggestion model can be used side-by-side or independently. Due to the platform requirements, only the service suggestion model is used.

We tested the speed of the Insieme recommendation system and measured the area under the receiver operating characteristic curve (AUC) on predictions obtained from a training and test set. In order to ensure satisfactory functioning of the website, the recommendation system had to make a prediction in less time than it takes for the services database to return a query. The results show that the system complies with this requirement and that its speed does not drastically affect loading times. The interaction data that was collected in the course of several months, was split into training (70 %) and test (30 %) data set. The model that uses services only (without tags) as the input, achieved an AUC score of 0.96 on the training set and an AUC score of 0.7 on the testing set. We expect this metric to improve in due time with further maturing of the website.

6 Conclusion

This paper presented the implementation of the recommendation system in the Insieme electronic and mobile health platform. The experimental results showed the ability of this system to recommend relevant services within given time constraints. Since the recommendation model was trained on data obtained from a time period of several months, constant usage of the platform and further user interactions are expected to increase the relevance of the suggestions. Furthermore, an additional model was trained for suggestion of service tags. In both models it can be observed that semantic meaning is captured through embeddings.

Acknowledgement

The paper was supported by the ISE-EMH project funded by the program Interreg V-A Italy-Slovenia 2014-2020.

References

[1] A. M. Alharbi, N. T. Alharbi, H. M. Alharbi, and D. M. Ibrahim (2019). Patient Assistance System: A Proposed Structure. *10th International Conference on Information and Communication Systems (ICICS)*, pp. 230–233, https://doi.org/10.1109/iacs.2019.8809136.

[2] R. D. Croon, L. V. Houdt, N. N. Htun, G. Štiglic, V. V. Abeele, and K. Verbert (2021). Health recommender systems: systematic review. *Journal of Medical Internet Research*, vol. 23, no. 6, article no. 18035, https://doi.org/10.2196/18035.

[3] A. Hommersom, P. J.F. Lucas, M. Velikova, G. Dal, J. Bastos, J. Rodriguez, M. Germs, and H. Schwietert (2013). MoSHCA-my mobile and smart health care assistant. *15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, pp. 188–192, https://doi.org/10.1109/healthcom.2013.6720664.

[4] ISE-EMH, Interreg Italia-Slovenia project, https://www.ita-slo.eu/en/ise-emh. Accessed July 13, 2022.

[5] M. Kula (2015). Metadata embeddings for user and item cold-start recommendations. *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015)*, vol. 1448, pp. 14–21, https://doi.org/10.1145/2792838.2798718.

[6] R. Manne and S. C Kantheti (2021). Application of artificial intelligence in healthcare: chances and challenges. *Current Journal of Applied Science and Technology*, vol. 40, no. 6, pp. 78–89, https://doi.org/10.9734/cjast/2021/v40i631320.

[7] J. G. D. Ochoa, O. Csiszár, and T. Schimper (2021). Medical recommender systems based on continuous-valued logic and multi-criteria decision operators, using interpretable neural networks. *BMC medical informatics and decision making*, vol. 21, no. 1, pp. 1–15, https://doi.org/10.1186/s12911-021-01553-3.

[8] Sudhanshu, N. S. Punn, S. K. Sonbhadra, and S. Agarwal (2021). Recommending best course of treatment based on similarities of prognostic markers. *International Conference on Neural Information Processing*, pp. 393–404, https://doi.org/10.1007/978-3-030-92270-2_34.

[9] T. N. T. Tran, A. Felfernig, C. Trattner, and A. Holzinger (2021). Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, vol. 57, no. 1, pp. 171–201, https://doi.org/10.1007/s10844-020-00633-6.

Ranking Effectiveness of Non-Pharmaceutical Interventions Against COVID-19: A Review

David Susič*[1], Janez Tomšič[1], and Matjaž Gams[1]
[1]Department of Intelligent Systems, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: david.susic@ijs.si, janez2001@gmail.com, matjaz.gams@ijs.si

Keywords: Non-pharmaceutical interventions, COVID-19, SARS-CoV-2

Received: May 13, 2022

In this review, we examine 34 studies based on experimental data that estimate and compare the effectiveness of 12 non-pharmaceutical government interventions against COVID-19 based on cases, deaths, and/or transmission rates to assess their overall effectiveness. The studies reviewed are based on daily country-level data and cover four to 200 countries and regions worldwide with varying time intervals, spanning the period between December 2019 and August 2021. We found that the overall most effective interventions are restrictions on gatherings, workplace closing, public information campaigns, and school closing, while the least effective are close public transport, contact tracing, and testing policy.

Povzetek: Predstavljen je pregled 34 objav, ki analizirajo uspešnost ukrepov proti kovidu.

1 Introduction

Looking back at the first months of 2020, it is clear that the pandemic COVID-19 caught the world unprepared. Initially, it was unclear how contagious the virus was, how quickly it would spread, how to protect against it, and how to prevent hospital overload. To combat the spread of the virus, governments began introducing various non-pharmaceutical interventions (NPIs). It quickly became clear that some NPIs had a stronger impact on containing the pandemic than others. As a result, researchers around the world have begun to study the effectiveness of NPIs in different geopolitical regions. Despite the vaccine being developed in the last half of 2020, the spread of COVID-19 and the number of infections are still a major burden to society. As of May 2022, there have been 6.25 million COVID-19-related deaths worldwide [1].

In this paper we extend our earlier work [2]. We review related work on the effectiveness of NPIs implemented in different countries and over different time periods, with the goal of assessing and ranking their overall effectiveness. There is some similar work in the literature [3, 4, 5], but in this work we only consider studies in which conclusive evidence of the effectiveness of at least two NPIs was found. In addition, we do not include simulation-based studies. Unlike the two reviews mentioned above, our review includes time intervals from the third and fourth waves and, to the best of our knowledge, is the most up-to-date review in this regard.

The rest of this paper is structured as follows. Section 2 presents methodology for selecting the papers and ranking the effectiveness of the NPIs. In section 3, we present and analyse the results. Section 4 describes the limitations of our study. We conclude the paper in section 5.

2 Methodology

The first step in our research was to establish the criteria for selecting the papers to be included and to create a unified ranking system that would allow us to compare the rankings of NPIs in related work.

2.1 Eligibility criteria

In this review, we considered 12 NPIs from the Oxford COVID-19 Government Response Tracker (OxCGRT) [6]: school closing (C1), workplace closing (C2), cancel public events (C3), restrictions on gatherings (C4), close public transport (C5), stay at home requirements (C6), restrictions on internal movement (C7), international travel controls (C8), public information campaigns (H1), testing policy (H2), contact tracing (H3), and facial coverings (H6). The letters C and H correspond to *containment and closure policies* and *health system policies*, respectively. The 12 selected NPIs were chosen because they have been implemented most frequently and therefore cover the majority of all measures implemented worldwide.

We searched for papers written in English and published up to May 2022. We searched Google Scholar for published studies and MedRxiv for preprints. For a study to be included in this review, it had to meet the following criteria:

- studies the effect on COVID-19 related deaths, cases or transmission rate,
- compares NPIs that map to at least two OxCGRT NPIs,
- is based on experimental data and not based on forecasts/simulatons, and
- was conducted on a geographical region level (one or more), meaning that studies that only focus on selected

groups of people (e.g. people from Universities only) [7, 8] were not included.

All papers included in this review are listed in Table 3 along with their respective study settings. In the cases where the study used NPI information from a database other than OxCGRT, the NPIs first had to be mapped from the other database to the OxCGRT, based on the descriptions of the interventions in both of the documentations. If multiple NPIs corresponded to one OxCGRT NPI, their scores were averaged. In contrast, if a single NPI corresponded to more than one OxCGRT NPI, its score was applied to all corresponding OxCGRT NPIs.

2.2 Ranking the effectiveness

To rank the effectiveness of the NPIs, we used a scale of one to four, with one and four representing the most and least effective NPIs studied, respectively. The effectiveness scores from each study were first ranked and then divided into four equally sized bins, with the most effective NPIs in bin one and the least effective NPIs in bin four. The bin number corresponds directly to the value on our effectiveness scale. Note that in some studies, some of the bins may be empty, resulting in this value not being assigned to an NPI.

In the Bendavid et al. study [9], the estimated impacts were reported separately for each country studied. In this case, the values were first averaged across countries and then ranked.

In the work of Askitas et al. [26], the NPIs were classified descriptively only. C1, C2, C3, and C4 were found to be the most effective NPIs and were given a value of one. The effect of C6 was judged to decrease over time and was therefore given a value of two. C8 was judged to be less effective and was given a value of three, while C5 and C7 were judged to be negligibly effective and were given a value of four.

Li et al. [10] calculated the estimated effects one, two, and three weeks after the implementation. In this case, the scores were averaged across all three cases.

In the work of Liu et al. [11], the effectiveness of NPIs was estimated in two scenarios, where NPIs are implemented at their maximum stringency or at any stringency. The NPIs were then described as either strong, moderate, or weak in both of the scenarios. The NPIs graded strong in at least any stringency scenario were assigned value one, NPIs graded strong in maximum stringency scenario only were assigned value two. All NPIs graded moderate were assigned value three, and all NPIs graded weak were assigned value four.

In the study by Wibbens et al. [12], the effectiveness of NPIs was assessed at different intensity levels. They were first rated separately at the highest intensity and at an intermediate intensity. Then, their overall ranking was calculated as the average of the two.

The estimated effects of NPIs from all studies reviewed are summarised in Table 3. In studies in which effects were estimated but could not be ranked [13, 14, 15, 16], all NPIs were assigned a value 2. In studies in which fewer than four NPIs were considered [13, 14, 17, 18, 19, 20, 21, 22, 23, 24, 25], values were also assigned on the basis of descriptive ranking.

3 Results

Among the 34 studies selected in this review, there are 14 works that deal with cases incidence [13, 14, 16, 21, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34], 11 works that deal with reproduction number [10, 11, 18, 20, 22, 34, 35, 36, 37, 38, 39], seven works that deal with infection growth rate [9, 12, 17, 19, 25, 40, 41], and nine works that deal with mortality [15, 16, 21, 23, 28, 29, 31, 40, 42]. Note that some works deal with more than one outcome and are thus mentioned more than once. Most of the works analyse time intervals before the vaccination, however two studies [31, 34] analyse time intervals when vaccines are used. Eventhough some papers consider only a few selected countries, 24 of the works include either all US states or at least 50 countries worldwide.

Boxplots of the effectiveness values of the NPIs are shown in Figure 1. Each box extends from the lower to the upper quartile of the NPI data, with a line at the median. The whiskers extending from the box show the range of the data. The most effective NPIs overall are restrictions on gatherings (C4), workplace closing (C2), public information campaigns (H1), and school closing (C1) with mean effectiveness value of 1.91, 1.92, 2.0, and 2.08, respectively. The NPIs with moderate effectiveness are stay at home requirements (C6), cancel public events (C3), restrictions on internal movement (C7), facial coverings (H6), and international traven controls (C8) with mean effectiveness value of 2.25, 2.54, 2.58, 2.63, and 2.75, respectively. The least effective NPIs are close public transport (C5), contact tracing (H3), and testing policy (H2), with mean effectiveness value of 3.33, 3.33, and 3.75, respectively. At this point it is important to note that Herby et al. [5] determined that lockdowns, which we find to have a moderate effect, only reduced deaths by 0.2–2.9 %.

4 Limitations

This review has the following limitations. Because the studies included in the review are based on experimental data, the NPIs are always used simultaneously, whereas the final results of the NPI effects are reported individually. Because combinations of NPIs active at the same time were very similar in different regions and time intervals, it is sometimes difficult to justify treating the effects separately.

In some papers, NPIs were not ranked, so these NPIs receive the same value in our study. In addition, some effectiveness values were assigned based on descriptive ranking.

Results are reported here as steady-state rankings, even though the effects of NPIs will change as they are imple-

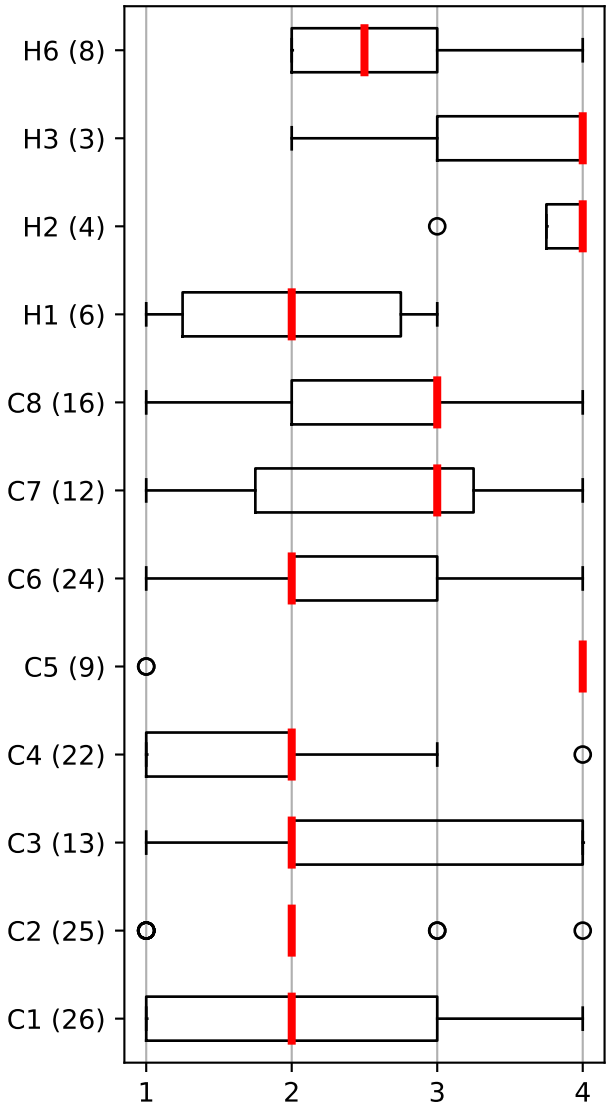


Figure 1: Boxplot of the NPIs' effectiveness. Value one corresponds to the maximum and four to the minimum effectiveness. The numbers in parentheses indicate the number of times the NPIs occurred in the studies examined.

mented (e.g., as people stop complying with restrictions on gatherings, as vaccines are developed, etc.). In addition, the time intervals studied vary in length, and the effects could differ between short and long intervals, as the effects of some NPIs diminish over time [43]. The NPIs are implemented with different stringency according to the Oxford database. This means that our results apply only to the average levels of stringency at which the NPIs can be implemented. Some NPIs may be much more effective (less effective) when implemented with higher (lower) stringency.

5 Conclusion

In this work, we reviewed 34 studies that assessed the effectiveness of 12 non-pharmaceutical interventions against COVID-19. The studies are all based on experimental data and cover up to 200 countries and regions worldwide with different time intervals covering time span between December, 2019 and August, 2021. We found that the overall most effective interventions are restrictions on gatherings, workplace closing, public information campaigns, and school closing. The interventions with moderate impact are stay at home requirements, cancel public events, restrictions on internal movement, facial coverings, and international travel controls. The interventions with the least amount of impact are close public transport, contact tracing, and testing policy.

Acknowledgement

The paper was supported by the ISE-EMH project funded by the program Interreg V-A Italy-Slovenia 2014-2020. The authors acknowledge the financial support from the Slovenian Research Agency, research core funding No. P2-0209.

References

[1] WHO COVID-19 Dashboard, Geneva: World Health Organization, https://covid19.who.int/. Accesed May 2022.

[2] J. Tomšič, D. Susič, and M. Gams (2021). Effectiveness of non-pharmaceutical interventions in handling the COVID-19 pandemic: review of related studies. In *Proceedings of the 24th international multiconference Information Society - IS 2021*, vol. D, pp. 46–51.

[3] A. Mendez-Brito, C. El Bcheraoui, and F. Pozo-Martin (2021). Systematic review of empirical studies comparing the effectiveness of non-pharmaceutical interventions against COVID-19. *Journal of Infection*, vol. 83, no. 3, pp. 281–293, https://doi.org/10.1016/j.jinf.2021.06.018.

[4] I. Ayouni, J. Maatoug, W. Dhouib et al. (2021). Effective public health measures to mitigate the spread of COVID-19: a systematic review. *BMC Public Health*, vol. 21, no. 1, pp. 10–15, https://doi.org/10.1186/s12889-021-11111-1.

[5] J. Herby, L. Jonung, and S. H. Hanke (2022). A literature review and meta-analysis of the effects of lockdowns on COVID-19 mortality. *Studies in Applied Economics*, vol. 200.

[6] T. Hale, N. Angrist, and R. Goldszmidt (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, vol. 5, pp. 4, pp. 529–538, https://doi.org/10.1016/j.jinf.2021.06.018.

[7] Z. Niu and G. Scarciotti (2021). Ranking the effectiveness of non-pharmaceutical interventions to counter COVID-19 in UK universities with vaccinated population. *medRxiv*, https://doi.org/10.1101/2021.11.07.21266028.

[8] H. H. Suh, J. Meehan, L. Blaisdell, and L. Browne (2021). Non-pharmaceutical interventions and COVID-19 cases in US summer camps: results from an American Camp Association survey. *Journal of Epidemiology & Community Health*, vol. 76, no. 4, pp. 327–334, https://doi.org/10.1136/jech-2021-216711.

[9] E. Bendavid, C. Oh, J. Bhattacharya, and J. P. A. Ioannidis (2021). Assessing mandatory stay-at-home and business closure effects on the spread of COVID-19. *European Journal of Clinical Investigation*, vol. 51, no. 4, article no. e13484, https://doi.org/10.1111/eci.13484.

[10] Y. Li, H. Campbell, D. Kulkarni et al. (2021). The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. *The Lancet Infectious Diseases*, vol. 21, no. 2, pp. 193–202, https://doi.org/10.1016/S1473-3099(20)30785-4.

[11] Y. Liu, C. Morgenstern, J. Kelly et al. (2021). The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC medicine*, vol. 19, no. 1, pp. 1–12, https://doi.org/10.1186/s12916-020-01872-8.

[12] P. D. Wibbens, W. W. Koo, and A. M. McGahan (2020). Which COVID policies are most effective? A Bayesian analysis of COVID-19 by jurisdiction. *Plos one*, vol. 15, no. 12, article no. e0244177, https://doi.org/10.1371/journal.pone.0244177.

[13] P. Jüni, M. Rothenbühler, and P. Bobos et al. (2020). Impact of climate and public health interventions on the COVID-19 pandemic: a prospective cohort study. *CMAJ*, vol. 192, no. 21, pp. 566–573, https://doi.org/10.1503/cmaj.200920.

[14] A. M. Jalali, S. G. Khoury, J. See et al. (2020). Delayed interventions, low compliance, and health disparities amplified the early spread of COVID-19. *medRxiv*, https://doi.org/10.1101/2020.07.31.20165654.

[15] C. T. Leffler, E. Ing, J. D. Lykins et al. (2020). Association of country-wide coronavirus mortality with demographics, testing, lockdowns, and public wearing of masks. *The American journal of tropical medicine and hygiene*, vol. 103, no. 6, pp. 2400–2411, https://doi.org/10.4269/ajtmh.20-1015.

[16] D. I. Papadopoulos, I. Donkov, K. Charitopoulos, and S. Bishara (2020). The impact of lockdown measures on COVID-19: a worldwide comparison. *medRxiv*, https://doi.org/10.1101/2020.05.22.20106476.

[17] V. Chernozhukov, H. Kasahara, and P. Schrimpf (2021). Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S. *Journal of Econometrics*, vol. 220, no. 1, pp. 23–62, https://doi.org/10.1016/j.jeconom.2020.09.003.

[18] P. Deb, D. Furceri, J. D. Ostry, and N. Tawk (2021). Policy interventions, social distancing, and SARS-CoV-2 transmission in the United States: a retrospective state-level analysis. *The American journal of*

the medical sciences, vol. 361, no. 5, pp. 575–584, https://doi.org/10.1016/j.amjms.2021.01.007.

[19] S. Ebrahim, H. Ashworth, and C. Noah (2020). Reduction of COVID-19 incidence and nonpharmacologic interventions: Analysis using a US county–level policy data set. *Journal of medical Internet research*, vol. 22, no. 12, pp. 575–584, https://doi.org/10.1016/j.amjms.2021.01.007.

[20] S. Flaxman, S. Mishra, A. Gandy et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, vol. 584, no. 7820, pp. 257–261, https://doi.org/10.1038/s41586-020-2405-7.

[21] P. R. Hunter, F. J. Colón-González, J. Brainard, and S. Rushton (2021). Impact of non-pharmaceutical interventions against COVID-19 in Europe in 2020: a quasi-experimental non-equivalent group and time series design study. *Euro Surveill*, vol. 26, no. 28, https://doi.org/10.2807/1560-7917.ES.2021.26.28.2001401.

[22] A. M. Olney, J. Smith, S. Sen et al. (2021). Estimating the effect of social distancing interventions on COVID-19 in the United States. *American Journal of Epidemiology*, vol. 190, no. 8, pp. 1504–1509, https://doi.org/10.1093/aje/kwaa293.

[23] D. Piovani, M. N. Christodoulou, A. Hadjidemetriou et al. (2021). Effect of early application of social distancing interventions on COVID-19 mortality over the first pandemic wave: an analysis of longitudinal data from 37 countries. *Journal of Infection*, vol. 82, no. 1, pp. 133–142, https://doi.org/10.1016/j.jinf.2020.11.033.

[24] P. D. Wibbens, W. W. Koo, and A. M. McGahan (2020). Evaluation on different non-pharmaceutical interventions during COVID-19 pandemic: an analysis of 139 countries. *J Infect*, vol. 81, no. 3, pp. 70–71, https://doi.org/10.1016/j.jinf.2020.06.

[25] X. Zhang and M. E. Warner (2020). COVID-19 policy differences across US states: shutdowns, reopening, and mask mandates. *Int J Environ Res Public Health*, vol. 17, no. 24, article no. 9520, https://doi.org/10.3390/ijerph17249520.

[26] N. Askitas, K. Tatsiramos, and B. Verheyden (2021). Estimating worldwide effects of non-pharmaceutical interventions on COVID-19 incidence and population mobility patterns using a multiple-event study. *Scientific reports*, vol. 11, no. 1, article no. 1972, https://doi.org/10.1038/s41598-021-81442-x.

[27] N. Banholzer, E. Van Weenen, A. Lison, A. Cenedese, A. Seeliger et al. (2021). Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave. *PLOS ONE*, vol. 16, no. 6, article no. e0252827, https://doi.org/10.1371/journal.pone.0252827.

[28] R. Chaudhry et al. (2021). A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. *eClinicalMedicine*, vol 25, article no. 100464, https://doi.org/10.1016/j.eclinm.2020.100464.

[29] P. Deb, D. Furceri, J. D. Ostry, and N. Tawk (2020). *The effect of containment measures on the COVID-19 pandemic.* Centre for Economic Policy Research.

[30] N. Islam, S. Sharp, G. Chowell et al. (2020). Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. *British Medical Journal*, vol. 370, article no. m2743, https://doi.org/10.1136/bmj.m2743.

[31] M. Sharma, S. Mindermann, C. Rogers-Smith et al. (2021). Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nature Communications*, vol. 12, no. 1, article no. 5820, https://doi.org/10.1038/s41467-021-26013-4.

[32] L. Y. Chan, B. Yuan, and M. Convertino (2021). COVID-19 non-pharmaceutical intervention portfolio effectiveness and risk communication predominance. *Sci Rep*, vol. 11, no. 1, article no. 10605, https://doi.org/10.1038/s41598-021-88309-1.

[33] Y. Gokmen, C. Baskici, and Y. Ercil (2021). Effects of non-pharmaceutical interventions against COVID-19: A cross-country analysis. *The International Journal of Health Planning and Management*, vol. 36, no. 4, pp. 1178–1188, https://doi.org/10.1002/hpm.3164.

[34] H. Li, L. Wang, M. Zhang, Y. Lu, and W. Wang (2022). Effects of vaccination and non-pharmaceutical interventions and their lag times on the COVID-19 pandemic: comparison of eight countries. *PLOS Neglected Tropical Diseases*, vol. 16, no. 1, article no. e0010101, https://doi.org/10.1371/journal.pntd.0010101.

[35] Y. Bo, C. Guo, C. Lin et al. (2021). Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020. *International Journal of Infectious Diseases*, vol. 102, pp. 247–253, https://doi.org/10.1016/j.ijid.2020.10.066.

[36] J. M. Brauner, S. Mindermann, M. Sharma et al. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science*, vol. 371, no. 6531, article no. eabd9338, https://doi.org/10.1126/science.abd9338.

[37] R. T. Esra, L. Jamesion, M. P. Fox et al. (2020). Evaluating the impact of non-pharmaceutical interventions for SARS-CoV-2 on a global scale. *medRxiv*, https://doi.org/10.1101/2020.07.30.20164939.

[38] N. Haug, L. Geyrhofer, A. Londei et al. (2020). Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat Hum Behav*, vol. 4, no. 12, pp. 1303–1312, https://doi.org/10.1038/s41562-020-01009-0.

[39] W. C. Koh, L. Naing, J. Wong et al. (2020). Estimating the impact of physical distancing measures in containing COVID-19: an empirical analysis. *International Journal of Infectious Diseases*, vol. 100, pp. 42–49, https://doi.org/10.1016/j.ijid.2020.08.026.

[40] Y. Li, M. Li, M. Rice et al. (2021). The impact of policy measures on human mobility, COVID-19 cases, and mortality in the US: a spatiotemporal perspective. *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, article no. 996, https://doi.org/10.3390/ijerph18030996.

[41] F. Pozo-Martin, H. Weishaar, F. Cristea et al. (2021). The impact of non-pharmaceutical interventions on COVID-19 epidemic growth in the 37 OECD member states. *European journal of epidemiology*, vol. 82, no. 1, pp. 1–12, https://doi.org/10.1016/j.jinf.2020.11.033.

[42] J. Stokes, A. J. Turner, L. Anselmi et al. (2020). The relative effects of non-pharmaceutical interventions on early Covid-19 mortality: natural experiment in 130 countries. *medRxiv*, https://doi.org/10.1101/2020.10.05.20206888.

[43] F. Zhou, T. J. Hu, X. Y. Zhang et al. (2022). The association of intensity and duration of non-pharmacological interventions and implementation of vaccination with COVID-19 infection, death, and excess mortality: natural experiment in 22 European countries. *Journal of Infection and Public Health*, vol. 15, no. 5, pp. 499-507, https://doi.org/10.1016/j.jiph.2022.03.011.

Table 1: Studies included in this review.

Authors	**NPI data source**	**Countries covered**	**Time interval**
Askitas et al. [26]	OxCGRT	175 countries	unclear
Banholzer et al. [27]	Collected by the authors	USA, Canada, Australia and 17 EU countries	Feb – May, 2020
Bendavid et al. [9]	COVID-19 policy data-bank	10 countries	Dec, 2019 – June, 2020
Bo et al. [35]	Collected by the authors	190 countries	Jan – Apr, 2020
Brauner et al. [36]	Collected by the authors	41 countries	Jan – May, 2020
Chan et al. [32]	WHO and John Hopkins University	50 countries	Dec, 2019 – June, 2020
Chaudhry et al. [28]	Collected by the authors	50 countries	Dec, 2019 – May, 2020
Chernozhukov et al. [17]	COVID Tracking Project	USA	Mar – June, 2020
Deb et al. [29]	OxCGRT	129 countries	Dec, 2019 – May, 2020
Dreher et al. [18]	unclear	USA	Dec, 2019 – Apr, 2020
Ebrahim et al. [19]	Hikma Health	1320 US counties	Mar – July, 2020
Esra et al. [37]	WHO-PHSM	26 countries and 34 US states	Dec, 2019 – May, 2020
Flaxman et al. [20]	unclear	11 EU countries	Feb – May, 2020
Gokmen et al. [33]	Our World in Data	4 countries	Dec, 2019 – June, 2020
Haug et al. [38]	CCCSL	56 countries	Dec, 2019 – Aug, 2020
Hunter et al. [21]	IHME	30 European countries	Dec, 2019 – Apr, 2020
Islam et al. [30]	OxCGRT	149 countries	Dec, 2019 – May, 2020
Jalali et al. [14]	Collected by the authors	30 US states	Mar – May, 2020
Jüni et al. [13]	Collected by the authors	144 worldwide regions	Dec, 2019 – Mar, 2020
Koh et al. [39]	OxCGRT	170 countries	Jan – May, 2020
Leffler et al. [15]	OxCGRT	200 countries	Dec, 2019 – May, 2020
Li et al. (a) [10]	OxCGRT	131 countries	Jan – July, 2020
Li et al. (b) [40]	NSF spatiotemporal center	USA	Mar – July, 2020
Liu et al. [11]	OxCGRT	130 countries and territories	Jan – June, 2020
Olney et al. [22]	Collected by the authors	USA	Feb – Apr, 2020
Papadopoulos et al. [16]	OxCGRT	151 countries	Jan – Apr, 2020
Piovani et al. [23]	OxCGRT	37 members of OECF	Jan – June, 2020
Pozo-Martin et al. [41]	OxCGRT and WHO-PHSM	37 members of OECD	Oct – Dec, 2020
Sharma et al. [31]	Collected by the authors	7 EU countries	Aug, 2020 – Jan, 2021
Stokes et al. [42]	OxCGRT	USA and 7 countries	Dec, 2019 – June, 2020
Wang et al. [34]	OxCGRT	139 countries	Dec, 2019 – Aug, 2021
Wibbens et al. [12]	OxCGRT	40 countries and US states	unclear
Wong et al. [24]	OxCGRT	139 countries	Mar – Apr, 2020
Zhang et al. [25]	NY Times and CNN	USA	Feb – Aug, 2020

Table 2: Estimation of effectiveness of NPIs in reviewed studies.

Study	C1	C2	C3	C4	C5	C6	C7	C8	H1	H2	H3	H6
Askitas et al. [26]	1	1	1	1	4	2	4	3				
Banholzer et al. [27]	2	2		1		4		3				
Bendavid et al. [9]	3	4	3	2	1	1	4	3				
Bo et al. [35]	1		1	1		3	4	4				2
Brauner et al. [36]	1	2		1		3						
Chan et al. [32]			4	4			1	1			2	
Chaudhry et al. [28]		2		2		2		3				
Chernozhukov et al. [17]		2				2						3
Deb et al. [29]	1	2	2	2	1	1	2	1				
Dreher et al. [18]	2	2				1						
Ebrahim et al. [19]		2				3						
Esra et al. [37]		3	3			1						2
Flaxman et al. [20]	4		4			3						
Gokmen et al. [33]	4	1	4	2	4	2	3	3				
Haug et al. [38]	1			1	4	3	3	2	2	3		
Hunter et al. [21]	1	2		3								
Islam et al. [30]	2	2		1	4	3	3					
Jalali et al. [14]	2											2
Jüni et al. [13]	2	2		2								
Koh et al. [39]		1				2	2	3				
Leffler et al. [15]	2		2					2				2
Li et al. (a) [10]	1	2	1	3	4	2	3	4				
Li et al. (b) [40]		2	2			3			1			
Liu et al. [11]	1	1	2	2	4	3	1	4	3	4	4	
Olney et al. [22]	2			1		1						
Papadopoulos et al. [16]	2	2						2	2			
Piovani et al. [23]	3			2								
Pozo-Martin et al. [41]	3	2		1						4		4
Sharma et al. [31]	4	1		2		3						3
Stokes et al. [42]	1	2		3				3				
Wang et al. [34]	3	3		2		2						
Wibbens et al. [12]	2	1	4	3	4	2	1	3	3	4	4	
Wong et al. [24]	3	2							1			
Zhang et al. [25]						2						3

An IoT-Based Pill Management System for Elderly

Boudrali Roumaissa *[1] and Boudour Rachid [1]
E-mail: boudraliromaissadoc@gmail.com, racboudour@yahoo.fr
* Corresponding author

[1] Department of Computer Science, University of Badji Mokhtar
Annaba, Algeria

Keywords: smart healthcare, IoT, pill dispenser, ambient assisted living, medication adherence

Received: June 14, 2022

The challenge to guarantee healthy aging has become a major social concern. Due to the cognitive deficits related to age, hectic daily activities, and multiple medications prescriptions, the elders often tend to forget their pills intakes. This has a colossal impact on their health and life. Moreover, the recent pandemic of COVID-19 has accentuated the importance to provide independent and autonomous living for the elderly. This paper presents a pill management system based on IoT intended for aged individuals. The proposed system is a smart pill dispenser associated with a mobile application. The main actors of the system are the patient, the doctor, the pharmacist, and the patient's caregiver and/or relative, each having restricted access to the system via specific credentials. The prescription is directly edited on the mobile application by the doctor and the scheduling and filling of the pillbox is done wirelessly by the pharmacist. The reminding of medications intakes in this system is done gradually to help the patient adhere to his/her prescription. First, it notifies the patient about his/her scheduled pill by a mobile notification then via the pill dispenser using LCD, LED, and buzzer. The implemented system also allows the doctor and caregiver to keep a tab of the patient's intakes. Furthermore, the pill dispenser is featured with a locking mechanism to ensure medication dosage control and prevent drugs abuse. Experiments show that the proposed system is appreciated by the elderly and encourages them to take their pills successfully and safely without causing any disturbance.

Povzetek: Zaradi kognitivnih omejitev, povezanih s starostjo, napornih dnevnih dejavnosti in množice predpisanih zdravil, starejši pogosto pozabijo vzeti predpisana zdravila. Ta članek predstavlja sistem za upravljanje jemanja zdravil, ki temelji na internetu stvari in je namenjen starejšim posameznikom. Predlagani sistem sestoji iz pametnega razdeljevalnika zdravil, ki je povezan z mobilno aplikacijo. Poleg pomoči pri pravilnem jemanju zdravil sistem omogoča tudi nadzor jemanja zdravil s strani svojcev, zdravnikov in negovalcev.

1 Introduction

In the last few years, the average age of the world population has been growing rapidly. Statistics affirm that the number of the elderly in the world is growing by 3.26% per year [1]. In addition, according to recent reports of health organizations, world societies are expecting more growth in communities of elderly individuals. Global Age watch index estimates that the aging population will continue to increase to reach 1.4 billion in 2030 and 2.1 billion by 2050 [2].

As the rate of aged individuals grows, the number of individuals suffering from chronic diseases such as diabetes, high blood pressure will continue to rise. These conditions usually require people to take multiple medications regularly to help them complete their daily life activities safely and autonomously. Thus, due to cognitive deficits correlated to age as memory deficits, older people with intricate medication routines often tend to forget about their intakes, timings, and doses. This frequently leads to wrong medication intakes, which can cause serious health consequences. Medication adherence is considered a major medical concern. To support them in medication intake, elderlies often appeal for help. However, relatives and caregivers who frequently help the elderly in remembering their medication intakes face a daily burden while dealing with their own lives and assisting these individuals. In addition, the recent pandemic of Covid-19 has emphasized the importance of providing independent autonomous living to seniors.

In this perspective, multiple tools and solutions were developed to support the elderly in this particular activity. It was found that senior individuals have high acceptability of using smart kits and technological solutions [3]. Smart pill dispensers are one of the most preferred solutions [4]. A study shows that the demand for smart pill dispensers will keep increasing [5].

The use of IoT paradigm in the healthcare industry has offered various efficient applications and systems with different architectures [29].

In this context, we present an IoT based smart pill management system. It consists of a connected pill dispenser connected with a mobile application installed for four main users. The main users of the proposed

system are the patient, the doctor, the pharmacist, and the caregiver. The developed system works on two main modes: the assistance mode and the programming mode. In the assistance mode, the pillbox continuously checks the current time using NTP (network timing protocol). Before 10 minutes of the prescheduled intake, a notification is sent to the mobile phone of the patient in order to notify him/her of his/her intake. When the intake time arrives, the pillbox emits sounds and lights using a buzzer and a LED to alert the patient. An LCD is also used to display directives for the patient. However, if the patient decided to take his/her pill in the 10 minutes that precede the right intake time, the pill dispenser alarms are automatically cancelled.

In the programming mode, the pharmacist schedules wirelessly the pill dispenser and refills the pills compartment one by one using a simple interface on his/her cellphone. The doctor is also able to write and edit prescriptions for his/her patient remotely anytime. For medication adherence monitoring, the proposed system provides all of the users with real-time intakes history. In addition to the gradual assistance in medication reminders, the presented system keeps safe the pills, the patient, and the relatives living with him.

The rest of this paper is structured as follows: Section 2 presents a brief review of the existing pill dispensers. The system architecture as well as its operating logic and modes are presented in Section 3. Section 4 presents the design and implementation of the mobile APP and the pill dispenser. Section 5 contains the results and discussion. Finally, conclusions and perspectives are presented in Section 6.

2 Literature review

Existing medication adherence solutions vary from simple electronic devices which are the traditional pill organizers [6] to complex intelligent systems such as pill dispensers. This paper mainly focuses on the review of the recently developed pill dispensers. Medication dispensers as electromechanical organizers are able to dispense the right pills with the right doses for the users [7]. By comparing different conceptual and implementation aspects, pill dispensers may be reviewed according to the used pill dispensing mechanism, the developed programming mode of the dispenser, the way pills intakes' reminders are delivered and the developed intakes monitoring method.

To remind the assisted individual of his/her medication, first developed pill dispensers mainly used visual and acoustic reminders such as buzzers to emit sounds and LEDs to turn on lights [8].

The authors of [9] used an eccentric rotating mass (vibrator) along with a buzzer to provide the alarms; while [10] [11] [12] implemented playback modules and speakers for pills reminding, in addition to LEDs and LCDs that provide visual assistance. Others opted for the use of vocal messaging to provide the intakes alarms. In [11] a voice message saying "take vitamin tablet" is announced whenever it is time for the scheduled pills.

The use of smartphones in healthcare as well as Android Applications is found to be a powerful and promising manner to improve consumer-oriented products [26] [27].

There are pill management systems that have associated mobile applications to the pillboxes [12] [13] [14]. These systems use mobile notifications to alert the patient about his/her intakes. Others, e.g., [15] used both mobile application notifications along with pill dispenser alarms to alert the patient. However, the user may be confused due to these double reminders and may take his/her pills twice. Furthermore, many of the proposed systems used LCDs or OLEDs screens to display medicines' relative information [16].

The second aspect to discuss pill management systems is the way the pill dispenser is programmed and filled. Most of the developed dispensers require that the caregiver or the patient repeatedly fill the pills and schedule the intake timings [17] [18] [8] [9] [12] [19]. Thus, they leave the programming and filling mode unsecured. In this case, the patient has access to the pill dispenser outside the intake hours. Also, anyone may fill the pills and even edit pill schedules. [20] and [21] point out the importance of respecting prescriptions and the consequences of mistaking pills, so it is very important to secure pill dispensers and make pill scheduling and filling process accessible only by health professionals.

To secure the pill dispenser, [15] used an identification protocol based on person identification via camera, thus no information was provided whether the programming mode of the pillbox is also secured. Others like [16] secure the scheduling process but ignore the security of the filling process. In their system, the scheduling is done by physicians through a web application. Thus, no information was given on how the filling is done. In [14] the patient himself/herself is responsible for pill filling whereas it is the doctor who does the scheduling of the intakes.

Another important detail in pill dispensers is the method used for dispensing the medication. The dispensing method used in medication dispensers is very important since it impacts the medication adherence of the patient and also the security of both the patient and the relatives living with him/her as well.

Some of the developed systems use vital signs to dispense the pills only when it is necessary [22] [23]. However, this is not adopted for elderly individuals with memory deficits since they do not emit any reminders. The authors of [15] and [24] use ultrasound sensors to detect the presence of the individual before dispensing the pills. If a presence is detected, the scheduled pill will automatically be dispensed. Yet, this may put at risk infants or illiterate adults living with the patient, since it blindly frees the medication. Similar solutions are described in [14] and [25] which automatically open when the intake time comes. In case of non-response from the patient, the pills remain accessible for all the individuals living with him/her.

In [18] an infrared sensor is used to detect if the pill has been taken or not. However, if the patient does not respond in 10 min the pillbox locks automatically.

A simple yet efficient technique for pill discard was used by [10] and [12]. The tablet compartment opens only if a

push-button is pressed. The authors of [15] used a more sophisticated technology, where a PIR sensor and a camera are used. Whenever a movement is detected around the pillbox, the camera is activated and identifies the person around. If it is a correct intake time, the pillbox unlocks and dispenses the medication. This method is the most secure technique for pill dispensing. However, in some cases when the senior is not ready to take his/her pill yet and passes around the pillbox, the pill will be dispensed and will remain accessible to the patient relatives which represents a risky situation. Also, medication remains exposed to ambient factors such as temperature and humidity. Furthermore, the use of cameras may be displeasing for some seniors as they see it as a privacy issue.

For the monitoring techniques, some of the presented systems such as [17] [12] [13] provide no track of the intakes. Others use a variety of methods to this end. The monitoring technique used in [8] [10] [18] relies on sending messages alerts to the patient's caretaker or doctor if he/she does not take his/her scheduled pill. In addition to sending an SMS to the caregiver, [8] uses IR sensors in the compartments, whenever the sensor detects a presence, the pill is supposed to be taken. These two methods are useful yet they do not offer detailed medication tracking.

The authors of [28] used body sensors to develop a health monitoring system for individuals with cardiac risk. Their findings proved the efficiency of using sensors in preserving peoples' health.

In pill management systems, biosensors are also used to report the state of the user and check whether the medications have been taken on time or not. The authors of [25] [13] developed a more suitable solution for intake supervisions. In their system, the pharmacist and the doctor monitor the consumption of the patient via their mobile application. However, the patient has no record of his/her intakes. To confirm the real consumption of pills a load sensor HX711 is used [19]. Whenever the patient takes the pill from the pill dispenser, the weight decreases, and the intake is confirmed. However, this pill dispenser is not adapted for elderly use.

Despite the proven efficiency of the cited pill dispensers, they do not completely meet the needs of elderly population. While developing a pill dispenser for elder individuals, it is very important to take into consideration the cognitive deficits that the elderly may suffer such as memory, initiation and planning deficits. Also, it has been noted that some of the used techniques require that the patient interacts directly with the system, but for illiterate seniors or those lucking for skills and abilities, this may not be suitable. In addition, to protect the pill dispenser from unauthorized individuals, particular attention to security and safety issues is required. Through the literature review, it was noticed that few or no smart pill management system regroups all the elderly needs, also no safety measures were developed. In addition, none of them offers a gradual reminding to the users.

The main objective of this paper is to offer the elderly with memory deficiencies a pill management system that offers gradual assistance for medication intake. Even though many pill dispensers were developed, no medication management system offered gradual assistance. Assisting the elderly with a gradual method is an important aspect that helps in medication adherence and technology acceptance. This aspect is often ignored in previous research. The existing pill dispensers use either smartphone notifications or pill dispensers' alarms, the few that reassemble both techniques often cause disturbance to the users. The assistance acts in this pill management system are provided by both the dispenser and the smartphone with gradation. The provided acts respect as well the acoustic and eyesight deficits the elderly may suffer from because of old age.

Through research, multiple pill dispensers are proposed, most of them are similar and do not pay attention to security aspects. Hence, in the present work, the security of the senior using the pillbox as well as the individuals living with him/her are taken into consideration.

Moreover, the existing pill dispensers require physical intervention to be programmed, while the dispenser of this system is wirelessly and securely programmed. The connected pill management system is fully connected and the intake tracking is done remotely.

We developed a pill management system adapted to elderly needs suffering from mild cognitive impairments related to age. In this system, we address multiple issues. Mainly, simple gradual assistance is provided for the users so as to increase their adherence to medication by giving them more time to take their scheduled pills. Our system uses multiple means of assistance. This way individuals with visual or acoustic deficits and also elderly with few skills in operating IT are able to handle productively. The implemented pill dispenser is connected with a mobile application that ensures full medication monitoring for the preauthorized users, and it is manipulated only by health professionals. A secured dispensing mechanism is implemented so that the pills stay out of reach of unauthorized individuals and also from authorized individuals outside their intakes timings intervals. In addition, no privacy-invading technique was used in the development of this system.

3 System design

Figure 1 presents the overall system units and the main users. The proposed system is used by four main users: the patient, the doctor, the pharmacist, and the relative of the patient. It is composed mainly of a mobile application developed under Android Studio and a connected pill dispenser. The communication between all users and the pillbox as well as the communication between the users are done wirelessly using their mobile application and Wi-Fi through WebSockets. Only two actors, the patient and the pharmacist, interact physically with the pill dispenser.

Figure 2 shows the block diagram of the designed system. It shows the main components and modules of the pill dispenser. The block diagram was designed to be used as a template while developing the system. The five main modules that compose our system, namely the hardware assistance modules, the control unit, the dispensing

module, and the software module will be discussed in Section 4.

The ESP8266 is considered as the communication unit of this system since it provides the connection between the mobile applications and as it is responsible for the communication between all actors and the mainboard. Arduino UNO controls all of the hardware components of the pill dispenser, i.e., the pill dispenser mechanism and the hardware assistance part. A power supply is used to make the pillbox portable. The software assistance part is the patient mobile application. It communicates mainly with the considered Wi-Fi module.

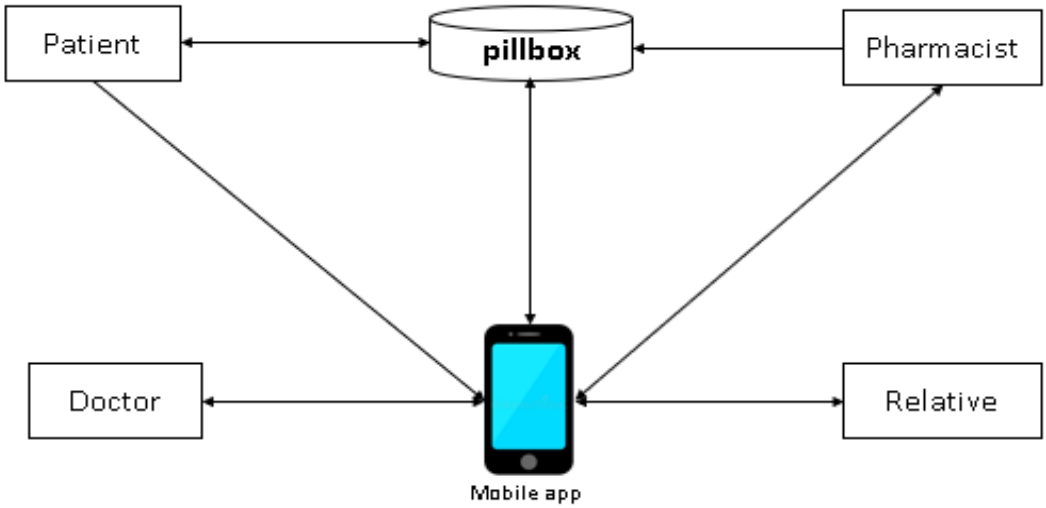


Figure 1: Overall system actors and components.

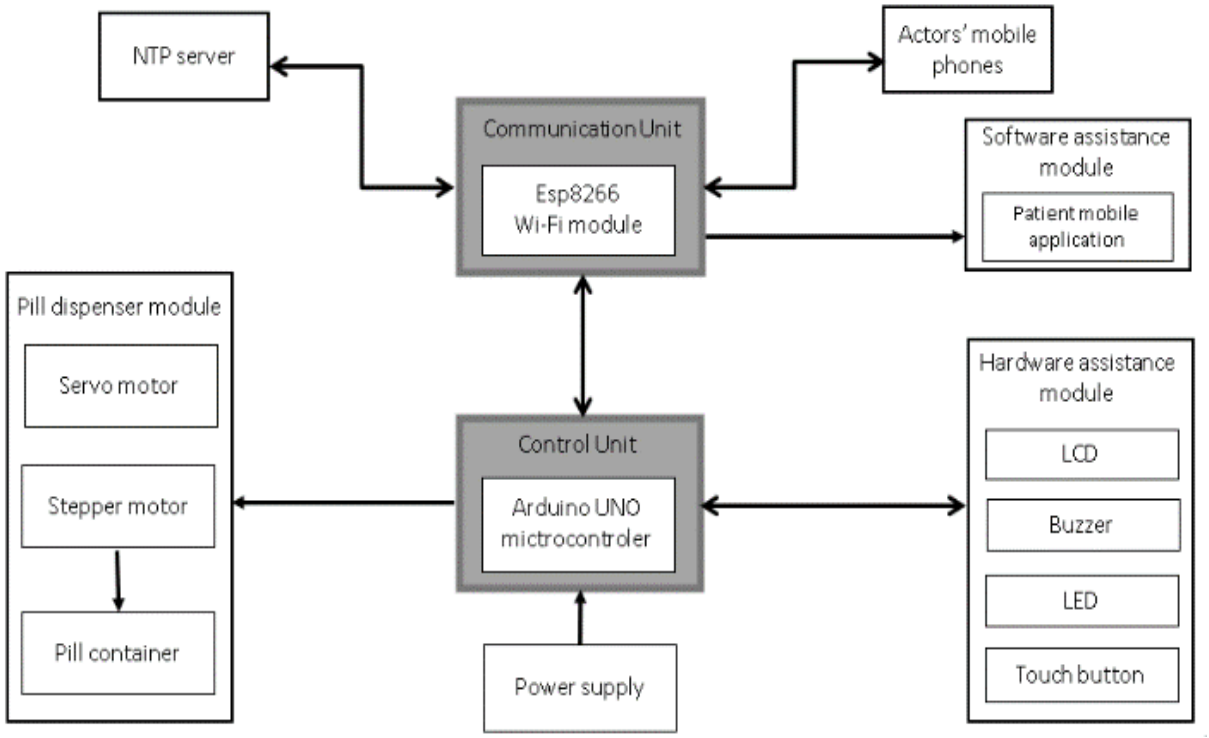


Figure 2: Block diagram of the system.

3.1 Operating logic

Figure 3 shows the operating logic of the proposed system.

To operate in assistance mode, the smart pillbox needs to be programmed and filled at the start.

The box checks the actual time using the internet. Before 10 minutes of a scheduled intake, the system emits a notification through the mobile application on the main user's smartphone to prepare him/her for his/her intake. In the 10 minutes that follow if the patient decides to take his/her pill, he/she can easily unlock the box by

confirming his/her presence using the touch button. If so, the pillbox opens the right compartment. If the user confirms his/her intake, the box locks up, sends feedback to the mobile application and passes to a standby mode where it keeps checking the actual local time. If the intake is not confirmed, the smart pillbox will automatically lock and notify the user through his/her phone to complete the intake process. If no response is given, the intake will be saved as a non-completed task.

In the case when the 10 minutes elapse and the patient did not confirm his/her presence, the smart pillbox automatically emits sounds, lights, and displays messages on its screen.

If after the pill dispenser alerts the patient does not confirm his/her intake, the smart pillbox goes into a standby mode and saves the intake as a non-completed task.

3.2 Operating modes

The developed pill dispenser operates in two main modes: scheduling and filling mode, and assistance mode.

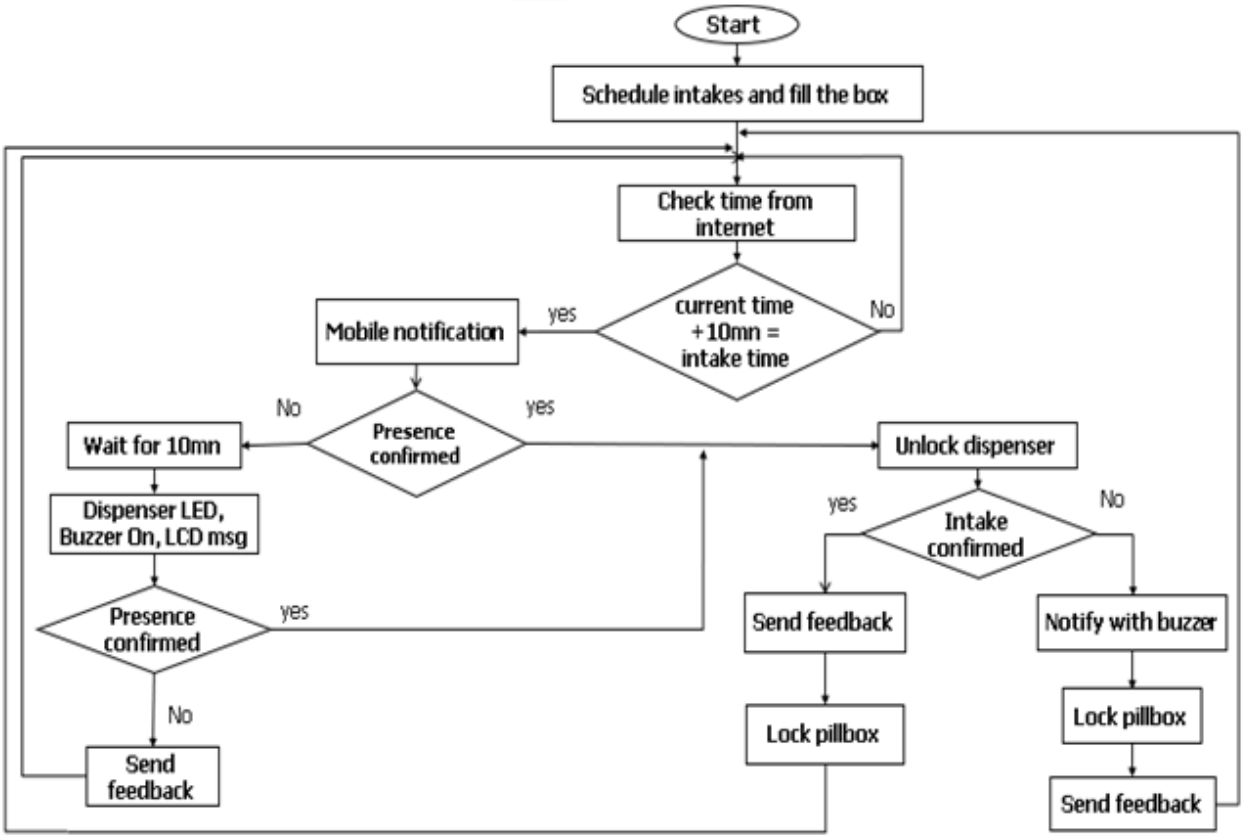


Figure 3: Activity diagram of the proposed system.

3.2.1 Scheduling and filling mode

The programming mode is strictly reserved for health professionals, in our scenario it is reserved for the pharmacist. After accessing his/her account on the mobile app through specific credentials, the pillbox switches automatically to the programming mode. The pill dispenser connects with the pharmacist's app on a local network. It shows directive messages on its screen in order to assist the pharmacist. Through a simple user interface, the pharmacist can easily manipulate the pill dispenser.

The mobile app shows at first the prescription for the pharmacist so that he/she can verify the availability of the medications before proceeding to the next step. Then, by a simple touch on the screen, the pill dispenser is unlocked.

The programming mode contains two modes: the filling part and the scheduling part.

When the pharmacist enters the filling mode, he/she can easily choose which compartment he/she wants to fill by entering its number on the interface shown on his/her phone. The prescribed doses of each medicine are taken into account while filling the pillbox.

When the filling mode is validated, the pharmacist accesses the scheduling mode where he/she plans the intakes timings. The pharmacist refers to the patient's preferences while scheduling the intakes.

To validate the process of filling and scheduling, the pharmacist is asked to confirm or cancel the new modifications. In the end, the pillbox locks up and switches automatically to the assistance mode.

3.2.2 Assistance mode

In the assistance mode, the pillbox is either in standby mode or in reminding mode. In standby mode, the pillbox screen is off to gain battery life and to not disturb the patient during his/her daily life activities. The smart dispenser remains connected to the internet.

When the right time of a pill comes, the screen turns on and starts displaying messages, a LED is turned on as well and a sound is emitted from the buzzer to draw the patient's attention.

The senior is asked to confirm his/her presence before unlocking the pill compartment. After confirmation, the senior is again asked to confirm his/her intake. Confirmations are done by a simple touch on the touch button implemented on the top of the pillbox.

The connected pillbox sends feedback to the mobile application using WebSockets. The feedback represents the history of the patient's intakes.

4 System implementation

The system we propose is composed of two main parts, i.e., the hardware part which is the smart pill dispenser, and the software part which is the mobile application. Both parts are connected which offers an intuitive interface and ease of configuration for the system.

4.1 Software development

The software part of the system consists of a mobile application developed under Android Studio. The mobile application contains multiple accounts for the four main users. Each one of them has access to it through their private accounts using preregistered credentials. The functionalities that the mobile app offers differ from a user to another according to his/her role. To understand the different functions of the application, the following section details the different possible interacts between the pill dispenser and the mobile app and also between the users and the system.

4.1.1 Interaction between patient and mobile app

The patient interacts with the mobile app through a simple interface adapted for elderly individuals. The patient can read his/her updated prescription and can consult his/her scheduled intakes.

Through his/her account, the patient may anytime ask for help by a simple click on the 'ASK for help' button. An important feature that our system offers is the possibility for the patient to check the history of his/her intakes. In case the elder forgets whether he/she took the pill or not, this feature turns out to be very helpful.

4.1.2 Interaction between doctor and mobile app

The mobile application offers the possibility for the doctor to directly write and edit the patient prescriptions remotely. In addition, the doctor is able to consult the history of the patient's intakes in real-time and also his/her pills schedule.

4.1.3 Interaction between relative and mobile app

One of the main purposes that smart pill dispensers were developed for is to reduce the burden of caregivers and relatives. With our mobile application, the caretaker is able to monitor the medication adherence of the elder while completing their regular daily life activities. Through a simple interface, the relative may consult the history of medication intakes, and also the scheduled intakes.

4.1.4 Interaction between pharmacist and mobile app

The mobile app on the pharmacist side provides him/her with the most interesting feature of our system. Through his/her specific credentials, the pharmacist is able to lock and unlock the pillbox, rotate the pills container, and also schedule the intake of the elder. The interfaces of the pharmacist account are all simple and intuitive. The mobile app on the pharmacist's phone connects locally with the pill dispenser. This offers more security for the filling and scheduling mode of the pillbox.

4.2 Hardware circuit and components

For the realization of the prototype, multiple components have been used. The overall used materials and their connection is shown in Figure 4.

The hardware of the proposed pill dispenser may be divided into assistance module, dispensing module, and communication module.

4.2.1 Assistance module

The Assistance module includes all the components responsible for the reminders:

- A LED will light up when it is time to take the pills and keep blinking when the user confirms his/her presence. The LED turns off when the user confirms that he/she took his pill.
- An LCD screen 20x4 shows preprogrammed messages for the user. The LCD screen displays messages such as: "Please confirm your presence" and "Please confirm your intake". For the energy economy, the used screen remains off unless the pill dispenser enters the programming mode or the assistance mode.
- A buzzer is used to emit sounds whenever it is time to take the medicines.
- A touch button helps the user interacts with the messages displayed on the LCD screen. When it is time to take his/her pills, the user unlocks the pillbox via a simple touch.

4.2.2 Dispensing module

The Dispensing module is responsible for controlling the dispensing mechanism and the assistance module.

A simple yet efficient safety mechanism is adopted for this prototype. To dispense the pills, multiple components have been utilized:

- A servo motor is responsible for locking and unlocking the container. When it is time to take the pills and the user confirms his/her presence, the servo motor will turn up to 45 degrees to give access to the right pill storage.
- A stepper motor and its driver are used to rotate the container. The stepper receives signals from the Arduino Uno. Those signals are the addresses of the different compartments. So, whenever an address is received, the stepper motor will rotate and stop at the compartment containing the prescheduled pill.
- A container: a basic round plastic component divided into multiple compartments is used to store the pills.

4.2.3 Communication module

The Communication module is composed of one component which is ESP8266 NODEMCU. This is a development board capable of acting as a Wi-Fi module and a microcontroller as well. In this project, the ESP8266 is used to guarantee the connection to the Internet. This is important to synchronize data through all the users' applications and also to get the local hour from NTP server. In addition, it is used to send feedback from the pill dispenser to mobile applications such as the intake state. The NODEMCU is also used to communicate with the Arduino UNO. It sends specific signals representing coded data. This data represents the compartment address of the container.

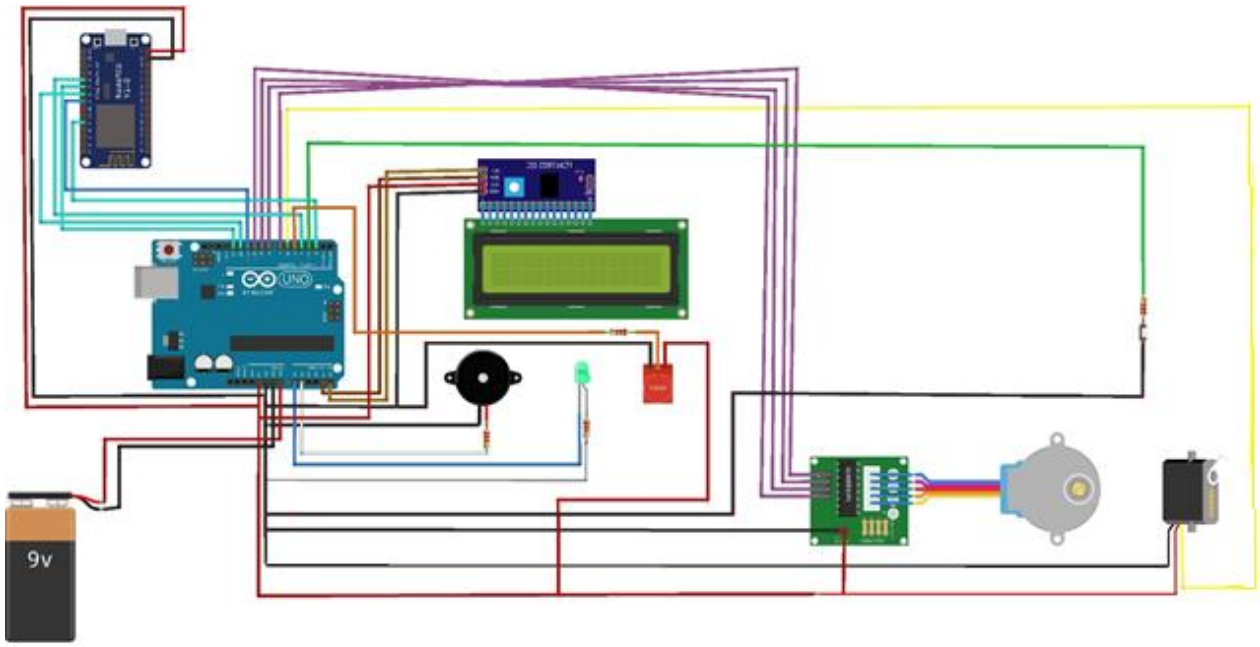


Figure 4: Circuit of the prototype.

4.2.4 Control unit

The Arduino Uno is a microcontroller that has multiple inputs and outputs. It is considered as the control unit of this prototype as it is the component responsible for controlling the dispensing mechanism and the assistance module.

5 Results

The proposed system was tested with the help of seven individuals, three seniors aged between 64 and 67 years old, one caretaker (a relative of subject C), one doctor, and two pharmacists. All related details are shown in Table 1.

The pill dispenser and the mobile application were given to the three seniors one by one for 3 days long. In addition to the chronic conditions, the seniors have small to mild acoustic and visual deficits.

In the testing process, the pill dispenser was first loaded with the pills of each patient respecting the prescribed amounts and timings. The intakes timings were scheduled with the help of each of the testers according to the respective daily routines. The prototype given to the seniors was initially placed in their living room. The participants were free to move it around as they exercise their daily life activities.

The mobile application was installed on each of the users' phones, and specific credentials were given to them. At the end of the testing period, the users reported full contentment about the experience. They stated that the gradual assistance provided by the mobile application and the dispenser helped them take their medication without

any disturbance. They were also enthusiastic about the different assistance acts that the system offers.

They also expressed their satisfaction with the ease of use of the device and the mobile app. In addition, the possibility for them to keep tabs on their intakes helped them in gaining confidence and relieving their stress.

As for the experience of the two pharmacists, they were given the pill dispenser and the mobile application for about 15 minutes each. They expressed right away their appreciation for the overall system. The filling and scheduling process was done efficiently and without any help from the developer.

Table 1: Subjects details.

Subject	Gender	Age	Profession	Health Condition
A	female	64	retired	High blood pressure
B	female	65	retired	High blood pressure
C	female	67	retired	High blood pressure and diabetes
Relative of C	female	36	teacher	
D	male		doctor	
E	male		pharmacist	
F	woman		pharmacist	

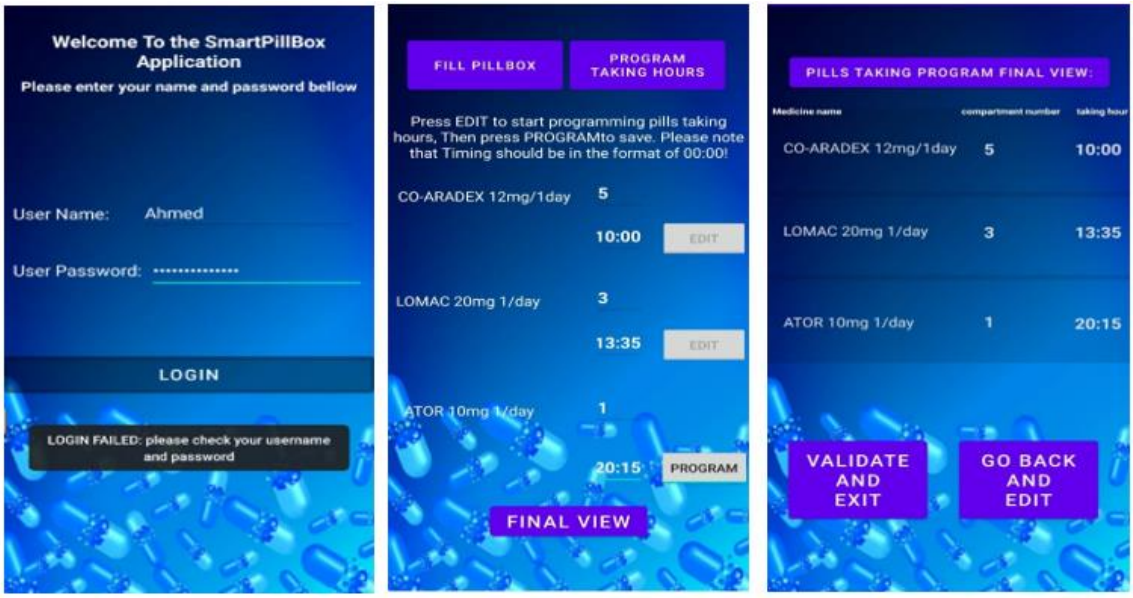


Figure 5: Mobile application screenshots.

The fact that the pill dispenser can be manipulated only by a health professional was very appreciated by them. Hence, suggestions like increasing the number of compartments in the prototype and taking into account pills conditioning were given by one of subjects (E).

To test our pill management system from a doctor's perspective, first the mobile application was installed on a doctor's phone. A specific username and password were then given to him. The doctor reported that he was able to easily fill and edit his patients' prescriptions without any confusion. No help was needed from the developers. He also stated that in some conditions like the recent Covid-19 pandemic, the possibility to edit his patients' prescriptions remotely is a very needed option. Small

suggestions like enhancing the aesthetic aspect of the prototype and adding biosensors to monitor health conditions were given.

For the relative of subject C, the pill management system was appointed as lifesaving. The caretaker of patient C declared his full appreciation for the possibility to track her relative intakes remotely without moving on site. Also, the possibility for the elder to contact his caretaker directly via his mobile application was very admired.

Some of the app's screenshots of the mobile application are shown in Figure 5. The first screenshot represents the welcome interface of the mobile application. An error message is displayed at the bottom when a non-registered user attempts to access the application. The screenshot in the middle represents the interface of the pharmacist programming mode. The interface shows the prescription pre-registered by the doctor. The pharmacist can switch between the programming mode and the filling mode using the buttons at the top of the interface. Besides the name of each medication, two text fields represent respectively the number of the compartment of the medication and the scheduled timing of its intake. The buttons "edit" and "program" are used by the pharmacist to modify these data. The button "final view" at the bottom of this interface leads to the third screenshot. In this screenshot a general view of the prescribed medication is shown, each with the corresponding compartment number and the scheduled intake timing. Using the two buttons at the bottom of the interface the pharmacist has two options: go back and edit the entered information or confirm the settings and exit the application.

6 Discussion

The results of the initial experiments show that the use of the pill dispenser is an efficient solution for pills intakes reminding. The developed system differs from the existing pill management systems on multiple sides. The existing pill dispensers often ignore or miss to address important aspects such as: security, privacy, acoustic and eyesight deficits, and the safety of the elders. To design this system, multiple acts of assistance were provided to ensure the implementation of all the important aspects of this specific daily activity. Using visual and sound alerts, the system offers a maximum assistance.

While most of the existing systems use either mobile notification or pill dispenser's alarm for medication reminders, the proposed system combines both. It uses both techniques with gradation. This constitutes the unicity of the designed system. Offering assistance gradually helps the elders prepare themselves for their intake without causing any perturbance. From the obtained results providing gradual assistance also encourages the elders to take their medication and hence improves their medication adherence and their technology acceptance.

The initial findings are very encouraging. The results open mainly a discussion on how simple assistive technologies can also improve medication adherence, which is an important medical challenge. Further studies and tests for this approach will be conducted.

7 Conclusion and future scope

This paper focused on medication adherence issues in old-age individuals. A pill management system based on IoT has been presented. It is mainly made for the elderly suffering from mild cognitive deficits mainly related to age such as: memory deficits, initiation deficits and planning deficiencies. The significant advantage of this system is that it offers gradual assistance through both software application and hardware components. In addition, the visual and acoustic deficits related to age that most elderly suffer from are considered. Thus, multiple acts of assistance through mobile notifications, LED lighting as well as messages display and sound alarms are provided. In order to keep the seniors and the individuals living with them safe, the pill dispenser is equipped with a locking and unlocking mechanism.

The proposed system permits not only the elderly to have an independent and secure life, but also to reduce the burden of caregivers, family members and doctors since it allows them to supervise and monitor the elderly's intakes remotely. The proposed pill management system is programmed wirelessly only by the pharmacist using specific credentials. Furthermore, the user interface is developed to be intuitive and adapted for both literate and illiterate individuals.

As ongoing works, more tests and experiments will be conducted. The pill dispenser, as well as the mobile application will be given to more seniors of different ages, and backgrounds suffering from different diseases. More health professionals, as well as caregivers will also be invited to test the dispenser and its application. The remarks and suggestions given during the experimental phase will be taken into consideration to enhance the user experience. Furthermore, the duration of the experiments will be extended.

In addition, the system will be enhanced with additional features and more attention will be given to the pill dispenser's look, as well as the size and conditioning of the compartments. To enrich the elderly experience, a speaker could be added to provide vocal assistance. Another interesting extension for this system would be the addition of biosensors such as heartbeat sensors to offer health monitoring of the user. In the smart city context, the pill dispenser compartment could be equipped with sensors such as weight sensors or IR sensors, so that the pill dispenser would be able to order pills before running out from the patient. Furthermore, a database is about to be connected for the whole system to be fully operational. Also, a specific printed circuit board (PCB) could be designed to embed all of the components and make the prototype easy to carry.

References

[1] S. Shahrestani (2018). Internet of things and smart environments. *Cham: Springer International* , pp 1–9. https://doi.org/10.1007/978-3-319-60164-9_1

[2] V. T. Taipale (2014). The global age watch index, GAWI 2013. *Gerontechnology*, vol. 13, pp. 16–20. https://doi.org/10.4017/gt.2014.13.1.010.00

[3] S. L. Smith, J. W. Archer, G. P. Timms, K. W. Smart, S. J. Barker, S. G. Hay, and C. Granet (2012). A millimeter-wave antenna amplitude and phase measurement system. *IEEE transactions on antennas and propagation*, vol. 60, pp. 1744–1757. https://doi.org/10.1109/tap.2012.2186218

[4] B. Reeder, G. Demiris, and K. D. Marek (2013). Older adults' satisfaction with a medication dispensing device in home care. *Informatics for Health and Social Care*, vol. 38, pp. 211–222. https://doi.org/10.3109/17538157.2012.74108

[5] Data Bridge Market Research (2021). *Global Smart Pill Dispenser Market – Industry Trends and Forecast to 2027*. https://www.databridgemarketresearch.com/reports/global-smart-pill-dispenser-market

[6] J. Joy, S. Vahab, G. Vinayakan, M. V. Prasad, and S. Rakesh (2021). SIMoP box–a smart intelligent mobile pill box. *Materials Today: Proceedings*, vol. 43, pp. 3610–3619. https://doi.org/10.1016/j.matpr.2020.09.829

[7] J. F. Pinto, J. L. Vilaça, and N. S. Dias (2021). A review of current pill organizers and dispensers. *9th IEEE International Conference on Serious Games and Applications for Health (SeGAH)*, IEEE, pp. 1–8.

[8] A. Jabeena, A. K. Sahu, R. Roy, and N. S. Basha (2017). Automatic pill reminder for easy supervision. *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, pp. 630–637. https://doi.org/10.1109/iss1.2017.8389315

[9] M. L. I. Goh, M. B. Garcia, P. L. Jay-ar, A. C. Lagman, H. N. Vicente, and R. M. De Angel (2019). A pocket-sized interactive pillbox device: design and development of a microcontroller-based system for medicine intake adherence. *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, IEEE, pp. 718–723. https://doi.org/10.1109/iccike47802.2019.9004276

[10] S. B. Lenin, S. Pushparaj, M. Adithya, S. Murugesan, N. Balaji, and J. Gurupriyan (2021). Smart medkit. *Journal of Physics: Conference Series*, IOP Publishing, vol. 1717, p. 012041. https://doi.org/10.1088/1742-6596/1717/1/012041

[11] V. B. Sree, K. S. Indrani, and G. M. S. Latha (2020). Smart medicine pill box reminder with voice and display for emergency patients. *Materials Today: Proceedings*, vol. 33, pp. 4876–4879. https://doi.org/10.1016/j.matpr.2020.08.400

[12] D. S. A. Minaam and M. Abd-ELfattah (2018). Smart drugs: improving healthcare using smart pill box for medicine reminder and monitoring system. *Future Computing and Informatics Journal*, vol. 3, pp. 443–456. https://doi.org/10.1016/j.fcij.2018.11.008

[13] P. N. J. Najeeb, A. Rimna, K. P. Safa, M. Silvana, and T. K. Adarsh (2018). Pill care-the smart pill box with remind, authenticate and confirmation function. *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, IEEE, pp 1–5. https://doi.org/10.1109/icetietr.2018.8529030

[14] C.-C. Hsu, T.-L. Chen, I.-F. Chang, Z.-Y. Wu, and C.-H. Liu (2020). Design and implementation a smart pillbox. *International Conference on 5G for Future Wireless Networks*, Springer, pp. 427–432. https://doi.org/10.1007/978-3-030-63941-9

[15] R. O. D. R. Carlos (2020). IoT-based smart medicine dispenser to control and supervise medication intake. *Intelligent Environments 2020: Workshop Proceedings of the 16th International Conference on Intelligent Environments*, IOS Press, pp. 39–48. https://doi.org/10.1109/ie49459.2020.9154933

[16] D. Karagiannis and K. S. Nikita (2020). Design and development of a 3D printed IoT portable pillbox for continuous medication adherence. *2020 IEEE International Conference on Smart Internet of Things (SmartIoT)*, IEEE, pp. 352–353. https://doi.org/10.1109/smartiot49966.2020.00066

[17] S. Jayanthi, S. Sindhuja et al. (2020). Smart pill dispenser. *Journal of Critical Reviews*, vol. 7, pp. 1481–1484.

[18] J. M. Parra, W. Valdez, A. Guevara, P. Cedillo, and J. Ortiz-Segarra (2017). Intelligent pillbox: automatic and programmable assistive technology device. *13th IASTED International Conference on Biomedical Engineering (BioMed)*, IEEE, pp. 74–81. https://doi.org/10.2316/p.2017.852-051

[19] P. Wadibhasme, A. Amin, P. Choudhary, and P. Saindane (2020). Saathi—a smart IoT-based pill reminder for IVF patients. *International Conference on Information and Communication Technology for Intelligent Systems*, Springer, pp. 697–705. https://doi.org/10.1007/978-981-15-7062-9_70

[20] D. I. Velligan and S. H. Kamil (2014). Enhancing patient adherence: introducing smart pill devices. *Therapeutic delivery*, vol. 5, pp. 611–613. https://doi.org/10.4155/tde.14.33

[21] N. A. Chaudri (2014). Adherence to long-term therapies evidence for action. *Annals of Saudi Medicine*, vol. 24, p. 221–222. https://doi.org/10.5144/0256-4947.2004.221

[22] U. Singh, A. Sharad, and P. Kumar (2019). IoMT based pill dispensing system. *10th International*

Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1–5.

[23] S. Jaipriya, R. Aishwarya, N. B. Akash, and A. P. Jeyadevi (2019). An intelligent medical box remotely controlled by doctor. *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, pp. 565–569. https://doi.org/10.1109/iss1.2019.8907996

[24] K. Arora and S. K. Singh (2019). IOT based portable medical kit. *International Journal of Engineering and Advanced Technology, Special Issue*, vol. 8, pp. 42-46. https://doi.org/10.35940/ijeat.e1012.0785s319

[25] B. Ayshwarya and R. Velmurugan (2021). Intelligent and safe medication box in health IoT platform for medication monitoring system with timely remainders. *7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, pp. 1828–1831. https://doi.org/10.1109/icaccs51430.2021.9442017

[26] M. Sinha, L. Fukey, K. Balasubramanian, M. H. Hanafiah, P. Kunasekaran, and N. A. Ragavan (2021). Acceptance of consumer-oriented health information technologies (CHITs): integrating technology acceptance model with perceived risk. *Informatica*, vol. 45, no. 6, pp. 45–52. https://doi.org/10.31449/inf.v45i6.3484

[27] M. Rathi, S. Sahu, A. Goel, and P. Gupta (2022). Personalized health framework for visually impaired. *Informatica*, vol. 46, no. 1, pp. 77–86. https://doi.org/10.31449/inf.v46i1.2934

[28] M. Depolli, V. Avbelj, R. Trobec, J. M. Kališnik, T. Korošec, A. P. Susič, U. Stanič, and A. Semeja (2016). PCARD platform for mHealth monitoring. *Informatica*, vol. 40, pp. 117–123.

[29] S. Nasiri, F. Sadoughi, A. Dehnad, M. H. Tadayon, and H. Ahmadi (2021). Layered architecture for internet of things-based healthcare system: a systematic literature review. *Informatica*, vol. 45, no. 4, pp. 543–562. https://doi.org/10.31449/inf.v45i4.3601

Applications of the Insieme Platform: A Case Study

Primož Kocuvan*[1], Erik Dovgan[1], Simon Ražman[2], Devid Palčič[2], Matjaž Gams[1]
E-mail: primoz.kocuvan@ijs.si, erik.dovgan@ijs.si, simon.razman@robotina.si, devid.palcic@robotina.si, matjaz.gams@ijs.si
*Corresponding author

[1]Deparment of Intelligent Systems, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia

[2]Robotina d.o.o.
OIC-Hrpelje 38, 6240 Kozina, Slovenia

Keywords: EMH, electronic and mobile health, prostate problems, time series, HEMS, Android application, ASPO, HEP-Y

Received: June 27, 2022

The information society has significantly changed the field of medicine. Several decades ago when a person got sick, a doctor examined a patient and prescribed some medicine with a patient more or less unaware of the true nature of the problem. Even if explained, many patients did not understand much due to the lack of medical knowledge. Today, knowledge is widely accessible through the web and many patients try to self-diagnose or at least get another opinion using popular search engines and specialized web applications. However, many of them provide misinformation due to the lack of proper education of the provider or the lack of understanding of the user. To overcome these issues, we developed a verified medical platform (Insieme platform) that includes medical data, applications and services. This article provides a list of applications in the Insieme platform, their descriptions and how to use them.

Povzetek: V prispevku so opisane zdravstvene aplikacije na platformi Insieme. Aplikacije nudijo uporabnikom različne funkcije, od osnovnega informiranja do zaznavanja zdravstvenih težav. Poleg tega aplikacije pripomorejo tudi k boljši strukturi in razdelanosti platforme Insieme.

1 Introduction

The Insieme platform is being developed as part of the ISE-EMH project [1], with the collaboration of three Italian and three Slovenian partners. The platform is a continuation of the EkoSMART project [2], where one of the segments was electronic and mobile health. EkoSMART was a three-year project where we explored the potentials of eHealth systems. The Insieme platform is available in English, Slovenian and Italian. The backbone of the platform are services. A service consists of a name, a short description, and sections. We defined 10 sections, but not all sections need to be included in the service.

In this paper we describe a list of applications in the Insieme platform, their descriptions and how to use them. They all belong to the field of electronic and mobile health (EMH). The Insieme platform includes data on around 40 applications. We list representative 15 of them and describe four in detail. The most relevant Insieme applications are:

- Motiphy [3]
- HEMS-based elderly behavioral change [4]
- DaVinci [5]
- Platomics [6]
- Depression app [7]
- Narcissistic personality disorder app [8]
- Bigorexia or muscular dysmorphia quiz [9]
- Schizophrenia app [10]
- Lymphoma app [11]
- Thyroid Cancer app [12]
- ASPO app [13]
- HEP-Y app [14]
- Senior helper app [15]
- Nala-care app [16]
- Skin vision - skin cancer app [17]

The four applications described more thoroughly are the following. The first application, Senior helper, is intended for elderly people and their caregivers. It enables the elderly to prolong stay in their homes. The second application is HEMS (Home energy management system) smart house that includes embedded sensors in rooms of a house or a flat to help users and physicians to detect potential prostate problems. The third one is ASPO (Application for informing about sexually transmitted diseases). It is a questionnaire which assigns a weight to the answers to evaluate the probability that a person has a sexually transmitted disease. Finally, the fourth one is HEP-Y, the questionnaire for evaluating the probability that a person has contracted hepatitis.

2 Senior helper

By the latest statistics, every fifth resident of Slovenia is older than 65 years [18]. Here we are denoting every person older than 65 years as an elderly. By the year 2050 every third person will be older than 65 years on the global scale [19]. As a consequence, the number of health problems in the population will rise, while at the same time the number of medical doctors and nurses will probably not increase accordingly (but might even decrease). To tackle this issue, we need a demographic solution or an ICT (Information communication technology) solution. We have developed the Senior helper application within the Insieme project that addresses this problem. The platform builds upon the previous H2020 project INLIFE [20].

The application consists of two integral parts, one for each user role. The first role is the caretaker and the second is the senior (the elderly person).

2.1 The functions of the Senior role

The senior-specific part of the application has five main components: Alarms, Settings, Fall detection, SOS function and Contacts. While most of the functions are self-explanatory, the fall detection needs more clarification. Every smartphone has integrated a special chip (MEMS - micro electro mechanical system) to detect acceleration. We developed an algorithm for fall detection, which measures acceleration in all three axes. When the application detects a high acceleration in any axis and shortly afterwards none, a fall is reported, i.e., an SMS is automatically sent to the caretaker that the fall has occurred [21]. In case of a false-positive event, the senior has an option to hold the button, which sends an SMS to the caretaker that the fall has not occurred.

2.2 The functions of the Caretaker role

The caretaker has a comprehensive view of the data of a specific senior. This includes additional features implementing the inaccessibility function, pedometer, location, and alarms. It also has options to view a senior's daily activity history.

The main purpose of the inaccessibility function is to enable the caretaker to disable regular messages for a particular senior or to set the alarms. The caretaker can select one of three options for inaccessibility: none, partial or complete. If the caretaker selects no inaccessibility, the senior can fully use the application. If he/she selects partial inaccessibility, the senior cannot use all the functions, but can unlock the partial inaccessibility if he/she remembers how to. This can be done if the senior holds the Settings button for five seconds or more. If the caretaker selects complete inaccessibility, the senior cannot unlock the selected functions in any way. In general, the caretaker can disable any kind of activities for the senior if the senior does not know how to use these functions.

In the interactions with elderly and caretakers on relevant events, such as the workshop in which the application was presented to elderly, significant interest was expressed in using this application, even by the older persons with basic ICT knowledge. Figures 1–2 show views, i.e., screenshots of the application.

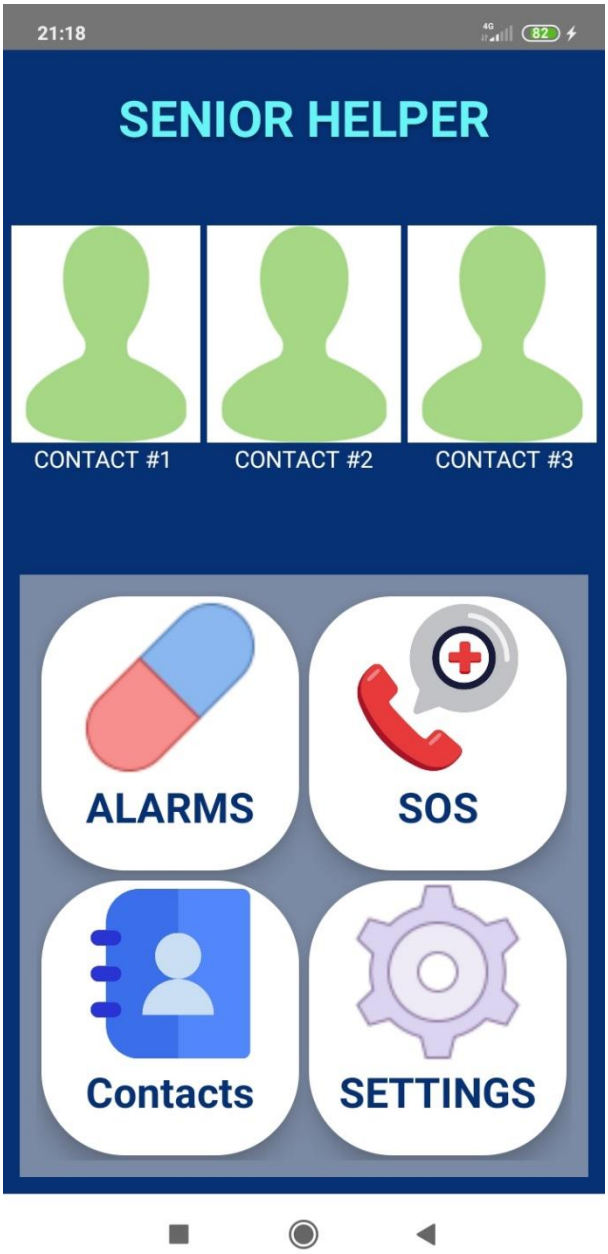


Figure 1: Basic view of the Senior role in the Senior Helper application.

3 Detecting prostate problems from the HEMS data

Medical issues with prostate affect men when they get older (in their last period of life). More than 50 % of men have an enlarged prostate when they are older than 60 years [22]. Enlarged prostate is not cancer and also, it does not lead to cancer. The exact causes of enlarged prostate are unknown. Also, some men have symptoms and some do not. Men who have these symptoms can either have an enlarged prostate or prostate cancer. For both diagnoses, the symptoms include difficulties urinating and frequent

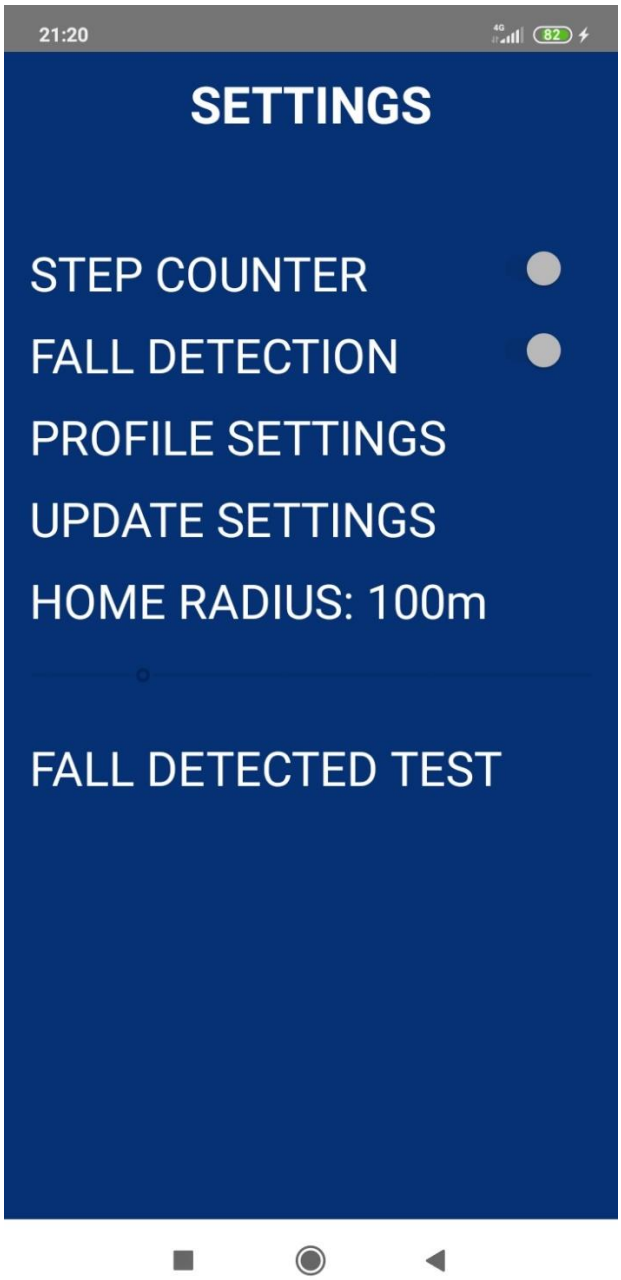


Figure 2: Settings view of the Caretaker role in the Senior Helper application. Home radius setting is intended as a geofencing option for people with dementia that are at risk of wandering off and getting lost. If the GPS detects the user outside the home radius, the application will send an alarm to the caregiver.

urinating (especially at nighttime). Additionally, blood in the urine or burning pain when urinating are the first signs of prostate cancer. This application aims at detecting such problems in the early stage thus preventing them from worsening the condition.

Several tests exist for diagnosing an enlarged prostate. If the patient's personal doctor cannot determine the cause of the problems, he/she can send the patient to the urologist for further examination and tests. In the following section, we will describe some typical tests urologist perform to either confirm enlarged prostate or prostate cancer.

Besides medical tests, ICT solutions for detecting prostate problems are also being developed. These solutions include, e.g., mobile applications or computer systems that assess the parameters from blood and urine results [23], [24], [25]. Such programs automatically assess the risk (i.e., multivariable risk calculator) for prostate cancer. Other systems assess non-medical data. For example, we have developed a system that asserts the risk of having a prostate problem based on HEMS data as described in Section 3.2.

3.1 Medical procedures for prostate problems

Several medical procedures for detecting prostate problems exist.

Urine test: For this test, a patient urinates into a cup. With a special piece of paper placed into the urine, the urologist can detect if a person has an infection or if there is some blood present in the urine. The blood in urine can be a cause of prostate cancer. Although other tests are needed, some latest research with a new type of biomarkers is promising [26].

Blood test: Through blood tests, specialists can detect abnormal levels of PSA (Prostate-Specific Antigen) [27]. PSA is a protein that is produced by the prostate. If a person has high levels of protein in the blood, this is a sign of prostate cancer or enlarged prostate. But typical medical specialists need to make other tests to confirm it.

Urodynamic tests: This is a group of tests that show how the person's urine is released from the bladder and how it is stored in the person's bladder. Urodynamics is the measurement of the relevant physiological parameters of the LUT to assess its (dys)function [28]. Some tests include flow measuring, where a person pees in a special container used to calculate the urine flow. Another option consists of a physician placing a thin tube into the urethra after the person pees and measuring how much urine is left in the bladder.

Transrectal ultrasound: Transrectal ultrasound was first developed in the 1970s. Transrectal ultrasound-guided biopsy, under local anesthetic and prophylactic antibiotics, is now the most widely accepted method to diagnose prostate cancer [29]. A technician places a transducer into the person's rectum. While he moves it around, it shows different parts of the bladder and prostate on the screen while the transducer emits ultrasound waves. The obtained images show if there is a tumor in the prostate, or if the prostate is bigger than normal size for a man that age.

Biopsy: The transrectal ultrasound-guided systemic biopsy is the recommended method in most cases with suspicion of prostate cancer. Transrectal periprostatic injection with a local anesthetic may be offered as effective analgesia [30]. For this test, the physician while taking an ultrasound or CT scan inserts a needle into the person's prostate and takes a sample of tissue for further laboratory exams. In the laboratory, a technician can see under the microscope if it is cancerous.

Do you have any problems when urinating?

- *No.*
- *Yes, burning urination.*
- *Yes, frequent urination.*
- *Yes, it often forces me to urinate.*

Figure 3: An example question from the questionnaire of the ASPO web application.

3.2 A HEMS approach to detect problems with prostate

An average healthy person who drinks about two liters of liquid (e.g., water) goes to the toilet about six to seven times per day (in 24 hours) [31]. Abnormal behavior starts when this number is above seven or eight. We developed a HEMS-based system for detection of prostate problems, which enables the caretakers to detect abnormal urinating behavior based on the electricity consumption in the bathroom, associated with the bathroom visits. This approach may be appropriate for older men who also have dementia or problems with memory and thus do not count/remember the number of toilet visits, and may serve as an early warning system for prostate problems.

4 ASPO

The Application for sexually transmitted infection risk assessment (ASPO) [13, 32] was designed as an informative questionnaire, in which the algorithm assigns weights to different answers from the user. Then the application sums up all the weights and returns an answer in a natural language. Basically, it provides the user an overall risk report. The strong point of the application is that it selects the following questions based on the answers from the previous questions. For example, if a person answered that he/she did not have any symptoms, the application will not ask a question regarding particular symptoms in the following. The application also consists of user stories (personal stories from users who also had such symptoms and problems). The application is available as a web app [13], and a person can access it through a mobile phone, computer or tablet. Figures 3–4 show an example question from the quiz and the possible answers.

5 HEP-Y

The liver is an organ that processes nutrients from food and drink which we consume. It also has a function to filter blood. Due to various reasons (e.g., alcohol abuse, medication, or infection by the hepatitis viruses) the liver may become inflamed. The main problem for patients is that the symptoms are not present at the beginning of inflammation, resulting in many patients only seeking medical assistance when the condition has already progressed very far or is even fatal, and thus treatment is difficult.

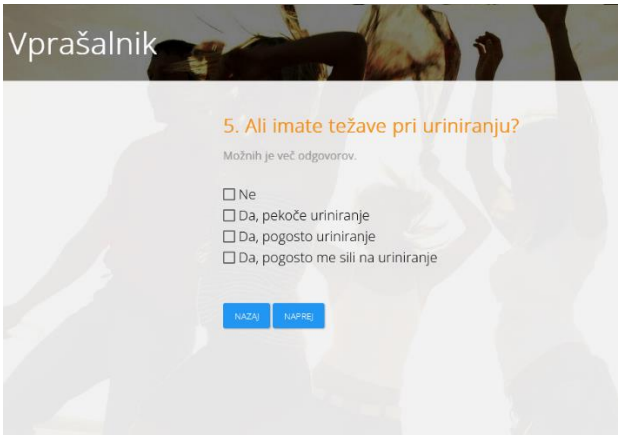


Figure 4: The user interface with an example question from the questionnaire of the ASPO web application (in Slovene).

To inform the general public about liver issues, the researches have created a platform called HEP-Y (Application for Viral Hepatitis Infection Risk Assessment) [14]. The platform is a continuation of the ASPO platform research [13]. As ASPO, HEP-Y also implements questionnaires. For this purpose, it uses a special data structure called a queue. Every question has an assigned id and an order property. When the user begins answering questions, the platform constructs a priority queue based on this order. Each question includes several (possible) answers. Every answer also includes a set of references to the questions that should not be provided to the user if he/she chooses this answer. For example, if the user selects an answer with reference to question no. 3, the system removes question no. 3 from the queue [33]. This approach allows the user to reach the risk assessment in the most straightforward and user-friendly way. A screenshot of the platform is shown in Figure 5.

6 Conclusion

We presented the most relevant applications available on the Insieme platform. This included the Android application for elderly, Senior Helper, which is fast, responsive, and easy to use. A HEMS-based approach for detecting problems with prostate was also presented. Finally, we described the ASPO and HEP-Y applications that implement questionnaires and share similar algorithm for generating questions. The ASPO web application estimates the risk of a person having a sexually transmitted disease, while HEP-Y does it for hepatitis. Overall, we believe that the Insieme platform has a potential for disseminating information about a large variety of health-related applications developed by several academic and private entities.

Acknowledgment

The paper was supported by the ISE-EMH project funded by the program Interreg V-A Italy-Slovenia 2014-2020. The authors also acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0209).

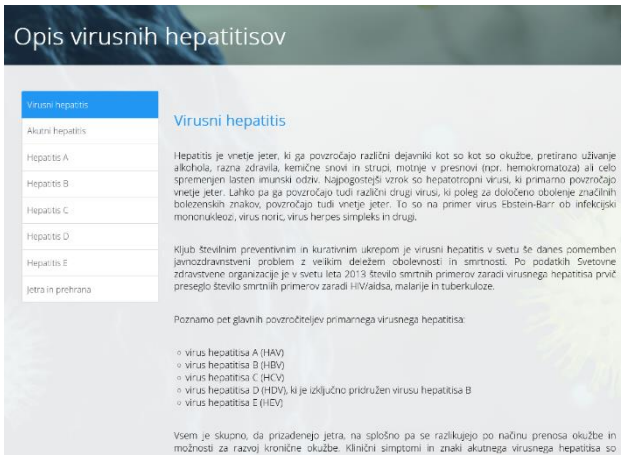


Figure 5: A screenshot of the HEP-Y platform (in Slovene), describing different types of hepatitis.

References

[1] ISE-EMH project, Insieme platform, https://ise-emh.eu/. Accessed July 11, 2022.

[2] EkoSMART, https://feri.um.si/raziskovanje/mednarodni-projekti-in-projekti-strukturnih-skladov/ekosistem-pametnega-mesta/. Accessed July 11, 2022.

[3] Clynx, Motiphy+ application, https://www.clynx.io/. Accessed May 10, 2022.

[4] Robotina, HEMS, https://mailchi.mp/robotina/ise-emh_elderly_hems_data_services. Accessed May 10, 2022.

[5] DaVinci Biolab app, DaVinci, https://davinci.biohub.solutions/. Accessed May 10, 2022.

[6] Platomics, Platomics application Biolab, https://davinci.biohub.solutions/platomics/. Accessed May 10, 2022.

[7] Depression application MoodTools, https://www.moodtools.org/. Accessed May 10, 2022.

[8] Narcistic personal disorder test, Do I have a personal disorder?, https://www.therecoveryvillage.com/mental-health/personality-disorders/do-i-have-a-personality-disorder/. Accessed May 10, 2022.

[9] Bigorexia, Body dysmphorphia quiz, https://eatingdisordersolutions.com/body-dysmorphic-eating-disorder-treatment/body-dysmorphia-quiz/. Accessed May 10, 2022.

[10] Schizophrenia self-testing app, https://www.mind-diagnostics.org/schizophrenia-test. Accessed May 10, 2022.

[11] Focus on Lymphoma web application, https://www.digiteum.com/portfolio/no-1-healthcare-app-for-lymphoma/. Accessed May 10, 2022.

[12] Clayman Thyroid Center, Thyroid cancer application, https://www.thyroidcancer.com/app. Accessed May 10, 2022.

[13] ASPO, https://aspo.mf.uni-lj.si/static/ASPO_new/#/. Accessed May 10, 2022.

[14] HEP-Y, https://hepy.mf.uni-lj.si/static/HEPY_new/#/opisi-virusnihhepatitisov/description-hb. Accessed May 10, 2022.

[15] Senior helper application for elderly, http://senior-helper-website.docker-e9.ijs.si/. Accessed May 10, 2022.

[16] Nala-care, Neuro-dermatitis app, https://nala.care/en/. Accessed May 10, 2022.

[17] Skin-vision, Skin cancer melanoma application, https://www.skinvision.com/. Accessed May 10, 2022.

[18] SURS, Število in sestava prebivalstva, https://www.stat.si/StatWeb/Field/Index/17/104. Accessed May 10, 2022.

[19] United nations, Our world is growing older, https://www.un.org/development/desa/en/news/population/our-world-is-growing-older.html. Accessed May 10, 2022.

[20] A. J. Astell, M. Panouand K. Touliou, et al. (2022). Developing a pragmatic evaluation of ICTs for older adults with cognitive impairment at scale: the IN LIFE experience. Univ Access Inf Soc, vol. 21, pp. 1–19, https://doi.org/10.1007/s10209-021-00849-5.

[21] J. Bizjak, A. Gradišek, L. Stepančič, H. Gjoreski, and M. Gams (2017). Intelligent assistant carer for active aging. *EURASIP J. Adv. Signal Process*, vol. 2017, no. 76, https://doi.org/10.1186/s13634-017-0511-y.

[22] R. I. Putu. (2020). *Integrated Scada-Tricoder: A Modern Hospital Room*. https://doi.org/10.13140/RG.2.2.24385.38247.

[23] N. M. Pereira-Azevedo and L. D. F. Venderbos (2018). eHealth and mHealth in prostate cancer detection and active surveillance. *Transl Androl Urol*, vol. 7, no. 1, pp. 170–181, https://doi.org/10.21037/tau.2017.12.22.

[24] N. Pereira-Azevedo, J. F. M. Verbeek, D. Nieboer, C. H. Bangma, and M. J. Roobol (2018). Head-to-head comparison of prostate cancer risk calculators predicting biopsy outcome. *Transl Androl Urol*, vol. 7, no. 1, pp. 18–26, https://doi.org/10.21037/tau.2017.12.21.

[25] N. Pereira-Azevedo, L. Osório, A. Fraga, and M. J. Roobol (2017). Rotterdam prostate cancer risk calculator: development and usability testing of the mobile phone app. *JMIR Cancer*, vol.3 no. 1, article no. e1, https://doi.org/10.2196/cancer.6750.

[26] ScienceDaily, University of East Anglia, Prostate cancer urine test identifies good prognosis patients, www.sciencedaily.com/releases/2021/11/211102210147.htm. Accessed May 10, 2022.

[27] D. Ilic, M. Djulbegovic, J. H. Jung, E. C. Hwang, Q. Zhou, A. Cleves, T. Agoritsas, and P. Dahm (2018). Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. *BMJ*, vol. 362, article no. k3519, https://doi.org/10.1136/bmj.k3519.

[28] S. M. Lenherr and J. Quentin Clemenm (2013). Urodynamics: with a focus on appropriate

indications. *Urol Clin North Am*, vol. 40, no. 4, pp. 545–557, https://doi.org/10.1016/j.ucl.2013.07.001.

[29] C. J. Harvey, J. Pilcher, J. Richenberg, U. Patel, and F. Frauscher (2012). Applications of transrectal ultrasound in prostate cancer. *Br J Radiol*, vol. 85, pp. S3–S17, https://doi.org/10.1259/bjr/56357549.

[30] S. F. Shariat and C. G. Roehrborn (2008). Using biopsy to detect prostate cancer. *Rev Urol*, vol. 10, no. 4, pp. 262–280.

[31] Frequency of urinating, https://www.bladderandbowel.org/bladder/bladder-conditions-and-symptoms/frequency/. Accessed May 10, 2022.

[32] A. Ajanović, J. Konda, G. Fele-Žorž, A. Gradišek, M. Gams, A. Peterlin, K. Počivavšek, M. Matičič (2017). Application for sexually transmitted infection risk assessment. *Informatica*, vol. 41, no. 2, pp. 253–254.

[33] A. Ajanović, A. Ulčar, A. Petelin, K. Počivavšek, G. Fele-Žorž, A. Gradišek, M. Gams, and M. Matičič (2018). Application for viral hepatitis infection risk assessment - HEPY. *Informatica*, vol. 42, no. 2, pp. 279–281.

How can Online Resources for Cancer Patients be Useful?

Chiara Cipolat Mis [1], Camilla Costa [2], Flavio Rizzolio [1,3] and Ivana Truccolo *[2]
E-mail: chiara.cipolatmis@gmail.com, kamy1769@gmail.com, flavio.rizzolio@unive.it, ivanatruccolo@gmail.com
* Corresponding author

[1] Scientific and Patient Library, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS
33081, Aviano, Italy

[2] Italian Association of Cancer Long-term Survivors (ANGOLO OdV)
33081, Aviano, Italy

[3] Department of Molecular Sciences and Nanosystems, Ca' Foscari University of Venice
30172, Venice, Italy

Keywords: online information, cancer patients, digital platform, patient education & empowerment, evaluation tools, forest bathing

Received: July 4, 2022

In the European Union, the necessity to help patients providing them with an equal, even if differentiated, kind of information about healthcare options is increasing. Patients have the same rights in any country to access prevention, diagnosis, and treatment of any type of disease. To achieve this end, the quality of information provided to the patients is crucial. Digital approaches aim at helping doctors and in general caregivers to reach all the patients to improve their empowerment, participation, and quality of life. In this paper, many questions are raised, and some solutions are provided including the INSIEME platform, where information about patient associations, companies, public and private services related to different types of pathologies in three languages are included.

Povzetek: Digitalne rešitve pomagajo zdravnikom in skrbnikom pri dostopu do pacientov z namenom izboljšanja kakovosti življenja pacientov. V članku je predstavljenih več vprašanj, ki se nanašajo na uporabo digitalnih tehnologij pri interakciji s pacienti. Poleg tega so predstavljene tudi izbrane rešitve, vključno s platformo INSIEME. INSIEME nudi podatke o združenjih bolnikov, podjetjih ter javnih in zasebnih službah, ki so povezani z različnim vrstami patologij. Vsi podatki so na voljo v treh jezikih.

1 Introduction

On 23 September 2020, during a virtual event organized by the European Cancer Organization (ECO), the European Code of Cancer Practice was launched. The Code is a citizen and patient-centered accessible statement of the core requirements for good clinical cancer practice, with ten key overarching rights of what a patient should expect from their healthcare system, supported by a plain language explanation [1].

Actually, the first Declaration on the Rights of Cancer Patients (Oslo) [2] dates 2002. Nevertheless, many principles are just a statement on paper and there is a long way to make them a reality.

The ten patients' rights of the 2020 Code are available in several European Community languages. Many of the points relate to the topic of patient information, in particular points 2 and 5, even if the information topic is crosscutting to all ten principles.

In detail, Principle 2 is about the Information and states: "You have a right to: Information about your disease and treatment from your medical team and other reliable sources, including patient and professional organizations. Patients should be informed that they can ask questions about the diagnosis, treatment, and the consequences of the disease and/or its treatment, as well as receiving information on nutrition, physical activity, psychological aspects, etc. The hospital should also refer the patient-to-patient organizations which can provide invaluable information and support at many levels" [3].

Furthermore, Principle 5 is about Shared Decision-Making and states: "You have a right to: Participate in Shared Decision-Making with your healthcare team about all aspects of your treatment and care. Increasingly in the era of patient-centered care, a shared or collaborative approach is being employed, in which a doctor recommends treatment but takes account of the patient's situation and views after careful discussion".

Also, the Picker Institute, a leading healthcare European charity researching patient and staff experience of care, has worked to promote the idea of person-centered care and defined the eight Picker Principles of Person-Centered Care, setting out a framework for understanding what matters most to most people, and what constitutes high-quality person-centered care [4]. One of the principles is about clear information and communication and enounce the importance for people to receive reliable, high quality and accessible information at every stage of the care process. The information should support people to make an informed decision and manage the care.

The rest of this paper is structured as follows. Section 2 presents an overview of the approach for designing IT healthcare solutions. The key questions that should be addressed when designing such solutions are discussed in Section 3. Finally, conclusions are given in Section 4.

2 Methods

Before designing an IT solution, it is necessary to analyze different key elements. Some of them are usually given for granted, e.g., the quality information evaluation and its impact on the solution. Therefore, the 5Ws for digital platforms have been considered when designing a platform for patients: why, what, who for, who with, where and how.

In the next section we will focus on the topic of health literacy and the evaluation of the information for lay people both written and online. We will also present the INSIEME platform, which contains the information that was chosen according to the considered key elements.

3 Key topics on information in e-health

There are several questions that need to be addressed when designing an IT healthcare solution. In the following sections we present the key questions and discuss about possible solutions.

3.1 Why do we talk about information in e-health?

The information, and communication as well, are part of the healthcare and the healthcare is becoming more and more digital. WHO defines eHealth as the cost-effective and secure use of information and communications technologies in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge, and research. Clear evidence exists on the growing impact that eHealth has on the delivery of healthcare around the world today, and how it is making health systems more efficient and more responsive to people's needs and expectations [5]. It is a fact that the Web is now part of patients' daily life, and/or that of their loved ones.

3.2 Why is it so important to be careful regarding online health information?

In the internet era, people have to disentangle themselves from the huge amount of information that they find on the Web. Information overload sometimes means making information useless, but a good information is necessary to empower patients and citizens in making decision about their health. Two important factors are to be considered related to quality information: **Health literacy** and **Evaluation tools of health information for people.**

3.3 What is health literacy?

Health literacy is the ability to access, understand, appraise, and use information to make healthy choices [6]. There are skills that make a person able to get the right information both on the prevention and health promotion, as well as on the treatment aspect. When a person has a good level of health literacy, he/she can better understand what is communicated to him/her by health services. Health literacy is a critical aspect for all the countries, but above all for the countries where the general level of literacy and numeracy competencies are low. This is because there is a correlation between low health literacy and low level of adherence to the care, and low consciousness regarding the prevention actions.

For cancer patients, literacy is even more important. As stated in the document *Health Literacy and Communication Strategies in Oncology: Proceedings of a Workshop* [7], health literacy is a critical skill for engaging in healthy behaviors to reduce disease risk and improve health outcomes across the continuum of cancer care. Low health literacy among patients with cancer is associated with poor health and treatment outcomes, including lower adherence to treatment, higher rates of missed appointments, and an increased risk of hospitalization. Low health literacy can also impede informed decision making.

Improvement of health literacy depends on the level of education, long life learning, and public policy on health literacy defined by the healthcare system/organizations. Health literacy on cancer information is challenging for frequent Internet users. As some authors state, health professionals, information specialists and librarians should orient people to reliable sources [8]. There are different tools in many languages for evaluating health literacy level of people related to different fields [9, 10].

3.4 Why and how should we evaluate the health information resources?

Evaluation of health information resources is a key process for increasing the information understanding by people. This task should be performed by healthcare organizations to guarantee the best quality of information to their users/patients. The Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, along with the IRCCS-AUSL from Reggio Emilia, Italy developed the ETHIC Evaluation Tool of Health Information for Consumers as an instrument to easily evaluate the formal aspects of the written and online information resources for patients [11]. This tool is now under a strict process of validation that includes different aspects of patient centered care, as part of a research project granted by the Italian Ministry of Health with the title "Changing the future: can we effectively improve patient education and its effectiveness in cancer care?". This project is the continuation of a very successful previous one about the power of patient education activities in empowering patients [12, 13].

Paying attention to the formal aspects of information resources [14] is one of the requirements for improving information literacy. It means evaluating the title and the authorship of the patient handouts, guides, booklets, webpages, the accuracy of date, the suitability of the

images to the text, the readability of tables and figures, the simplicity (but not banality) of the language, the transparency of possible sponsorship etc. Actually, plain language is writing designed to ensure the reader understands as quickly, easily, and completely as possible. Plain language strives to be easy to read, understand, and use [15].

ETHIC includes the utilization of an application to evaluate the Gulpease Index, a measure that calculates the readability of a text based on the length of words (measured in number of letters), the number of words, and the length and complexity of sentences.

To sum up, we can state that formal aspects of information is not only important when we are dealing with scientific publications [16], but also when we are dealing with information resources for lay people are not only formal but real indicators of quality information.

Furthermore, it is important to remember that, since 1998, the Health On the Net Foundation (HON), a no-profit organization of UNESCO, proposed the HONcode certification based on eight ethical principles and a verification process [17] to address the quality of the medical Internet. The HONCode certification is certainly one of the most successful initiatives in term of quality information warranty also because of its revisions [18].

Integrating the above tools in building a portal of online information for patients and citizens should be a must in the current Internet era.

3.5 Which are the patients' and citizens' frequently asked questions?

The e-health solutions should also consider the questions that people are asking on the Web about their health conditions. Some of the frequently asked questions are:

- Where can I find an inexpensive accommodation near the hospital where I have to go for treatment?
- What can I do to help my loved one not to give up...?
- What can I eat to help me get better?
- Where can I find someone who can tell me exactly what my rights are in relation to my work?
- Who can tell me whether this supplement is safe or there is a risk of interaction with my cancer treatment?
- Can I easily book online an exam?
- What physical activity is suitable to my condition?

Asking for the previous questions, a person can find many reliable heath resources on the Web. In addition, by talking with people who have experienced some treatments such as forest bathing, one can find information about this opportunity near him/her. Forest bathing is a practice, known in Japan as shinrin-yoku, consisting of slow walk into the forest, being calm and quiet amongst the trees, observing nature around you whilst breathing deeply. This practice can help both adults and children de-stress and boost health and wellbeing in a natural way. There is literature also related to cancer patients [19] and experiences in our interregional area as well [20].

There is a need to integrate in one place all these different resources and also to let people find them exactly when they need them.

There is a huge amount of associations and authoritative bodies offering online heath information such as websites, guides, FAQs, interactive booklets, forums, services, commercial sites, chats etc. But very often, people rely on the Google's page ranking and read the Google's knowledge panels and featured snippets [21]. However, the question is whether what is found online is really helpful to patients when they need help [22].

According to the 2022 Italian Censis Annual Report on Welfare and Health, 66.9 % of the respondents to their survey search autonomously on the Web about their own health condition, 41.6 % have a dialogue with their clinicians, 94.3 % still hope for a real patient centered care, and 93 % is expecting tailored care based on the patients' needs, focused on the continuity of care. Nevertheless, more than 92 % of the respondents rely on their doctors and healthcare professionals [23].

Therefore, the fragmentation of the resources that people can find on the Web, even if optimal resources, is a poor answer to the patients' needs.

3.6 What is really missing?

Currently, patients really miss an accessible digital web app ecosystem ensuring that they can be always up to date with their medical information but also find information about important aspects such as nutrition, sexual life, physical activity, communication with loved ones, rights etc. Furthermore, it would be important for a patient to see all his/her upcoming and past appointments in one place.

It is very useful also to have the option, in the same electronic place, to record his/her own symptoms or side effects at any time and share them with his/her treating team during next visit. Virtual and physical aspects can walk together with patients can feel more empowered when using patient-centered mobile applications, and mobile applications have potential for improving collaboration with healthcare professionals and care coordination [24].

There is a long way to go but this is the right direction. The INSIEME platform (https://ise-emh.eu/) developed by an interdisciplinary team within the ISE-EMH (Italian-Slovene ecosystem for electronic and mobile health) Italian-Slovene Interreg project [25], is a prototype having the potential to give a correct answer to the patient's needs. The platform aims to share good practices between the two countries to increase the mutual benefits. Other than information related to the primary care, the contents are enriched with secondary aspects that are important during and after the therapy. Even specific applications were developed to help people (e.g. elderly patients) during daylife. INSIEME includes the following categories: Hospital services, Social services, Physiotherapy services, Physical activities opportunities and related opportunities, Psychological services, Accommodation services for patients and/or family members, Information and counselling services about health topics and patients' rights, Administrative services, Local social and health services – screening, Local social and health services - palliative care, Voluntary associations, Independent information on cancer,

Independent Dr Information Desk and Information on fake news. Some aspects are peculiar such as the program Forest Bathing, and the module for appropriate communication with patients during the normal life based on cognitive science, which aims at preventing the development of pathologies.

The creators of the INSIEME platform were instructed about the above described issues and principles, and they acted to assure that the data provided on the platform are reliable. The information was collected by medical professionals and cancer information experts to guarantee that the information is good both about content and formal aspects.

4 Conclusion

Online health information is a resource only if it is of good quality and well-integrated in an ecosystem tailored on the patients' needs. Most of the time people use search engines or friends as a source of information to find a solution to their problems. In this context, there is an urgent need to provide free services for the patients where professionals take care both of content and formal aspects of information to bridge the gap between healthcare providers and citizens. The INSIEME platform could be a solution in which all the actors of the healthcare system (private and public) are included.

Acknowledgement

This paper was supported by the ISE-EMH project funded by the program Interreg V-A Italy-Slovenia 2014-2020.

References

[1] M. Lawler, K. Oliver, S. Gijssels, M. Aapro, A. Abolina, T. Albreht, S. Erdem, J. Geissler, J. Jassem, S. Karjalainen, C. La Vecchia, Y. Lievens, F. Meunier, M. Morrissey, P. Naredi, S. Oberst, P. Poortmans, R. Price, R. Sullivan, G. Velikova, E. Vrdoljak, N. Wilking, W. Yared, and P. Selby (2021). The European code of cancer practice. *Journal of Cancer Policy*, vol. 28, p. 100282, https://doi.org/10.1016/j.jcpo.2021.100282.

[2] Association of European Cancer Leagues (ECL), Declaration of Intent, https://www.cancer.eu/declaration-of-intent/. 2021, accessed June 07, 2022.

[3] European Cancer Organisation, European Code of Cancer Practice: Translations & Resources, https://www.europeancancer.org/2-standard/68-european-code-of-cancer-practice-translations-resources. Accessed June 09, 2022.

[4] Picker, The Picker Principles of Person-Centred care, https://picker.org/who-we-are/the-picker-principles-of-person-centred-care/. 2022, accessed June 09, 2022.

[5] World Health Organization - Regional Office for the Eastern Mediterranean, eHealth, http://www.emro.who.int/health-topics/ehealth/. Accessed June 09, 2022.

[6] I. Kickbusch, J. M. Pelikan, F. Apfel, and A. D. Tsouros (2013). *Health Literacy: The Solid Facts*. World Health Organization Regional Office for Europe.

[7] National Cancer Policy Forum, Roundtable on Health Literacy, Board on Health Care Services, Health and Medicine Division, and National Academies of Sciences, Engineering, and Medicine (2020). *Health Literacy and Communication Strategies in Oncology: Proceedings of a Workshop*. National Academies Press, https://doi.org/10.17226/25664.

[8] P. Serçekuş, H. Gencer, and S. Özkan (2020). Finding useful cancer information may reduce cancer information overload for Internet users. *Health Information and Libraries Journal*, vol. 37, no. 4, pp. 319–328, https://doi.org/10.1111/hir.12325.

[9] P. Zotti, S. Cocchi, J. Polesel, C. Cipolat Mis, D. Bragatto, S. Cavuto, A. Conficconi, C. Costanzo, M. De Giorgi, D. A. Drace, F. Fiorini, L. Gangeri, A. Lisi, R. Martino, P. Mosconi, A. Paradiso, V. Ravaioli, I. Truccolo, P. De Paoli, and ICPEG (2017). Cross-cultural validation of health literacy measurement tools in Italian oncology patients. *BMC Health Services Research*, vol. 17, no. 1, article no. 410, https://doi.org/10.1186/s12913-017-2359-0.

[10] V. Lastrucci, C. Lorini, S. Caini, Florence Health Literacy Research Group, and G. Bonaccorsi (2019). Health literacy as a mediator of the relationship between socioeconomic status and health: a cross-sectional study in a population-based sample in Florence. *PLOS ONE*, vol. 14, no. 12, article no. e0227007, https://doi.org/10.1371/journal.pone.0227007.

[11] S. Cocchi, M. Mazzocut, C. Cipolat Mis, I. Truccolo, E. Cervi, R. Iori, and D. Orlandini, ETHIC – Evaluation tool of health information for consumers. Development, features and validation, http://eprints.rclis.org/23241/. 2014, accessed January 17, 2021.

[12] I. Truccolo, C. Cipolat Mis, S. Cervo, L. Dal Maso, M. Bongiovanni, A. Bearz, I. Sartor, P. Baldo, E. Ferrarin, L. Fratino, M. Mascarin, M. Roncadin, M. A. Annunziata, B. Muzzatti, and P. De Paoli (2016). Patient-centered cancer care programs in Italy: benchmarking global patient education initiatives. *Journal of Cancer Education*, vol. 31, no. 2. pp. 405–412, https://doi.org/10.1007/s13187-015-0805-4.

[13] C. Cipolat Mis, I. Truccolo, V. Ravaioli, S. Cocchi, L. Gangeri, P. Mosconi, C. Drace, L. Pomicino, A. Paradiso, P. Paoli, and M. Apostolico (2015). Making patient centered care a reality: a survey of patient educational programs in Italian cancer

research and care institutes. *BMC Health Services Research*, vol. 15, no. 1, article no. 298, https://doi.org/10.1186/s12913-015-0962-5.

[14] A. Belkacem and Z. Houhamdi (2022). Formal approach to data accuracy evaluation. *Informatica*, vol. 46, no. 2, pp. 243–258, https://doi.org/10.31449/inf.v46i2.3027.

[15] Wikipedia, Plain language, https://en.wikipedia.org/w/index.php?title=Plain_language&oldid=1082781245. 2022, accessed June 10, 2022.

[16] O. Azeroual, M. J. Ershadi, A. Azizi, M. Banihashemi, and R. E. Abadi (2021). Data quality strategy selection in CRIS: using a hybrid method of SWOT and BWM. *Informatica*, vol. 45, no. 1, pp. 65–80, https://doi.org/10.31449/inf.v45i1.2995.

[17] HON Projects and Initiatives, Health On the Net Code of Conduct (HONcode), https://www.hon.ch/Project/HONcode.html. Accessed June 10, 2022.

[18] C. Boyer, V. Baujard, and A. Geissbuhler (2010). Evolution of health web certification through the HONcode experience. *Swiss Medical Informatics*, vol. 69, pp. 53–55, https://doi.org/10.4414/smi.26.00233.

[19] A. M. Ross and R. J. F. Jones (2022). Simulated forest immersion therapy: methods development. *International Journal of Environmental Research and Public Health*, vol. 19, no. 9, article no. 5373, https://doi.org/10.3390/ijerph19095373.

[20] Facebook, ForestBathing Potenziato FVG, https://www.facebook.com/forestbathingpotenziatoFVG. Accessed June 13, 2022.

[21] P. Lacey (2022). Google is goodish: an information literacy course designed to teach users why Google may not always be the best place to search for evidence. *Health Information and Libraries Journal*, vol. 39, no. 1, pp. 91–95, https://doi.org/10.1111/hir.12401.

[22] A. Scull (2020). Dr. Google will see you now: Google's health information previews and implications for consumer health. *Medical Reference Services Quarterly*, vol. 39, no. 2, pp. 165–173, https://doi.org/10.1080/02763869.2020.1726151.

[23] Censis and Janssen, Welfare e Salute. The Italian Health Day, https://www.censis.it/welfare-e-salute. 2022, accessed June 13, 2022.

[24] K. Mohsen, J. Kildea, S. D. Lambert, and A. M. Laizner (2021). Exploring cancer patients' perceptions of accessing and experience with using the educational material in the opal patient portal. *Supportive Care in Cancer*, vol. 29, no. 8, pp. 4365–4374, https://doi.org/10.1007/s00520-020-05900-4.

[25] ISE-EMH, Interreg Italia-Slovenia project, https://www.ita-slo.eu/en/ise-emh. 2020, accessed July 15, 2022.

A Novel Group Mobility Model for Software Defined Future Mobile Networks

A. Sureshkumar[1], D.Surendran[2]
[1]Assistant Professor, Department of Information Technology, Sri krishna College of Technology, Coimbatore, Tamilnadu, India.
E-mail: a.sureshkumar@skct.edu.in
[2]Professor, Department of Computer Science and Engineering, KPR Institute of Engineering and Technology. Coimbatore, Tamilnadu, India.

Keywords: Group mobility management (GMM), Software defined network (SDN), Future Mobile Networks, MoMo model, handover.

Received: May 11, 2021

Nowadays, a massive amount of data leads to cause network traffic and inflexible mobility in future mobile networks. A new Group Mobility Model (GMM) named MoMo is introduced that addresses the issue of the aforementioned problems. Even though, software defined network (SDN) is functional with network-rooted mobility protocols that enhance the network efficiency. Some existing network-rooted mobility administration methods still undergo handover delay, packet loss, and high signaling cost through handover processing. In this research work, SDN-based fast handover for GMM is proposed. Here, the neighbor number of evolving node transition probabilities of the mobile node (MN) and their obtainable resource probabilities are estimated. This makes a mathematical framework to decide the preeminent number of the evolving nodes and then allot these to mobile nodes virtually with all associations finished by the exploit of Open-Flow tables. The performance examination demonstrates that the proposed SDN rooted GMM technique has the enhanced performance than the conventional handover process and further technique by handover latency, signaling cost, network throughput, and packet loss.

Povzetek: Predstavljen je nov sistem MoMo kot model mobilnosti grup v modernih omrežjih.

1 Introduction

According to the statistics of mobile technology, an extensive increase has been made in the past 20 years and it is continued drastically in the near future [1]. Nowadays, the mobile communication network handles immense amount of data which leads to cause network traffic. Thus the network uses advanced technologies such as Artificial Intelligent, Internet of Things, and Software Defined Network (SDN), which eagerly support the network bandwidth [2]. In future mobile network, mobile video traffic is the main factor that should be compensated by virtual architecture. This infrastructure is based on Software Defined Network which provides flexible and on-demand service to the future 5G/6G mobile network. A software defined networking (SDN) is a programmable network architecture composed of three layers, namely, infrastructure, control, and application layer, respectively. Open flow is a bidirectional link which is used to direct the signalling message between the underlying network planes and SDN controller. Higher flexibility, better resource allocation, and improved performance are the potential benefits that should be governed by SDN [3]. In 5G/6G, the dependency of SDN's physical network is being reduced to generate high Quality of Service (QoS).

Ultra-dense network, a dense coverage model which supports high bandwidth in future mobile network (5G/6G) [4-6]. However, the dense network is not suitably able to perform the handover process using the conventional mobility model. Thus the network is required to change the mobility model to enhance the network performance. This effectiveness not only improves the handover process but also maintains high on-demand resource allocation and better QoS. To improve handover efficiency, mobility model selection plays an essential role in it. A Beam forming concept is a flexible operation meant to increase the available resources usage to sustain QoS [7].

However, the aforementioned technologies such as SDN and ultra-dense network perform data forwarding task with centralized mobility management (CMM), i.e., mobile anchor (MA) for home agent and local mobile anchor (LMA) for network routing respectively [8]. It is a central agent, that serves to sustain MN locations and redirect the traffic to them. CMM approaches handles some obstacles such as low scalability, high overhead, single node failure, etc. The following issue have been resolved using a new

paradigm named as Distribution Mobility Management (DMM) [9]. DMM is a distributed through the mobility agent to overcome the single node failure because if a mobility agent is failed to perform their job, then the other mobility agent within the network will take over the job of the failed node. This arrangement is being reduced the mobile data traffic by improving the handover delay, scalability, etc. More investigation was performed to confirm the SDN based DMM technology and it is noted in several literatures [10-12].

In a general viewpoint, the mobility model is categorized as individual and group mobility model [13]. These two agents are continuously making a handover service to mobile nodes (MN). Group mobility model (GMM) is a mobility pattern described to predict the movement of the mobile node in terms of continuous changeover time such as velocity, location, and acceleration [14]. In this model, each of the mobile nodes moves randomly together within the group using Random Waypoint approach [15] and Reference Point Group Mobility Pattern (RPGP) [16]. Both group mobility and individual mobility are used DynnaMo approach which provides a memory related model. The MoMo model combines memory-based individual model and flexible group model which increases the accuracy of the mobility model [17]. In this paper, we proposed software defined network based MoMo model for future mobile network to reduce the handover latency, signaling cost, and packet drop ratio.

The contribution of the proposed SDN based GMM work comprises of:

- The proposed model incorporates the mobility management module (MME) and admission control module. The MME includes the evolving node (EN) transition probability estimation and evolving node selection engines which are in the handover preparation phase.
- In the handover preparation phase, EN transition is computed for each neighbor node so that EN_ID is updated periodically. Therefore, the estimated EN_ID is transferred to the Open Flow table of the mobile node. If the current EN duration is expired, then automatically the MN node checks the optical flow (OF) table and chooses the next evolving node. The target EN influence OF table and it can find that the respective MN_ID is included in this table, then it will send handover acknowledgement to the MN. This concludes that both EN and MN nodes start preparing access to exchange their messages.
- In the handover phase, the grouping of each node is checked periodically based on two condition, namely, free state and forced state process.
- Finally, the handover performance such as handover latency, throughput, signaling cost, and packet loss are assessed for the proposed work which shows outperformed efficiency than the exiting GMM technique.

The association of this work is offered as pursues: Section 2 explains the literature review on recent SDN based mobility models. Section 3 describes the working of the proposed SDN based handover technique for future mobile networks. Section 4 portrays the simulation results and discussions. Lastly, section 5 ends the paper chased by the references.

2 Related work for research

Several existing works of mobility model with its issues, improvisation and challenges were discussed in this section.

Chung-Ming Huang et al. [18] presented a Bursty Multi-Node Handover using partially DMM (BMH–DMM) which uses a control scheme to tackle the handover problem. The proposed technique applies three procedures, namely, choosing MAAR candidate, handover preparation, and handover phase to simultaneously exchanging the query message to destination. Therefore, the approach uses a set of control schemes to minimize the handover delay and redundant message in Layers 2 and 3. In phase 1, Link_Going_Down (LGD) event is initiated from the set of mobile nodes (MN) which directs their message to MAAR1. After receiving the message, MAAR1 precisely needs to select proper destination among$MAAR_i$. Thus, MAAR1 derive score value for each $MAAR_i$ by using scoring function. After identified the score function from$MAAR_i$, MAAR1 establish a handover preparation to MAAR2. Finally, MAAR2 registers the CMD and handover the packet to the destination using a bidirectional tunnel. The benefit of this framework is to reduce the handover delay, packet loss, and signaling bandwidth but suffer from data integrity problem.

Luca Cominardi et al. [19] presented a more realistic network named as SDN with DNN technique. Packet delivery route is being optimized by reconfigured VLAN which shows a lack of scalability. Therefore, a new packet delivery procedure introduced by Sunghong Wei [20] processed through a soft anchor. It becomes more reasonable when combining the SDN technology with hybrid DMM method. Possibly, the approach of SDN based hybrid DMM (S-hDMM) is very effective to minimize high signaling cost. It performs two modes of procedure, namely, initial registration and handover. In registration procedure, the Optical Flow Switch (OFS) start receives the MN's message and forwards to S-hDMM. The controller creates the binding catch entries to update the MN's IP address to the OFS table. After the ended registration, the handover process occurs. Choosing the best anchor is an important factor which

reduces packet loss during training. OFS, a soft anchor is the best choice to reconfigure the packet delivery route for higher efficiency.

Battulga, Davaa, sambuu et al. [21] developed a distributed mobility model for selecting a mobility anchor (MA). The paper implemented some factors to select the mobility anchor based on cost function. Handover procedure provides the serving and target cells to transfer context information in which the handover commands return back to the subscriber at the end. Indispensible to note that the respective model is much effective to enhance the load balancing, packet delivery cost, and proper handover procedure but suffer from signaling performance.

Yong-hwan Kim et al. [22] presented a software-designed network (SDN) based on DMM for improving the LTE/EPC network performance. In conventional LTE/EPC networks, improper separation of data and control plane functionality may degrade the handover latency and radio resource allocation in Inter-technology. These issues are being resolved by the proposed architecture in the way of distributing the data plane through the gateway closer to the UEs, a centralized control plane virtually, and by clearly separating the control plane and data plane. The following procedure has to solve the latency problem, but the architecture becomes more complex than the other network because the system is virtually in-built. Another approach of the newly upgraded model named group mobility tolerates a valuable solution to the data exchanging mechanism. Cherry Ye Aung et al. [23] systematically reviewed a GMM by designing an accurate mobile ad hoc network (MANET). In this scenario, the author focuses on the categories of grouping model. Based on the movement of group members, the model is classified into four classes which independently control the location of an adjacent regions while moving. The review paper explains the movement of each group and their design features. Vehicular cognitive radio node is an application-oriented task utilized to make secure communication based on group mobility management (GMM) and is developed by Mani Shekhar Gupta et al. [24]. Potentially, the model chooses an improved network resource to achieve high throughput. Nevertheless, the network occupies congestion while performing mobility.

3 Proposed methodology

With the elevated challenging and the rising of mobile users (MU), the mobile network undergoes several issues such as handover (HO), data traffic, network routing, reliability, scalability, network signaling etc. To gather the inclination of increasing MUs, a novel group-based mobility management method is proposed to resolve the most challenging HO issue. Here, the particulars of the proposed SDN-based Fast HO control technique for GMM (SDN-GMM) are presented. A novel Group based mobility management (GMM) method, named MoMo is proposed by utilizing distributed mobility management (DMM). Different from the existing methods, the backward fast HO method is employed, which permits SDN to have an elevated possibility to end the HO training processing prior to detaching from the present sub-network, to diminish the packet loss rate and HO latency. The proposed novel GMM has three phases of operation namely Initial Phase, HO Preparation phase and HO Phase.

3.1 Initial phase

The MN is still associated with previous access points (AP) that detect the next AP. It established a signal strength which is advanced than the predefined threshold and is the uppermost one amid the sensed APs; then simultaneously the MNS threw the report message enclosing the next AP information to the initial one.

3.2 Handover preparation phase

This phase comprises of two modules such as the mobility management module and admission control module. This phase administers the dummy small cells and MNs in the data plane for mobility management. Furthermore, the mobility management entity (MME) is made to handle the HO procedure. In this way, the required mobility related data, for instance, the MN subscription data, mobile identification, and tackling trail area posts are attained. The mobile node id (MN_ID) and Evolving Node id (EN_ID) parameters are employed for MNs and small cells [25]. In the proposed SDN –GMM, MoMo model, the number of evolving nodes in the network contains a hexagonal architecture. This utilizes the automatic neighbor relation (ANR) function of evolving node count by which the neighbor relation tables are updated. Accordingly, the proposed SDN-GMM MoMo model reaches the valid neighbor relations of the evolving new nodes from these tables. Figure 1 demonstrates the proposed network architecture.

After calculating the transition possibilities for the neighbor evolving nodes, an obtainable resource probability of these neighbor evolving nodes is projected. According to the outcomes of this procedure, the next evolving node are estimated and allotted virtually to the MN before the movement. Thus, the predicted EN_ID is moved to the Open Flow (OF) table of the MN. Moreover, the time for this evolving node is computed and given to the value OF table of the MN. Every aforesaid method is performed for all evolving nodes situated on the MN movement path.

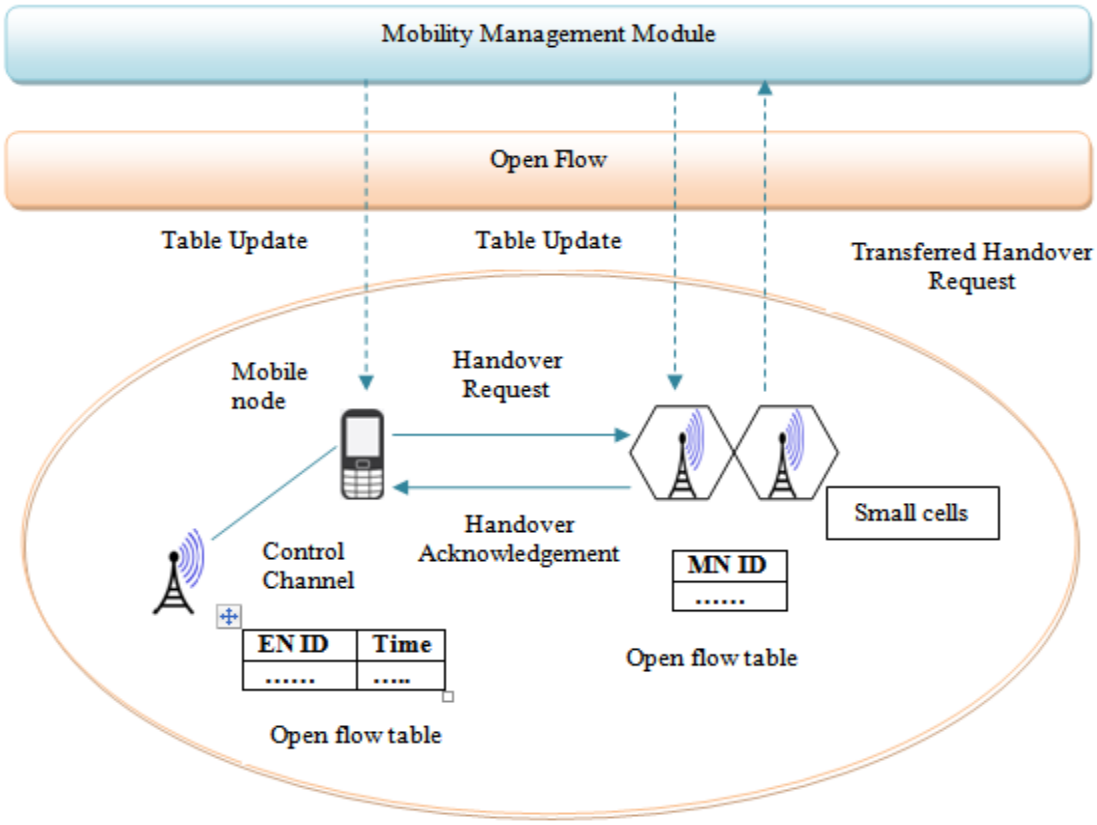


Figure 1: The proposed network architecture

If the time of the present evolving node closes, then MN views OF table to estimate the next evolving node. Accordingly, as revealed in Fig. 1, the MN propels a HO appeal to the identified target evolving node. The contact amid the MN and target evolving nodes is performed on the random-access channel (RACH). In future mobile networks, the RACH is utilized by MN to start the session with an arbitrary access preamble throughout the primary movement of the connect process. In addition, this preamble comprises the MN_ID. Next, goal evolving node manages OF table to discover this arriving MN_ID. If this MN_ID is integrated in the table, HO response is given to the MN. This response specifies that the HO appeal is established by the evolving node. Then the connect process continues amid the MN and evolving nodes
with the equivalent message series as downlink shared channel (DL-SCH), uplink shared channel (UL-SCH). Here, the delays observed in the HO training stage are investigated. If the MN_ID is not identified, this request is transferred to the controller. The controller updates the set OF tables accordingly.

3.3 Handover phase

The HO phase is followed by the HO preparation phase. This phase describes the connectivity steps of the nodes in the SDN based intelligent future network. The proposed SDN-GMM based MoMo model defines the binding conditions related tophysical proximity between the nodes. The binding condition between two nodes namely i and j are in same group, are referred to as group mates, and is defined as in (1)

$$d_{ij} \leq D_c \tag{1}$$

Where, $d_{ij} = d_{ji}$is the distance among nodes$i$and$j$. Ifthe binding condition in Eq. (1) is satisfied, the two nodes are said to be distance $D_c$connected. Consider a group of size M. For the generic node $j$the set of M_j^cgroup mates that the node detects as connected is called its connected set. The ratio between $M_j^c$and the total number M of group mates is called grouping factorρj in (2):

$$\rho j = \frac{M_j^c}{M-1} \tag{2}$$

The behavior of node j depends on the grouping condition (GC) defined on ρj in (3):

$$\rho j \geq \rho_{min} \quad (3)$$

where the grouping threshold ρ_{min} is a tunable model parameter. Every node occasionally verifies whether the GC is satisfied, with period ∇u based on the result, the node penetrates in anyone of the two subsequent states like free and forced:

- Free State occurs when the GC is satisfied. Here the node freely moves based on the boundless mobility model;
- Forced state occurs when the GC is not satisfied. Here the node travels towards the closest group mate, k, is not part of its connected set, to improve its grouping factor. The speed variables v and θ are set as in (4) and (5):

$$v = v_{max} \quad (4)$$

$$\theta = \begin{cases} min(\beta_{kj}, \theta_{old} + \gamma_{min} T_{lu}), if\, \beta_{kj} \geq \theta_{old} \\ max(\beta_{kj}, \theta_{old} - \gamma_{max} T_{lu}), otherwise \end{cases} \quad (5)$$

Where, $\beta_{kj} = arctan\left(\frac{y_k - y_j}{x_k - x_j}\right)$,$(x_k, y_k)$ and $(x_j, y_j)$are the locations of nodes k and j, respectively,θ_{old} is the previous direction, andT_{lu} is the time elapsed while the final position update1. Equations (4) and (5) ensure that node j attains the chosen group mate k in the shortest probable time frame, while evading although, destruction of the limitation on linear and angular speed.

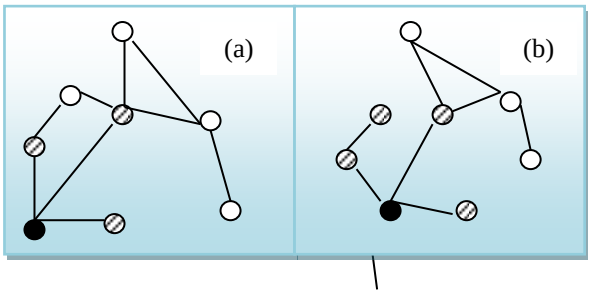


Figure 2: Example of the function of the SDN – GMM MoMo mobility model for a group of 8 nodes

In Fig 2 a), the measured node (black node) is linked to three striped nodes in its collection, out of its 7 group mates, and the grouping factor ρ= 3/7 calculated by the node is lesser than the necessaryρ_{min}. As a result, the node goes in forced mode and travels towards the nearby group mate among its connected set. The node preserves this behavior until the GCis is satisfied, evolving towards the situation in 2. b), where the range of the connected set is increased to 4, corresponding to$\rho \geq \rho_{min}$.

The behavior defined in the forced mode, that is, moving towards the closest group mate not in the connected set, is not the only possible one. More complex behaviors, e.g., moving towards the centroid of the points of group mates, can be easily established in the framework of the proposed SDN –GMM MoMo model. Figure 2demonstrate an instance of the application of the SDN – GMM MoMo model in the case of a group of 8 number of nodes with ρ_{min} = 0.5. Lines between nodes indicate connectivity for the SDN –GMM MoMo mobility model. In Figure 2a), the size 3 of the connected set for the black node leads to a grouping factor ρ= 0.43. The GC is thus not satisfied, and the node moves to the closest group mate not part of its connected set, until the condition is satisfied as in the configuration shown in Figure 2b, in which M_c = 4, ρ= 0.57).

The definition of connectivity, and the corresponding meaning of the threshold D_c , is a key feature in SDN – GMM MoMo. The replicate for a flexible definition of the theory of associated, rooted on the application scenario:

- connectivity related to radio communications - here two nodes are associated either during a straight radio connection (physical layer associativity), while they are in the radio range, or during conveying, assured by additional group mates (network associativity);
- connectivity is rooted on a radio-independent constraint - for instance, if a collection will become a security team, material visibility may correspond to association: a team member will go freely until it is capable to view a minimum count of team members, and travels nearer to the extra members of the team when the circumstance is not met any longer.

4 Simulation results and discussions

In this section, the HO presentation is evaluated with the proposed SDN- GMM model and estimates performance metrics including signaling cost, average HO latency, and packet loss rate. The simulation result of the proposed method is estimated by using NS-3 network simulator version 3.26. It is a discrete-event network simulator using the Open Flow module. By using it, the HO process working is verified. The simulation environment comprises of four sorts of network essentials with 50 MUs, 10 routers that manage 10 APs, 10 corresponding nodes, and 802.11n infrastructure. The simulation is continuously performed for 30 times to discover the regular outcomes.

4.1 Signaling cost

The signaling cost is described as the HO mobility binding update overhead acquired throughout the HO processing. Hence, the signaling message delivery cost of the mobility management protocol is reliant on the result of the count of network hops, the dimension of a signaling message, and the weighting aspect in a wired and wireless network. In the proposed SDN-GMM method, the worldwide assessment allows SDN Controller to gain the agent router and manage signaling messages in MNs' HO processing. The MME and the MODULE are termed to hold mobility administration and handle MNs' registrations for having the GMM service. Thus, the signaling cost of the SDN-GMM technique is articulated as in (6)

$$cost = \mu \begin{bmatrix} hL_{inform} hop_{MN-router} \\ +\rho L_{hi} hop_{router-router} \\ +\rho L_{hack} hop_{router-router} \end{bmatrix} \quad (6)$$

Where h and ρ are wired and wireless link's weight factor, μ is the arrival rate of each MN via router, Linform refers to the dimension of handover message, Lhi and Lhacks indicates the dimension of HI and Hack message respectively, hop specifies the connection between two nodes.

An advantage of the proposed SDN-GMM process is the decrease of the control packets for MN's and over processing. Figures 3 and 4portray the difference of the signaling cost rooted on diverse MN's velocities and different counts of hop counts. When the MN velocity is improved from 5 to 30 m/s and the count of hop count is augmented from 1 to 10. The count, MN velocity, and radius of cells of necessary binding update messages by the proposed SDN-GMM method is less than that of the traditional approach such as CMM [26], S-hDMM [20] and GMM [27]. The proposed technique achieves 187, 243, 300, 357, 419, and 498 signaling costs per packet versus m n's velocity from 5 to 30 m/s. Similarly, considering the signaling cost versus hop count and it shows that the proposed method attains 250, 272, 295, 309, 332, 340, 375, 391, 427, and 439signalling cost per packet from 1 to 10 hop counts respectively.

The result of radius of the cell is pursued: the larger radius indicates the lower HO occurrence rate. Thus, the necessary binding update message decrease when the cell's radius enhances. Although the cell's radius is improved from 100 to 350 m, the count of signaling messages of the proposed SDN-GMM technique is also lower than the existing techniques. In this analysis result, the proposed SDN-GMM model attains an improved output of 990, 442, 384, 216, 206, and 185signalling cost efficiency for 100, 150, 200, 250, 300 and 350 radius of cell in meters respectively. The comparison graph of handover signaling loss versus radius of cell using different techniques is revealed in Figure 5.

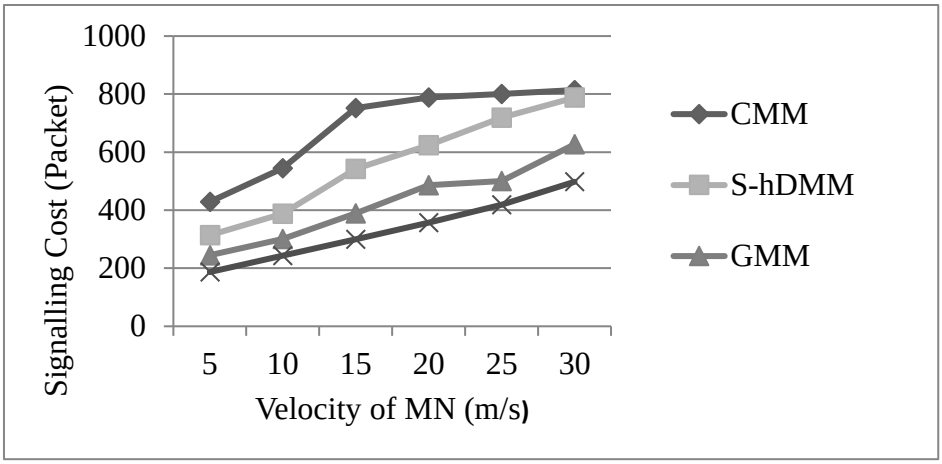


Figure 3: Handover signaling cost versus velocity of MNs

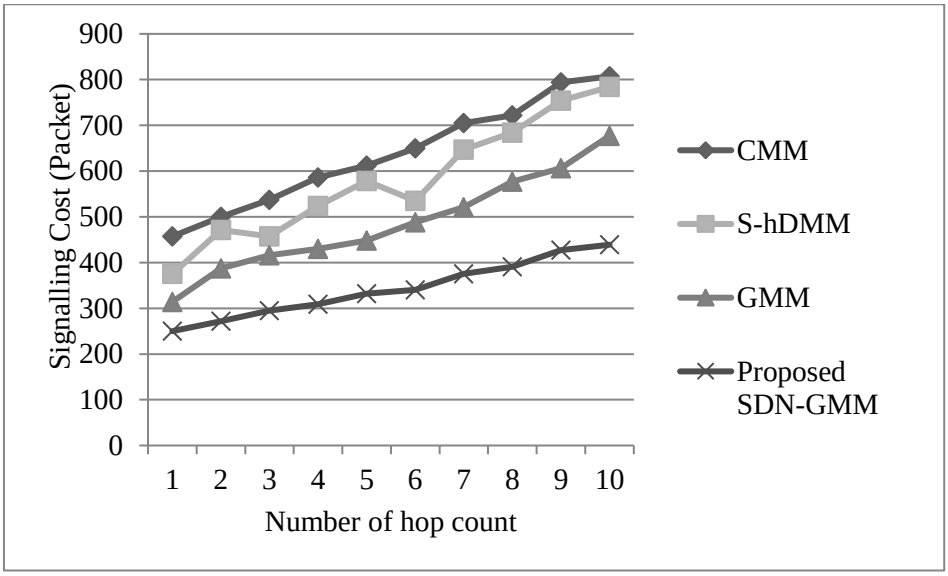


Figure 4: Handover signaling cost versus number of hop count

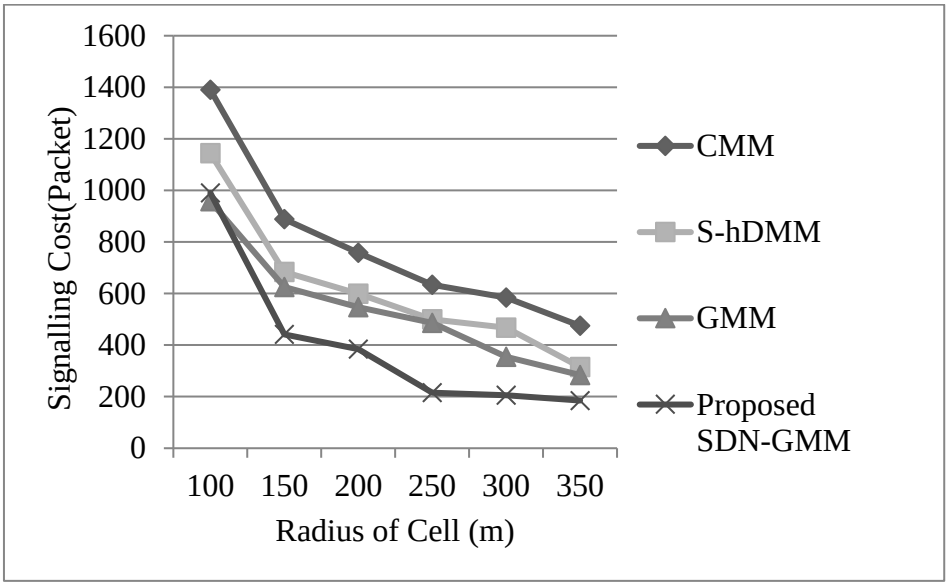


Figure 5: Handover signaling cost versus radius of cells

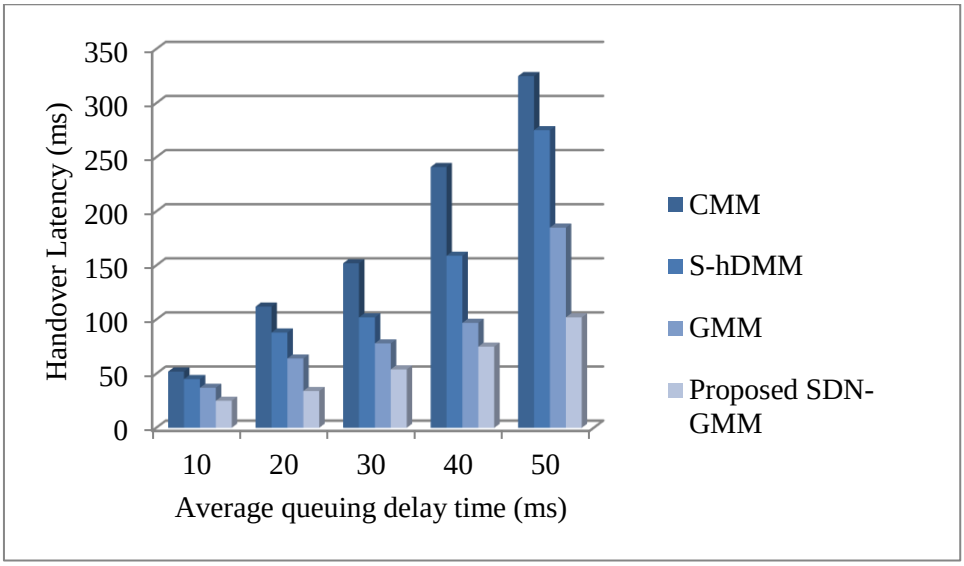


Figure 6: Handover delay versus average queuing delay

4.2 Handover latency

The further metric for mobility management protocols is HO latency. The HO latency is described as the delay from the time when MN begins the HO process to the time when it accepts the first packet from its novel mobility router. The delay time is required to drive a packet over wired and wireless links and it contains propagation time, transmission delay, and queuing delay. The HO delay relies on the hop-count distance from the source to the destination. It is tacit that the wired links are steady and reliable. Then,

the delay time of a packet of size p sent from u to v is represented as in (7):

$$Dt_{m,n} = hopcount * \left(\frac{S_{msg}}{BW_{bnd}} + L_{pl} + W_{aq}\right) \quad (7)$$

WhereS_{msg}, L_{pl}, hop count,BW_{bnd} and$W_{aq}$are the average manage message volume, propagation latency, hop distance connectivity, the accessible bandwidth, and the average queuing delay at every router in wired links.

From Fig. 6, the proposed SDN-GMM method achieves 25, 34,54, 75, and 102 handover latency (ms) for average queuing delay time 10ms, 20ms, 30ms, 40ms and 50ms respectively. It is clear that the average queuing delay at each router is amplified through the HO latency of the proposed SDN-GMM method which is lower than that of the traditional handover techniques.

4.3 Throughput

Throughput ($thrt$) is described as the whole volume of transmitted data packets in a session, which is$S_{ad} * L_u$over the session delivery time and is formulated in (8).

$$thrt_{sdn-gmm} = S_{ad} * L_u / \left[\left(PDT_{sdn-gmm} + (S_{ad} - 1) * \rho\right) + T_{HO-sdn-gmm} * t_{HO-no}\right] \quad (8)$$

Where,S_{ad} is the average count of transmitted packets in a session, $L_u$is the average packet size from a corresponding node to MNs, and $\rho$is the packet-to-packet gap time. $PDT_{sdn-gmm}$is packet delivery time for transmitting L_u packets. PDT is the packet delivery time from the corresponding node to MNs, which is calculated as the total delay time of transmitted packets given in formula (8). $T_{HO-sdn-gmm}$is calculated as the ratio among the HO μ and the average session coming charge φ.

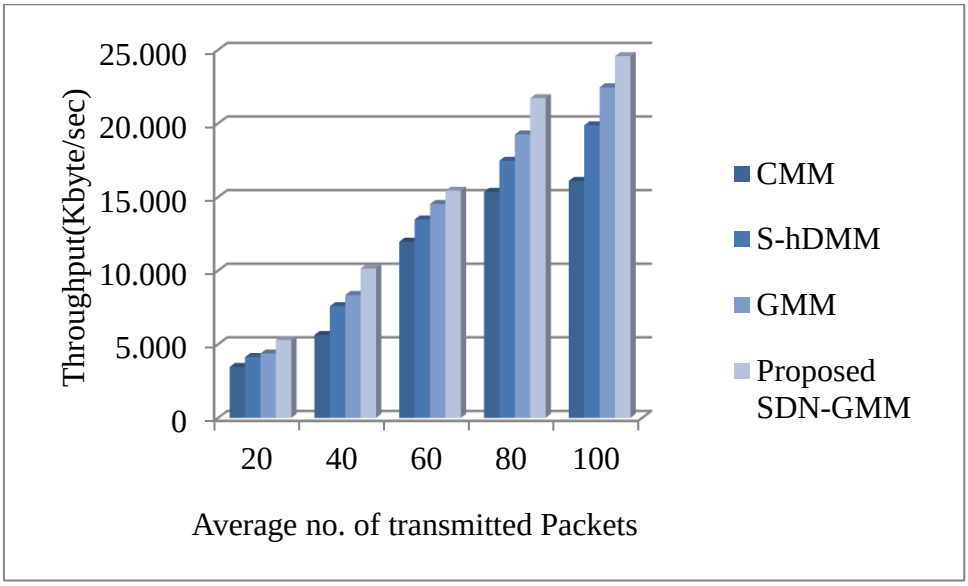


Figure 7: Throughput versus count of packets in a session

Figure 7 shows the comparison of throughput performance between the proposed SDN-GMM technique and the existing techniques rooted on the count of packets. It shows the throughput of the proposed method is advanced than the existing methods when the average count of packets is improved from 20 to 100 packets. It shows that the proposed technique has high throughput and is listed as 5292, 10147, 15427, 21719, 24572 Kbyte/sec of throughput accomplished for 20, 40, 60, 80 and 100 number of transmitted packets respectively.

5.4 Packet loss

When MN performs handover, it observes whether the packet loss can affect the HO procedure. MN's packet

loss rate is proportional to either HO delay or session arrival rate φ in unit of packet per second. Let Sad be the average count of delivered data packets during a session. Then, the packet loss rates of the use of the proposed SDN-GMM HO method are calculated in (9) as:

$$C_{sdn-gmm} = \varphi * S_{ad} * T_{HO-delay-sdn-gmm} \quad (9)$$

Whereas$T_{HO-delay-sdn-gmm}$denote the average HO delay of the proposed SDN-GMM method.

Figure 8 depicts the packet loss situations of using the proposed SDN-GMM method and the traditional methods in terms of HO rates. Referring to Fig. 8, when the HO rate is improved from 0.12 to 0.28, the count of missing packets by the proposed method is partly of that of that by the existing processes. It is experimentally proved that the total of lost packets by the proposed technique is smaller than that of the traditional techniques when the HO rate is improved. It is for the reason that the proposed SDN-GMM technique has the inferior HO latency time than the existing HO techniques and the momentary pre-established tunnel, which can be detached when all on-the-fly packets have been promoted to MN.

The proposed technique achieves 87, 91, 106, 175, and 192 packet loss amounts which are compared against the mobility rate of 0.12, 0.16, 0.20, 0.24, and 0.28 respectively.

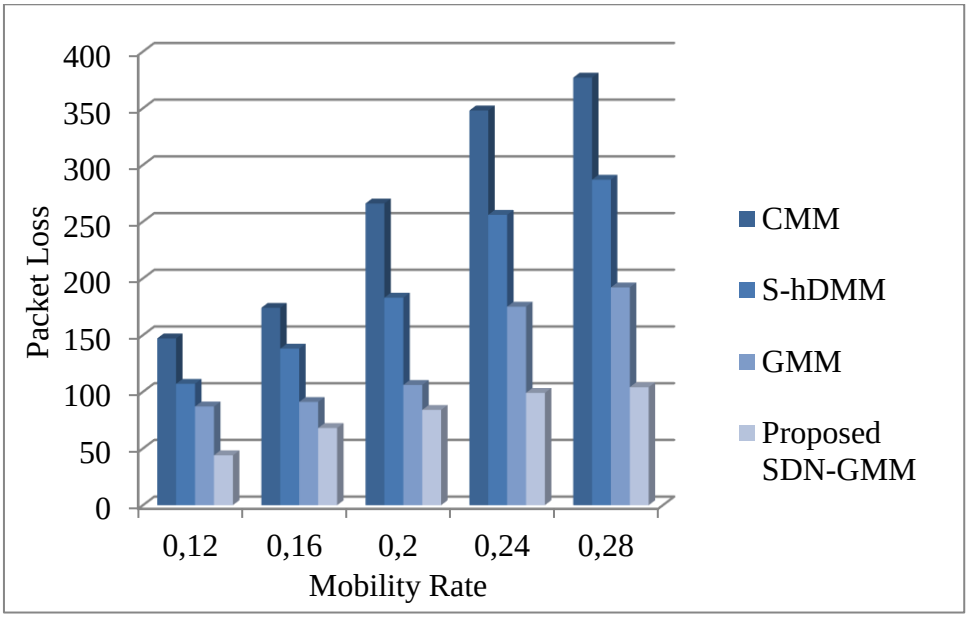


Figure 8: Packet loss versus mobility rat

5 Conclusion

In this work, an improved handover organized method for SDN -based fast handover for Group Mobility Model is proposed. The major contribution of the proposed SDN-GMM method are to allow MN be proficient to earlier finish the HO training processing and to attach to a new mobile router before MN disconnecting to the preceding mobile router. The packet loss of the proposed SDN-GMM method is condensed as of less HO latency than the existing GMM method and the handling of the temporarily established bidirectional tunnel, which can be detached when all on-the-fly packets have been forwarded. To prove the proposed method, it is compared with the traditional handover techniques over the NS-3 simulation environment. In future, the proposed SDN-GMM method can be practical in the vehicular setting.

Funding: There is no funding source.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

[1] Mansoor Shafi, Andreas F. Molisch, Peter J. Smith, Thomas Haustein, Peiying Zhu, Prasan De Silva, Fredrik Tufvesson, Anass Benjebbour, and Gerhard Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice", *IEEE journal on selected areas in communications*, vol.no.35, no. 6, pp. 1201-1221, 2017

[2] Catherine NayerTadros, Mohamed RM Rizk, and Bassem Mahmoud Mokhtar. "Software defined network-based management for enhanced 5G network services." *IEEE Access* vol.no.8, pp. 53997-54008, 2020

[3] Wenfeng Xia, Yonggang Wen, ChuanHengFoh, DusitNiyato, and HaiyongXie, "A survey on software-defined networking", *IEEE Communications Surveys & Tutorials*,vol.no.17, no. 1, pp. 27-51, 2014.

[4] Peng Hao, Xiao Yan, Yu-NgokRuyue, and Yifei Yuan. "Ultra dense network: Challenges enabling technologies and new trends." *China Communications*, vol.no.13, no. 2, pp. 30-40, 2016

[5] Zhonglin Chen, Shanzhi Chen, Hui Xu, and Bo Hu. "A security authentication scheme of 5G ultra-dense network based on block chain", *IEEE Access*, vol.no.6, pp. 55372-55379, 2018

[6] Xiaohu Ge, Song Tu, Guoqiang Mao, Cheng-Xiang Wang, and Tao Han. "5G ultra-dense cellular networks", IEEE *Wireless Communications*, vol.no.23, no. 1, pp. 72-79, 2016

[7] SteliosTimotheou, IoannisKrikidis, Gan Zheng, and Bjorn Ottersten, "Beamforming for MISO interference channels with QoS and RF energy transfer", *IEEE Transactions on Wireless Communications*, vol.no.13, no. 5, pp. 2646-2658, 2014

[8] Javier Carmona-Murillo, Ignacio Soto, Francisco J. Rodríguez-Pérez, David Cortés-Polo, and José Luis González-Sánchez, "Performance evaluation of distributed mobility management protocols: Limitations and solutions for future mobile networks." *Mobile Information Systems*, 2017.

[9] Yarisley Peña Llerena, Paulo RL Gondim, and Jaime Lloret, "Improving throughput in DMM with mobile assisted flow mobility," *Transactions on Emerging Telecommunications Technologies*, vol.no.29, no. 3, e3257, 2018.

[10] Michelle Perras, Juan Carlos Zuniga, Alexander Reznik, Carlos Jesus Bernardos, and Hao Jin, "Software defined networking distributed and dynamic mobility management" *U.S. Patent No. 9,998,967*, issued June 12, 2018.

[11] Jeong-A. Kim, Do Gun Park, and JongpilJeong, "Design and performance evaluation of cost-effective function-distributed mobility management scheme for software-defined smart factory networking", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-17, 2019

[12] Chung-Ming Huang, Duy-Tuan Dao, and Meng-Shu Chiang, "SDN-FHOR-DMM: a software defined network (SDN)-based fast handover with the optimal routing control method for distributed mobility management (DMM)", *Telecommunication Systems*, vol.no.72, no. 2, pp. 157-177, 2019

[13] Tracy Camp, Jeff Boleng, and Vanessa Davies. "A survey of mobility models for ad hoc network research." *Wireless communications and mobile computing*, vol.no.2, no. 5, pp. 483-502, 2002

[14] Subodh Kumar, G. S. Agrawal, and Sudhir Kumar Sharma, "Impact of Mobility Models on MANETs Routing Protocols", *INROADS-An International Journal of Jaipur National University*, vol.no.3, no. 1, pp. 142-147, 2014

[15] Subodh Kumar, G. S. Agrawal, and Sudhir Kumar Sharma, "Impact of Mobility on MANETs Routing Protocols Using Group Mobility Model", *International Journal of Wireless and Microwave Technologies*,vol.no.7, no. 2, pp. 1-12, 2017

[16] Dayan AdionelGuimarães, EdielsonPrevatoFrigieri, and Lucas Jun Sakai. "Influence of node mobility, recharge, and path loss on the optimized lifetime of wireless rechargeable sensor networks." *Ad Hoc Networks*, vol.no.97, pp. 102025, 2020

[17] Luca De Nardis, and Maria-Gabriella Di Benedetto, "MoMo: a group mobility model for future generation mobile wireless networks", *arXiv preprint*, 1704.03065, 2017.

[18] Chung-Ming Huang, Duy-Tuan Dao, and Meng-Shu Chiang, "A Bursty Multi-node Handover scheme for mobile internet using the partially Distributed Mobility Management (BMH–DMM) architecture", *Telecommunication Systems*, vol.no. 69, no. 1, pp. 113-130, 2018

[19] Luca Cominardi, Fabio Giust, Carlos J. Bernardos, and Antonio De La Oliva, "Distributed mobility management solutions for next mobile network architectures", *Computer Networks*, vol.no.121, pp. 124-136, 2017

[20] SunghongWie, "SDN-based Hybrid Distributed Mobility Management", *Journal of information and communication convergence engineering*, vol.no .17, no. 2, pp. 97-104, 2019.

[21] BattulgaDavaasambuu, TumneeTelmuun, Dominik Sasko, Yu Keping, and ShirmenSodbileg, "A Novel Anchor Selection Scheme for Distributed Mobility Management", *Computer Science*, vol.no. 22, no. 1, 2021.

[22] Yong-hwan Kim, Hyun-kyo Lim, Kyoung-han Kim, and Youn-Hee Han, "A SDN-based distributed mobility management in LTE/EPC network", *The Journal of Supercomputing*, vol.no. 73, no. 7, pp. 2919-2933, 2017

[23] Cherry Ye Aung, Boon Chong Seet, Mingyang Zhang, Ling Fu Xie, and Peter Han Joo Chong, "A review of group mobility models for mobile ad hoc networks", *Wireless Personal Communications*, vol.no.85, no. 3, pp. 1317-1331, 2015.

[24] Mani Shekhar Gupta, and KrishanKumar,"Application aware networks' resource selection decision making technique using group mobility in vehicular cognitive radio networks", *Vehicular Communications*, vol.no.26, 100263, 2020

[25] TugceBilen, BerkCanberk, and Kaushik R. Chowdhury, "Handover management in software-defined ultra-dense 5G networks", *IEEE Network*, vol.no.31, no. 4, pp. 49-55, 2017.

[26] Mohammed Balfaqih, Zain Balfaqih, Vladimir Shepelev, Soltan A. Alharbi, and Waheb A.

Jabbar, "An analytical framework for distributed and centralized mobility management protocols", *Journal of Ambient Intelligence and Humanized Computing*, pp.1-13, 2020.

[27] Subodh Kumar, G. S. Agrawal, and Sudhir Kumar Sharma, "Impact of Mobility on MANETs Routing Protocols Using Group Mobility Model", *International Journal of Wireless and Microwave Technologies*, vol. no. 7, no. 2, pp. 1-12, 2017.

EEG Signal Feature Extraction and Classification for Epilepsy Detection

Cherifi Dalila [1], Falkoun Noussaiba [1], Ouakouak Ferial [1], Boubchir Larbi [2], Nait-Ali Amine [3]
[1] Institute of Electrical and Electronic Engineering, University of Boumerdes, Algeria.
[2] LIASD Laboratory, University of Paris 8, France.
[3] LISSI Laboratory, University of Paris-Est Créteil, France.
Email: da.cherifi@univ-boumerdes.dz, noussaibafalkoun@gmail.com

Keywords: EEG, Epilepsy, Feature Extraction, DCT, DWT, Classification, K-NN, SVM, ANN.

Received: September 09, 2021

Epilepsy is a neurological disorder of the central nervous system, characterized by sudden seizures caused by abnormal electrical discharges in the brain. Electroencephalogram (EEG) is the most common technique used for Epilepsy diagnosis. Generally, it is done by the manual inspection of the EEG recordings of active seizure periods (ictal). Several techniques have been proposed throughout the years to automate this process. In this study, we have developed three different approaches to extract features from the filtered EEG signals. The first approach was to extract eight statistical features directly from the time-domain signal. In the second approach, we have used only the frequency domain information by applying the Discrete Cosine Transform (DCT) to the EEG signals then extracting two statistical features from the lower coefficients. In the last approach, we have used a tool that combines both time and frequency domain information, which is the Discrete Wavelet Transform (DWT). Six different wavelet families have been tested with their different orders resulting in 37 wavelets. The first three decomposition levels were tested with every wavelet. Instead of feeding the coefficients directly to the classifier, we summarized them in 16 statistical features. The extracted features are then fed to three different classifiers k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), and Artificial Neural Network (ANN) to perform two binary classification scenarios: healthy versus epileptic (mainly from interictal activity), and seizure-free versus ictal. We have used a benchmark database, the Bonn database, which consists of five different sets. In the first scenario, we have taken six different combinations of the available data. While in the second scenario, we have taken five combinations. For Epilepsy detection (healthy vs epileptic), the first approach performed badly. Using the DCT improved the results, but the best accuracies were obtained with the DWT-based approach. For seizure detection, the three methods performed quite well. However, the third method had the best performance and was better than many state-of-the-art methods in terms of accuracy. After carrying out the experiments on the whole EEG signal, we separated the five rhythms and applied the DWT on them with the Daubechies7 (db7) wavelet for feature extraction. We have observed that close accuracies to those recorded before can be achieved with only the Delta rhythm in the first scenario (Epilepsy detection) and the Beta rhythm in the second scenario (seizure detection).

Povzetek: Opisana je metoda zaznavanje epilepsije preko EEG signalov.

1 Introduction

The human brain is the most complex and mysterious organ of the human body, consisting of billions of neurons. It is considered as an electro-chemical machine because neurons exploit chemical reactions to generate electrical signals. These electrical signals can be monitored through different scientific techniques such as Electroencephalography (EEG), Magnetic Resonance Imaging (MRI), functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET). EEG is the most used technique to capture brain signals due to its ease of use, its excellent resolution and its low cost. It is used in the medical environment more precisely in the diagnosis and treatment of mental and neurological disorders (Alzheimer, Dementia....) and more particularly in the case of Epilepsy. According to an estimate of the World Health Organization (WHO), Epilepsy affects around 50 million people worldwide. Epilepsy is characterized by recurrent and sudden seizures. These seizures are the result of a transient and unexpected electrical disturbance of the brain and an excessive neuronal discharge that is evident in EEG. The detection of epileptic seizures by visual scanning of a patient's EEG data is a tedious and time-consuming process. In addition, it requires an expert to analyze the entire length of the EEG recordings. Moreover, the diagnosis of Epilepsy is nearly impossible from the seizure-free EEG recordings. As a result, it is necessary to develop a robust and a reliable automatic classification and detection system for Epilepsy diagnosis. For this aim, several automated EEG signal classification methods, using different approaches, have been proposed. However,

most of them deal with seizure detection only. In this work, an analysis of EEG signal is performed to detect Epilepsy during both ictal and interictal states. This is executed using three different techniques of feature extraction and three distinct classification algorithms. In order to compare the performance of these methods, each algorithm is tested on a real dataset which consists of three subject groups: healthy subjects (normal EEG), epileptic subjects during a seizure-free interval (interictal EEG), and epileptic subjects during a seizure (ictal EEG). To carry out this work, the article has been divided into four parts, briefly described as follows: the first section aims to introduce the EEG signal and the Epilepsy. The second section explains the three steps of the EEG signal analysis, which are respectively: the preprocessing step, the feature extraction step where three techniques are described, and the classification step where three classifiers are presented. Ultimately, section three illustrates the experimental part applied on the Bonn dataset and the statistical analysis for various methods proposed as well as their performances. Finally, conclusions about this work and possible perspectives are drawn.

2 EEG based methodology for epilepsy diagnosis

EEG is the most common test used to diagnose Epilepsy. The electrodes attached to the scalp, with a paste-like substance or a cap, record the electrical activity of the brain. If a person has Epilepsy, it is common to have changes in the normal pattern of brain waves, even when there is no seizure. However, the changes are more noticeable during seizure activity. The doctor may monitor patients on video when conducting an EEG while they are awake or asleep, to record any seizures they experience in order to determine their kind. The test may be done in a doctor's office or the hospital. If appropriate, an ambulatory EEG, which the patient wears at home, may be used. The EEG records seizure activity over the course of a few days. The doctor may give some instructions to trigger the seizures [1]. Recently, many researches are conducted in order to make the process of detecting Epilepsy automatic by means of machine learning. That is also the topic of interest in this work.

2.1 Literature review

Electroencephalography (EEG) records brain activities by measuring the voltage fluctuation on the scalp. This signal has a great potential for diagnosis and treatment of brain disorders. However, it is very difficult to get useful information from raw EEG signals directly. Hence, preprocessing and feature extraction steps are necessary in the EEG signal analysis. Numerous methods of feature extraction and classification have been proposed throughout the years. The Bonn database is used as a benchmark data set in many of the cited works. It consists of five sets denoted A, B, C, D and E. Sets A and B recordings belong to healthy subjects. Sets C and D recordings belong to epileptic patients during seizure-free intervals. Set E corresponds to seizure recordings. Gandhi et al. [2] used the DWT to extract three features from the EEG signals, energy, entropy and standard deviation. As classifiers, they used SVM and Probabilistic Neural Networks (PNN) to obtain a maximum accuracy of 95.44% for the ABCD-E case [3]. Nicolaou et al. [4] extracted a single feature, which is the permutation entropy from EEG signals and used the SVM classifier to report 93.5% accuracy for the A-E data sample whereas the maximum accuracy for other data samples such as B-E, C-E, D-E and ABCD-E is 86.1% [3]. M. Z. Parvez and M. Paul [5] presented an approach based on the high frequency components of The DCT for feature extraction, which are combined with the bandwidth feature extracted from the Empirical Mode Decomposition (EMD). They used the Least Square SVM (LS-SVM) classifier to identify the ictal and interictal periods of epileptic EEG signals from different brain locations. The maximum achieved accuracy on the Freiburg database was 79%. V. Bajaj and R. B. Pachori[6] proposed a novel method to detect the seizures using the Hilbert transformation of Intrensic Mode Functions (IMFs). The classification achieved an accuracy of 90% [7]. R. J. Martis et al. [8] used a decision tree classifier with energy, fractal dimension and entropy as features. The achieved accuracy is 95.7%. N. Ahammed et al. [9] used the Daubechies order 2 wavelets to extract the coefficients. The parameters fed to a linear classifier are energy, entropy, mean, maximum and minimum. They used three sets from the Bonn database, set A, set D and set E. The overall accuracy obtained is 84.2%. Juarez-Guerra et al. [10] extracted statistical features such as mean, median and variance from the EEG signals and used the feed-forward neural networks to report an accuracy of 93.23%. Zakariya Lasfer et al. [11] used only sets A and E from the Bonn database for seizure detection. They extracted the wavelet coefficients as features and calculated the energy of each wavelet coefficient. They obtained a maximum accuracy of 98.1%, a sensitivity of 97.8% and a specificity of 98.1% with the ANN classifier. A.B.Peachap and D.Tchiotsop[12] decomposed the EEG signal using Laguerre polynomials based wavelets. They reduced the dimensionality with Principal Component Analysis (PCA) and performed the classification using SVM and pattern recognition ANN. They tested multiple cases from the Bonn database. The lowest classification accuracy obtained with ANN was 94% and with SVM, it was 90%, which corresponds to data sample C-E. The best classification accuracy with ANN was 100% and with SVM, it was 98%, which corresponds to data sample B-E. They also pointed out that the scheme they used constitutes a classic case of overfitting, such as all the reported accuracies were 100% before the cross-validation.

2.2 Methodology

The Block diagram of the steps applied to EEG signal analysis in our study is presented in figure 1. We first used for the preprocessing step a Butterworth low-pass filter to correct and remove artifacts. Then, for the feature extraction step, three methods are proposed. The first one is to extract directly eight features from the original signal. In the second and third methods, features are extracted from the EEG signal after applying respectively Discrete Cosine Transform and Discrete Wavelet Transform. Concerning the classification step, we have used three classifiers, which are k-Nearest Neighbors, Support Vector Machine and Artificial Neural Network.

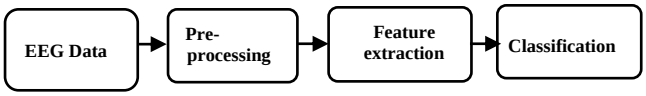
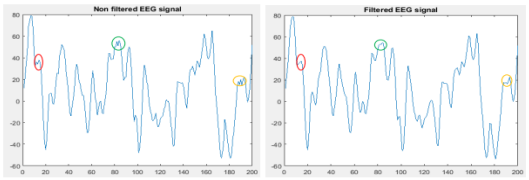


Figure 1: Block diagram of the basic steps applied to EEG signal analysis.

2.2.1 Preprocessing

EEG recording is highly susceptible to various forms of noise and artifacts, such as blinking or muscle movement, that can contaminate the data and distort the picture. Therefore, an initial task of any EEG data analysis is noise and artifact removal, which consists of separating the relevant neural signals from random neural activity that occurs during EEG recordings. This is done in the step of preprocessing, which is a procedure of transforming data into a format that is more suitable for further analysis and interpretable for the user [13]. For this preprocessing step, a filtering is done using a second-order low-pass Butterworth filter to cut off all the frequencies above 60Hz which are viewed as noise.



(a) (b)

Figure 2: One portion of an EEG signal (a) before and (b) after the filtering process.

2.2.2 EEG Signal feature extraction

After the preprocessing stage, a filtered EEG signal suitable for extracting the needed features is obtained. In this study, three methods of feature extraction are used. In the first method, we extract statistical features directly from the filtered time-domain signal. In the second method, we transform the signal to the frequency domain using DCT. While in the third method, the signal is transformed to the time frequency domain by the DWT.

A) Feature extraction using statistical parameters

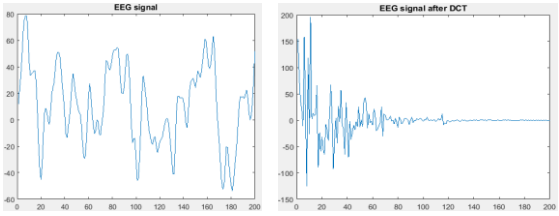
Throughout our study, eight statistical features have been introduced. They are maximum, mean, standard deviation, median, mode, first quartile, third quartile and interquartile range.

B) Feature extraction using Discrete Cosine Transform (DCT)

The Discrete Cosine Transform (DCT) is very similar to the Fourier Transform (FT), but DCT involves the use of just Cosine functions and real coefficients, whereas FT makes use of both Sine and Cosine functions and requires the use of complex numbers. Both FT and DCT are transformation methods used for converting a time series signal into basic frequency components and their respective inverse functions convert things back the other way. A DCT expresses a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. An important feature of DCT is that it takes correlated input data and concentrates its energy in just the first few transform coefficients. If the input data consists of correlated quantities, then only the first few coefficients are large and the other coefficients are zeros or small numbers. Therefore, they can be negligible. The one-dimensional DCT for a signal is given by [14]:

$$G_f = \sqrt{\frac{2}{n}} C_f \sum_{t=0}^{n-1} p_t \cos \frac{(2t+1)f\pi}{2n} \qquad (1)$$

The input is a set of n data values p_t, and the output is a set of n DCT transform coefficients (or weights) G_f.



(a) (b)

Figure 3: An EEG signal (a) before and (b) after DCT.

Figure.3(a) shows a 200 points EEG signal in time domain. While figure 3(b) shows the same sample after applying DCT on it. In frequency domain, figure 3(b), it is clear that the energy is compressed into the first few coefficients while the remaining are either null or close to zero.

C) Feature extraction using Discrete Wavelet Transform (DWT)

The DWT is computed by successively passing x[n] through a series of low-pass and high-pass filters. Each stage consists of two digital filters and two down-samplers by 2 to generate the digitized signal. The first filter, H_0, is the discrete mother wavelet, which is a high-pass filter, and the second, G_0, is a low-pass filter. The downsampled outputs of the first high-pass filter produce the detail information $d_1[n]$, while the downsampled outputs of the first low-pass filter produce the coarse

approximation, $a_1[n]$. The first approximation, $a_1[n]$, is again decomposed and this process is repeated at each stage. In this work, we have used six different families of wavelets, which are Haar, Daubechies, biorthogonal, Coiflet, Symlet and discrete Meyer [15-16].

2.2.3 EEG signal classification using machine learning

Signal classification means to analyze different characteristic features of a signal, and based on them, decide to which grouping or class the signal belongs. The resulting classification decision can be then mapped back into the physical world to reveal information about the physical process that created this signal. In order to have a broad understanding of classification, this section mainly provides an overview of used machine learning and classification algorithms.

Machine Learning is a branch of artificial intelligence based on the idea that systems can automatically learn and improve from experience without being explicitly programmed. The process of learning begins with observations or training data in order to look for patterns in that data and make better decisions in the future based on the provided data [17]. There are three types of learning approaches, namely, supervised, unsupervised and reinforcement learning. In a nutshell, reinforcement learning is dynamic programming that trains algorithms using a system of reward and punishment. Unsupervised learning is when the model is given training based on unlabeled data without any guidance while in supervised learning, the machine learns from a labeled dataset with guidance. Supervised learning uses classification algorithms and regression techniques to develop predictive models. Several algorithms have been developed [18]. In this section, the three algorithms used in the context of our study to perform binary classification are briefly explained.

A) k-Nearest Neighbor (k-NN)

The k-nearest neighbor's algorithm is a non-parametric and supervised machine learning method used for classification and regression. In classification, k-NN is based on similarity measure among the training and the testing sets. Given a point x_0 to be classified into one of *N* groups, the k nearest data points to x_0 must be found. The classification rule is to assign x_0 to the population that has the most observed data points out of the k nearest neighbors. Points for which there is no majority are either classified to one of the majority populations at random, or left unclassified [19]. The advantage of k-NN classification is its simplicity. There are only two important concepts that should be taken into consideration [20]: the parameter k, and the choice of a method to measure the distance between the attributes in both the training and the testing sets. The k-NN classification process is usually based on the following steps [21]:

- Determine parameter k as the number of nearest neighbors.
- Calculate the distance between each testing sample and all the training set element by element.
- Sort the distances and determine the k nearest neighbors.
- Determine the classes of each of the k nearest neighbors.
- Apply majority voting to decide the class of the new data.

B) Support Vector Machine (SVM)

Support vector machine, or SVM, is a machine learning algorithm initiated by Vladimir Vapnik. It was developed to solve linear or nonlinear classification and regression problems. The basic idea of the SVM classification algorithm is to construct a hyperplane that separates two groups if possible. The optimal hyperplane must have the largest distance to the nearest training-data points of the two classes in order to reduce the misclassification error. These points are called support vectors and the distance between the hyperplane and the support vectors of each class is called the margin. The goal of the SVM algorithm is to find the optimal separating hyperplane which maximizes the margin [22]. There are two types of SVMs, namely linear SVM and nonlinear SVM.

C) Artificial Neural Network (ANN)

Artificial neural networks are computing systems, in which a computer learns to perform tasks by analyzing training examples, generally without being programmed with task-specific rules [23-25]. ANNs take inspiration from the learning process of human brain. This latter is composed of cells called neurons interconnected with links (or axons). Similar to the brain, an ANN is composed of processing units called artificial neurons and interconnections. A graph of a network consists of a number of nodes connected through directional links. Each node represents a processing unit, and the links between nodes specify the causal relationship between connected nodes [21].

3 Experiments and results

This section describes and compares the performance of three methods, at the level of the feature extraction stage, proposed for Epilepsy detection from EEG signals during both *ictal* and *interictal* intervals. The raw EEG signal goes through a preprocessing step, then feature extraction and finally the classification. The same procedures are used for both experiments. The difference lies in the way we divide the data. All the details are provided later on.

3.1 Data set description

The used data set was developed by the Department of Epileptology, University of Bonn, Germany. It is made publicly available in [26]. The database consists of five separate sets denoted set A, B, C, D and E. Each containing 100 single-channel EEG samples of length 23.6s and sampled at 173.6 Hz using 12-bit resolution, resulting in 4097 data points per each signal. The amplitude is in microvolts. All the recordings were made with the same 128-channel amplifier system. Set A and set B were collected from surface EEG recordings of five healthy subjects with eyes open and eyes closed respectively. Sets C, D and E correspond to EEG records

of five epileptic patients. The samples in the first two sets are collected during seizure-free intervals (interictal), from the hippocampal formation of the opposite hemisphere of the brain and from within the epileptogenic zone respectively. Set E samples are collected during seizure activity (ictal). The properties of each set are summarized in table 1.

Table 1: Summary of the main properties of each set within the database.

	Subject state	Electrode type	Electrode placement
Set A	Healthy Eyes open	Surface	International 10-20 system
Set B	Healthy Eyes closed	Surface	International 10-20 system
Set C	Epileptic Interictal	Intracranial	Opposite to epileptogenic zone
Set D	Epileptic Interictal	Intracranial	Within epileptogenic zone
Set E	Epileptic Ictal	Intracranial	Within epileptogenic zone

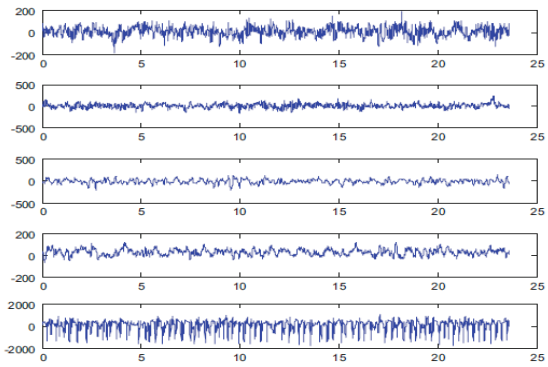


Figure 4: Example of an EEG signal from (a) Set A (b) Set B (c) Set C (d) Set D (e) Set E [27].

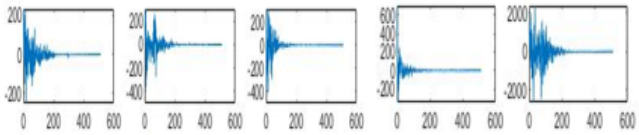
Figure 4 depicts five samples of the EEG recordings from the five different sets in the Bonn database. The y-axis corresponds to the amplitude in microvolts and the x-axis corresponds to the time in seconds.

3.2 Experimental procedure

Three methods are proposed for two experiments. In the first experiment, Epilepsy is detected mainly from the interictal intervals and the implemented scenario is healthy vs. epileptic. Therefore, all the samples in the dataset fall in two classes: healthy, for sets A and B, and epileptic for sets C, D and E. In the second experiment, Epilepsy is detected from ictal intervals and the implemented scenario is seizure-free vs. seizure. Since the database has only one set with ictal samples, sets A, B, C and D fall in the first class which is seizure-free (regardless of whether the subject is healthy or epileptic) while set E samples belong to the second class, ictal. For simplicity, we will refer to the first experiment as Epilepsy detection and the second as seizure detection throughout the whole section. In order to have good training and validate the results with a test dataset, the Bonn database is quite limited. To tackle this issue, an augmentation scheme is proposed. Each EEG signal is divided into 8 samples using a window length of 512 data points with no overlap. The resulting samples are treated as independent instances. Therefore, the augmented database has 800 signals per each set, which sums up to a total of 4000 samples.

3.2.1 Feature extraction step

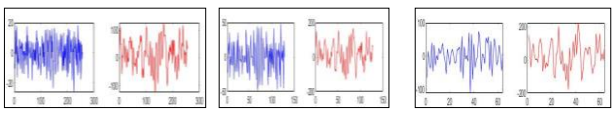
The choice of the right features plays a major role in classification problems. In the first method, eight statistical features are extracted directly from the signal to summarize the relevant information contained in it. Hence, this method relies only on time-domain information. The used statistical features are maximum amplitude, mean, mode, median, standard deviation, first quartile, third quartile and interquartile. The second method relies solely on frequency domain information using the DCT, which is a widely used data compression technique. Since energy is concentrated in low frequencies, we keep only the first 150 coefficients (29.3% of the signal after the transformation). Then, we extract four features, which are mean of the absolute value of the coefficients, interquartile, energy and entropy. We will later show that further reduction is possible on the number of input features.



(a) (b) (c) (d) (e)

Figure 5: DCT of a signal from (a) Set A (b) Set B (c) Set C (d) Set D (e) Set E.

The third method is based on the DWT, which captures both frequency and location in time information. The first three decomposition levels are tested separately. Figure 6 illustrates the plots of detail (in red) and approximation (in blue) coefficients, using the Haar wavelet on a sample from set A. Instead of directly feeding the coefficients to the classifier, we summarize the relevant information in 16 statistical features, 8 for the detail coefficients and 8 for the approximation coefficients. These features are maximum, mean of the absolute value of the coefficients, mode, median, standard deviation, first quartile, third quartile and interquartile.



(a) (b) (c)

Figure 6: Discrete Haar wavelet coefficients on a set A signal at (a) level 1 (b) level 2 (c) level 3.

3.2.2 Classification step

After extracting the selected features depending on the used method, they are fed to three different classifiers to compare their performances. The first classifier is k-NN, the second is SVM, and the last is ANN. To train both k-NN and SVM models, we used the software Matlab R2018b. The two classifiers are already implemented in the Statistics and Machine Learning Toolbox as the two functions *fitcknn* and *fitcsvm*. To train the ANN classifier, the model was built with Python 3.6. It is made exclusively of dense layers from the Keras library as we are using a simple MLP.

The model consists of four hidden layers; the first layer has 30 neurons, while the remaining three were implemented with 20 neurons each. The ReLU activation function was used for the hidden layers, and the sigmoid activation function was chosen for the output layer.

3.2.3 Evaluation parameters

The data is divided into 75% for the training and 25% for testing. The performance metrics used for the evaluation of the model are accuracy, sensitivity, and specificity. The accuracy (acc) of a classifier is its ability to differentiate between positive and negative cases correctly. Mathematically, it is expressed as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

Where:

- TP (true positive) is the number of cases correctly identified as positive (unhealthy).
- TN (true negative) is the number of cases correctly identified as negative (healthy).
- FP (false positive) is the number of cases incorrectly identified as positive.
- FN (false negative) is the number of cases incorrectly identified as negative.

The sensitivity (sen) of a binary classification model is its ability to determine the positive cases correctly, whereas, the specificity (spe) measures its ability to identify negative cases correctly. They are calculated as follows:

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

3.2.4 Experiment 1: epilepsy detection

In this experiment, the goal is to identify whether a subject has Epilepsy or not mainly from interictal intervals. Several data samples of the Bonn database are tested. First, a pair from the four sets (excluding set E) is taken each time (a healthy set and an epileptic set) resulting in four combinations: A-C, A-D, B-C, and B-D. Then, sets A and B are grouped to form the healthy class while sets C and D form the epileptic class. Finally, set E is added to the latter. For each pair, 1200 samples are used for the training, and 400 samples for the testing. In each train and test dataset, the positive and negative cases are equal. The data sample AB-CD is divided into 2400 samples for the training and 800 samples for the testing. Here again, the epileptic portion and the healthy portion are of equal size. The last data sample, which includes the whole database, is divided into 3000 samples for training, from which 1200 are healthy cases and 1800 are epileptic cases, and 1000 samples for testing, where 400 are negative cases and the remaining 600 are positive cases.

A) Method 1: Feature extraction using statistical parameters

As mentioned before, the first method is based on the extraction of statistical features directly from the original signal in the time domain. The results are recorded in table 2, table 3 and table 4 for the k-NN classifier, SVM classifier and ANN classifier, respectively.

Table 2: The obtained results for Epilepsy detection with the k-NN classifier using the first method (statistical features applied on the original signal).

		A-C	A-D	B-C	B-D	AB-CD	AB-CDE
k = 3	Acc (%)	74.75	74.75	67.75	72	66.12	69.8
	Sen (%)	67	69	49	62.5	56.5	67.17
	Spe (%)	82.5	80.5	86.5	81.5	75.75	73.75
k = 5	Acc (%)	77	76.25	70	71.25	68	71.2
	Sen (%)	70	67	51	60	55.25	65.67
	Spe (%)	84	85.5	89	82.5	80.75	79.5
k = 8	Acc (%)	78.25	76.25	70.5	74.75	68.12	72.3
	Sen (%)	73.5	71	54	67	60.5	62
	Spe (%)	83	81.5	87	82.5	75.75	87.75

Table 3: The obtained results for Epilepsy detection with the SVM classifier using the first method (statistical features applied on the original signal).

	A-C	A-D	B-C	B-D	AB-CD	AB-CDE
Acc (%)	82.75	81.7	73	78.75	74.87	75.2
Sen (%)	80.5	71.5	56.5	69	66.25	75
Spe (%)	85	92	89.5	88.5	83.5	75.5

Table 4: The obta-ined results for Epilepsy detection with the ANN classifier using the first method (statistical features applied on the original signal).

	A-C	A-D	B-C	B-D	AB-CD	AB-CDE
Acc (%)	72.5	78.5	71	74.75	69.5	76
Sen (%)	65	66	63.5	74	55	79.33
Spe (%)	80	91	78.5	75	84	75.5

When using the k-NN classifier, changing the parameter k affects the accuracy, such that it increases when we increase the number of nearest neighbors. The average accuracy is 73.36% for k = 8, which makes the k-NN classifier the least performing in this case, followed by ANN with an average accuracy of 73.7%. The SVM classifier has the best performance with an average accuracy of 77.72%. Generally, the pairs with set A as the healthy set give better results than with set B. It is worth noting that the two resting states eyes-open and eyes-closed have different impacts on the brain activity, which results in the observed difference. Mostly, the recorded specificity is higher than the sensitivity. In other words, the three models tend to misclassify the epileptic cases more than the healthy cases. The first method resulted in poor performance. The time-domain

information alone is far from enough for Epilepsy detection. We remarked from the different features used for the four samples (set A, set B, set C and set D) that there is no obvious distinction between the healthy and epileptic cases, which would explain the confusion of the classifiers. However, since set E signals are recorded during the seizure, they are distinguishable from the rest.

B) Method 2: Feature extraction using DCT

Since extracting the statistical features directly from the original signal resulted in a bad performance, we moved to the frequency domain with the DCT to see if that leads to any improvement. Using the four features mentioned in section 3.2.1, the results are recorded in tables 5-6 for the classifiers k-NN and SVM respectively.

Table 5: The obtained results for Epilepsy detection with the k-NN classifier using the second method (four statistical features applied on the DCT coefficients).

		A-C	A-D	B-C	B-D	AB-CD	AB-CDE
k= 3	Acc (%)	91.5	90.75	81.5	89.75	85.25	88
	Sen (%)	85.5	86.5	65.5	84	78	84.83
	Spe (%)	91.5	95	97.5	95.5	92.5	92.75
k= 5	Acc (%)	91.5	92	81	91.5	87	88.8
	Sen (%)	85.5	87.5	64	86.5	80	86
	Spe (%)	97.5	96.5	98	96.5	94	93
k= 8	Acc (%)	91.75	92.75	84.25	91	87.87	88.7
	Sen (%)	86	91	70.5	87	82.75	84.33
	Spe (%)	97.5	94.5	98	95	93	95.25

Table 6: The obtained results for epilepsy detection with the SVM classifier using the second method (four statistical features applied on the DCT coefficients).

	A-C	A-D	B-C	B-D	AB-CD	AB-CDE
Acc (%)	94.25	92.25	81.5	91.5	87.87	89.2
Sen (%)	90	86.5	64	84	77	84.83
Spe (%)	98.5	98	99	99	98.75	95.75

The best average accuracy with the k-NN classifier, 89.39%, was again achieved with parameter k=8. The SVM model performed barely better with an average accuracy of 89.43%. The performance was especially bad with data sample B-C compared to the other pairs where the accuracy was greater than 90%. The correctly classified cases are not equally distributed over the two classes with both models, as they tend to "favor" the healthy class. The specificity recorded with the SVM classifier was generally greater than 98% (except with data sample AB-CDE), unlike the sensitivity, which was quite low. k-NN was slightly better, as it offers more balance between the two metrics.

To see if there were any redundant features in the input vector, we removed one feature at a time and observed the results. We concluded that the dimensionality could be reduced to half the original one. Both energy and entropy were redundant and therefore removed. The results are shown in table 7, table 8 and table 9 for the classifiers k-NN, SVM and ANN, respectively.

Table 7: The obtained results for Epilepsy detection with the k-NN classifier (k=8) using the second method after the dimensionality reduction of the input vector.

	A-C	A-D	B-C	B-D	AB-CD	AB-CDE
Acc (%)	93.5	92	86.75	91.75	88.75	89.1
Sen (%)	90.5	88.5	75	85.5	82.5	84.17
Spe (%)	96.5	95.5	98.5	98	95	96.5

Table 8: The obtained results for Epilepsy detection with the SVM classifier using the second method after the dimensionality reduction of the input vector.

	A-C	A-D	B-C	B-D	AB-CD	AB-CDE
Acc (%)	94.25	92.25	84	90.75	87.75	89.3
Sen (%)	90.5	87	69	82.5	77.5	85.33
Spe (%)	98	97.5	99	99	98	95.25

Table 9: The obtained results for Epilepsy detection with the ANN classifier using the second method after the dimensionality reduction of the input vector.

	A-C	A-D	B-C	B-D	AB-CD	AB-CDE
Acc (%)	93	92	87.75	88.75	89.37	90.1
Sen (%)	88.5	86	76	78.5	80.75	85
Spe (%)	97.5	98	99.5	99	98	97.75

After reducing the size of the input vector, the best average accuracy recorded is 90.30% with the k-NN classifier (a gain of almost 1%), followed by ANN with an average accuracy of 90.16%, then SVM with an average accuracy of 89.72%. Relying on the frequency domain information with the DCT improved the performance considerably compared to the first method. The gain is 16.94% with k-NN, 16.46% with ANN and 12% with SVM. Nevertheless, the results for some data samples are still not satisfying, especially the sensitivity, which is quite low in most cases.

C) Method 3: Feature extraction using DWT

As an attempt to farther improve the performance for Epilepsy detection, we have used a powerful mathematical tool, which is the DWT, to extract the statistical features from the generated approximation and detail coefficients. We have recorded the results for 37 wavelets from six different families, which are Haar, Daubechies, Biorthogonal, Coiflet, Symlet and discrete Meyer. We tested the three first decomposition levels separately, but only the best accuracy was recorded with the corresponding level. The table 10 and table 11 show only 6 wavelets for which the accuracy was highest with k-NN and SVM classifiers respectively. Table 12 refers to the results achieved with ANN.

Table.10: The obtained results for epilepsy detection with the k-NN classifier using the third method (extracting statistical features from the DWT coefficients).

Data sample	Wavelet	Level	Acc (%)	Sen (%)	Spe (%)
A-C	Db5	2	92	84.5	99.5
	Db7	1	93	88.5	97.5
	Bior2.4	1	92.5	86	99
	Bior2.6	1	91.75	85	98.5
	Bior5.5	2	91.75	84.5	99
	Coif4	2	92.75	87.5	98
A-D	Db10	3	93	88.5	97.5
	Bior2.4	3	93.25	88	98.5
	Bior3.3	3	93	86.5	99.5
	Bior4.4	3	93.25	88.5	98
	Bior5.5	3	93	89	97
	Sym5	3	92.75	86.5	99
B-C	Db3	3	93.75	87.5	100
	Db5	3	93	87	99
	Db7	3	93	86	100
	Db9	3	93.5	87.5	99.5
	Db10	3	93.75	88	99.5
	Sym3	3	93.75	87.5	100
B-D	Db6	3	97.75	95.5	100
	Db10	3	98	96	100
	Bior4.4	3	97.75	95.5	100
	Bior5.5	3	97.75	97	98.5
	Coif4	3	97.75	96.5	99
	Sym8	3	98.75	97.5	100
AB-CD	Db5	3	91	84	98
	Db7	3	91.37	84.5	98.25
	Db10	3	92.25	86.25	98.25
	Bior6.8	3	91.5	84.5	98.5
	Coif4	3	91.62	85	98.25
	Sym5	3	91.12	83	99.25
AB-CDE	Db3	3	92	88.5	97.25
	Db5	3	91.7	87	98.75
	Db10	3	91.8	89	96
	Coif3	3	91.6	86.83	98.75
	Sym3	3	92	88.5	97.25
	Sym5	3	92.3	88.5	98

Table 11: The obtained results for Epilepsy detection with the SVM classifier using the third method (extracting statistical features from the DWT coefficients).

Data sample	Wavelet	Level	Acc (%)	Sen (%)	Spe (%)
A-C	Db2	2	95.75	93	98.5
	Bior1.3	1	94.75	93.5	96
	Bior2.4	1	94.5	94	95
	Bior2.6	1	94.5	94	95
	Sym2	1	95.75	94	97.5
	Sym6	1	94.5	93	96
A-D	Db3	3	93.5	90.5	96.5
	Db5	2	93.75	88.5	99
	Bior2.4	3	93.5	92.5	94.5
	Bior2.8	3	93.75	91	96.5
	Bior5.5	3	93.5	89.5	97.5
	Sym3	3	93.5	90.5	96.5
B-C	Db1	2	84.25	68.5	100
	Db2	3	83.5	67	100
	Db3	3	84.5	69	100
	Sym2	3	83.5	67	100
	Sym3	3	84.5	69	100
	Sym7	3	83.5	67	100
B-D	Db1	2	95.25	93.5	97
	Db10	3	93.75	87.5	100
	Bior1.3	1	95.75	94	97.5
	Bior1.5	1	96.5	94.5	98.5
	Coif4	3	94	88	100
	Sym8	3	93.25	86.5	100
AB-CD	Db1	1	91	84.75	97.25
	Db3	3	88.25	77.75	98.75
	Bior1.3	1	89.37	79.75	97
	Bior1.5	1	89.5	81.25	97.75
	Coif3	3	88	76.75	99.25
	Sym3	3	88.25	77.75	98.75
AB-CDE	Db3	3	94.6	92	98.5
	Db5	3	94.3	90.83	99.5
	Bior1.3	1	93.8	90	99.5
	Sym3	3	94.6	92	98.5
	Sym4	3	93.9	90.83	98.5
	Sym5	3	94.1	91	98.75

Table 12: The obtained results for Epilepsy detection with ANN classifier using the third method (extracting statistical features from the DWT coefficients).

Data sample	Wavelet	Level	Acc (%)	Sen (%)	Spe (%)
A-C	Bior2.4	1	90.5	83	98
A-D	Bior2.4	3	93.75	88.5	99
B-C	Db3	3	94.5	90.5	98.5
B-D	Coif4	3	98	96	100
AB-CD	Db10	3	94	90	98
AB-CDE	Db3	3	93.5	91.17	97

We observe from the obtained results that there is no "best wavelet" for EEG data, which would give the highest accuracy for all cases. It depends on both the data sample and the selected classifier. However, the db10 wavelet achieved the best average accuracy of 93.26% with k-NN. SVM was especially sensitive to the change in the training data such that the performance drops drastically with the sample B-C. It is also the least performing classifier with an average accuracy of 92.68%. k-NN was more stable and the least sensitive to data change, wavelet and level change. The average accuracies for the two classifiers, k-NN and ANN were 93.88% and 94.04% respectively. Probably, better results could be obtained with the latter since we tested the model with only one wavelet for each data sample. The choice of the wavelet for ANN was based on the results obtained with the two other classifiers. We choose one with which the accuracy was high for both classifiers. The DWT has indeed improved the overall performance. All the samples have a higher accuracy than 90% (except with SVM). The sensitivity is still lower than the specificity, but considerably high compared to the previous method.

After carrying on the experiment with the whole EEG signals and deducing that the DWT based method has the best accuracy for Epilepsy detection, we decided to test the performance on the separate EEG rhythms and see whether we can achieve close results with only one rhythm. The rhythms were obtained from filtering the original signal using a second-order Butterworth filter. The wavelet used throughout the whole experiment is db7 (Daubechies order 7). The wavelet choice was not random, it was obtained empirically, but there is no guarantee that this is the best choice. It is worth noting that unlike when using the whole signal, changing the wavelet when dealing with the rhythms separately could lead to very different results (up to 20% difference in the accuracy was observed when testing different wavelets). The used classifiers are SVM and k-NN; however, we only recorded the results obtained with the latter, as shown in table 13, since it had a better performance.

Table 13: The obtained results for Epilepsy detection with the k-NN classifier using the DWT coefficients after decomposing the EEG signal into 5 rhythms.

		A-C	A-D	B-C	B-D	AB-CD	AB-CDE
Delta Rhythm	Acc (%)	92.75	93.25	94.25	96.75	93.12	92.9
	Sen (%)	86	87.5	91	94	88.25	90.17
	Spe (%)	99.5	99	97.5	99.5	98	97
Theta Rhythm	Acc (%)	87.5	87.25	88.5	91.5	88.12	90.8
	Sen (%)	79	78.5	85.5	89.5	83	88.17
	Spe (%)	96	96	91.5	93.5	93.25	97.75
Alpha Rhythm	Acc (%)	76.5	84.25	88.5	91.75	82.12	82.1
	Sen (%)	64	76	78.5	84.5	73	78.5
	Spe (%)	89	92.5	98.5	99	91.25	87.5
Beta Rhythm	Acc (%)	78	81.25	84.25	91.25	81.62	83
	Sen (%)	63.5	69	71	83.5	70.5	77
	Spe (%)	92.5	93.5	97.5	99	92.75	92
Gamma Rhythm	Acc (%)	80	83.5	88.5	85.5	81.12	83.1
	Sen (%)	71	72.5	78.5	77.5	70.75	78.17
	Spe (%)	89	94.5	98.5	93.5	91.5	90.5

We observe from the obtained results that Epilepsy is better detected in low frequency elements (<8Hz). The best performance was recorded with the Delta rhythm, which has the lowest frequency band (<4Hz) and the highest average accuracy, 93.84%, followed by the theta rhythm (4Hz< frequency <8Hz) with an average accuracy of 88.95%. Then, Alpha, Gamma and Beta rhythms with average accuracies 84.20%, 83.62% and 83.23% respectively. The best accuracy was achieved with data sample B-D, 96.75%, which also has the highest sensitivity and specificity, 94% and 99.5% respectively. Using only the Delta rhythm instead of the whole EEG signal leads to almost the same results, with a loss of only 0.04% in accuracy, a gain of 0.03% in sensitivity and a loss of 0.2% in specificity. Using a different method does not forcibly lead to the same conclusions.

3.2.5 Experiment 2: Seizure detection

This experiment aims to identify epileptic seizures from EEG data. Several samples of the Bonn database are tested. First, we take set E, which represents the *ictal* class, with one of the remaining four sets each time, resulting in four combinations: A-E, B-E, C-E and D-E. Then, we use the whole database where sets A, B, C and D form the seizure-free class and set E forms the ictal class. Table 14 shows how the data was divided between training and testing the models.

Table 14: Data division to train and test the models for seizure detection.

Data sample	Purpose	EEG recordings	Seizure-free cases	Ictal cases
Pairs (A-E, B-E, C-E and D-E)	Training	1200	600	600
	Testing	400	200	200
ABCD-E	Training	3000	2400	600
	Testing	1000	800	200

A) Method 1: Feature extraction using statistical parameters

After extracting the features from the original signal in time-domain, the results are recorded in table 15 with the k-NN classifier, table 16 with the SVM classifier, and table 17 with the ANN classifier.

Table 15: The obtained results for seizure detection with the k-NN classifier using the first method (statistical features applied on the original signal).

		A-E	B-E	C-E	D-E	ABCD-E
k = 3	Acc (%)	99.75	96	97.75	94.25	97.1
	Sen (%)	99.5	93	98.5	94.5	90.5
	Spe (%)	100	99	97	94	98.75
k = 5	Acc (%)	99.75	96.25	97.75	95.5	97.2
	Sen (%)	99.5	93	98	96	89.5
	Spe (%)	100	99.5	97.5	95	99.12
k = 8	Acc (%)	99.75	95.75	98.25	94.25	96.5
	Sen (%)	99.5	92.5	99	96	90
	Spe (%)	100	99	97.5	92.5	98.12

Table 16: The obtained results for seizure detection with the SVM classifier using the first method **(**statistical features applied on the original signal).

	A-E	B-E	C-E	D-E	ABCD-E
Acc (%)	100	95.25	98.5	93.75	95.8
Sen (%)	100	93	99	95	89.5
Spe (%)	100	97.5	98	92.5	97.37

Table 17: The obtained results for seizure detection with the ANN classifier using the first method (statistical features applied on the original signal).

	A-E	B-E	C-E	D-E	ABCD-E
Acc (%)	99.75	95.75	98.5	94.75	96.8
Sen (%)	99.5	92.5	99	96	89.5
Spe (%)	100	99	98	93.5	98.62

The performance of the three classifiers was quite good, unlike the obtained results for Epilepsy detection. This is due to the remarkably high peaks in the EEG data, which results from the hyper-activity of the brain during seizure intervals. The results illustrate clearly the big difference in statistical features between set E samples and the other sets. It also justifies why we obtained the lowest accuracy with the data sample D-E. The best set used with set E in the training was set A, which represents the EEG recordings of healthy subjects with eyes open. It resulted in an accuracy of 100% with SVM and 99.75% with both k-NN and ANN. The effect of varying the parameter k in the k-NN model is barely noticeable. The best average accuracy of 97.29%, was recorded with k=5. The least performing classifier was SVM with an average accuracy of 96.66% followed by ANN with an average accuracy of 97.11%. When using the whole database, the sensitivity was especially lower than the specificity compared to the values obtained with the pairs. This is probably due to the unbalance of the positive and negative cases in the training data set. The negative class was 4 times bigger than the positive class, which resulted in lower sensitivity.

B) Method 2: Feature extraction using DCT

In the previous experiment, Epilepsy detection, the two features, energy and entropy, were proved redundant in the input vector. However, since we did not want to generalize the observation to this experiment, we observed the results with both 2 and 4 features with the SVM classifier. Once again, the energy and entropy were found to be unnecessary. Therefore, table 18, table 19, and table 20 refer to the obtained results with 2 features, mean and interquartile, with k-NN, SVM and ANN classifiers, respectively.

Table 18: The obtained results for seizure detection with the k-NN classifier using the second method (two statistical features applied on the DCT coefficients).

		A-E	B-E	C-E	D-E	ABCD-E
k = 3	Acc (%)	100	97.5	96.5	95.25	96.7
	Sen (%)	100	97.5	98.5	97	92.5
	Spe (%)	100	97.5	94.5	93.5	97.75
k = 5	Acc (%)	100	97.75	97	96	97.1
	Sen (%)	100	98.5	99.5	98.5	95.5
	Spe (%)	100	97	94.5	93.5	97.5
k = 8	Acc (%)	100	97.5	97.25	95.75	97.3
	Sen (%)	100	98.5	99.5	99	97
	Spe (%)	100	96.5	95	92.5	97.37

Table 19: The obtained results for seizure detection with the SVM classifier using the second method (two statistical features applied on the DCT coefficients).

	A-E	B-E	C-E	D-E	ABCD-E
Acc (%)	99.75	96.75	98.25	96	96.9
Sen (%)	99.5	96	99	98	92
Spe (%)	100	97.5	97.5	94	98.12

Table 20: The obtained results for seizure detection with the ANN classifier using the second method (two statistical features applied on the DCT coefficients).

	A-E	B-E	C-E	D-E	ABCD-E
Acc (%)	99.75	97.25	98.25	96.25	97.5
Sen (%)	100	97	99.5	97.5	96.5
Spe (%)	99.5	97.5	97	95	97.75

Relying on the frequency domain information slightly improved the overall performance. The recorded accuracies for data samples B-E and D-E are higher compared to the previous method. Although, the best data combination is still A-E and the worst is still D-E. The best classifier was ANN with an average accuracy of 97.8% followed by k-NN and SVM with an average accuracy of 97.57% (k=5) and 97.53%, respectively. The main advantage of applying the DCT to the original signal before feature extraction over the previous method is the high sensitivity recorded when using the whole database, such that both sensitivity and specificity are greater than 96% with the best classifier ANN.

C) Method 3: Feature extraction using DWT

As in the previous experiment, Epilepsy detection, 37 different wavelets from 6 families were tested with k-NN and SVM. Table 21 and table 22 refer to the obtained results, using the DWT coefficients, with the best 6 performing wavelets in each data sample, with k-NN and SVM, respectively. Table 23 refers to the results obtained with the ANN classifier using only a single wavelet per data sample.

Table 21: The obtained results for seizure detection with the k-NN classifier using the third method (extracting statistical features from the DWT coefficients).

Data sample	Wavelet	Level	Acc (%)	Sen (%)	Spe (%)
A-E	Db1	3	100	100	100
	Db4	3	100	100	100
	Bior2.2	3	100	100	100
	Coif1	3	100	100	100
	Sym2	3	100	100	100
	Dmey	3	100	100	100
B-E	Db1	2	97	94	100
	Db2	1	97	94	100
	Bior2.2	2	97	94	100
	Bior2.4	3	96.75	94	99.5
	Sym2	1	97	94	100
	Sym4	3	96.75	93.5	100
C-E	Bior2.2	3	99.5	100	99
	Bior2.8	3	99.25	99	99.5
	Bior3.3	2	99.75	99.5	100
	Bior3.7	2	99.5	99	100
	Coif1	3	99.25	99.5	99
	Sym4	2	99.25	99	99.5
D-E	Db1	2	98.25	97.5	99
	Db3	3	98.5	99	98
	Db5	3	98.25	98.5	98
	Coif2	3	98.25	99.5	97
	Sym3	3	98.5	99	98
	Sym5	3	99	99.5	98.5
ABCD-E	Db3	3	97.8	91.5	99.37
	Bior2.2	2	97.9	92.5	99.25
	Bior5.5	2	97.9	91	99.62
	Coif1	1	97.8	91.5	99.37
	Sym3	3	97.8	91.5	99.37
	Sym5	3	98	92	99.5

Table 22: The obtained results for seizure detection with the SVM classifier using the third method (extracting statistical features from the DWT coefficients).

Data sample	Wavelet	Level	Acc (%)	Sen (%)	Spe (%)
A-E	Db1	3	100	100	100
	Db5	3	100	100	100
	Bior2.6	3	100	100	100
	Coif2	3	100	100	100
	Sym5	3	100	100	100
	Dmey	3	100	100	100
B-E	Db1	2	97.75	95.5	100
	Db2	3	97.75	96	99.5
	Bior2.4	2	98	96.5	99.5
	Bior2.6	2	98.25	96.5	100
	Coif4	3	97.75	95.5	100
	Sym2	3	97.75	96	99.5
C-E	Bior2.2	1	99.5	100	94
	Bior2.4	1	99.75	100	99.5
	Bior2.6	1	99.5	100	99
	Bior2.8	2	99.5	99.5	99.5
	Bior3.1	3	99.5	100	99
	Coif1	1	99.5	100	99
D-E	Db1	3	96.5	100	93
	Db7	3	96.75	99.5	94
	Bior2.6	3	96.75	99.5	94
	Bior3.1	3	97	99	95
	Coif1	3	97.25	100	94.5
	Coif5	3	96.5	99.5	93.5
ABCD-E	Db1	1	97.5	95.5	98
	Bior2.6	1	97.4	94.5	98.12
	Coif1	1	97.6	95.5	98.12
	Coif2	3	97.6	94.5	98.37
	Sym2	2	97.4	95.5	97.87
	Sym5	2	98	97.5	98.12

Table 23: The obtained results for seizure detection with ANN classifier using the third method (extracting statistical features from the DWT coefficients).

Data sample	Wavelet	Level	Acc (%)	Sen (%)	Spe (%)
A-E	Db1	3	100	100	100
B-E	Db1	2	97.25	94.5	100
C-E	Bior3.3	2	98.75	97.5	100
D-E	Coif2	3	97.25	95	99.5
ABCD-E	Sym5	3	98.2	91.5	99.87

Table 24: The obtained results for seizure detection with the SVM classifier using the DWT coefficients after decomposing the EEG signal into 5 rhythms.

		A-E	B-E	C-E	D-E	ABCD-E
Delta Rhythm	Acc (%)	92.75	93.25	94.25	96.75	92.9
	Sen (%)	86	87.5	91	94	90.17
	Spe (%)	99.5	99	97.5	99.5	97
Theta Rhythm	Acc (%)	99.75	99	96.5	95.75	97.9
	Sen (%)	99.5	98.5	96	93.5	91.5
	Spe (%)	100	99.5	97	98	99.5
Alpha Rhythm	Acc (%)	100	91.5	98.25	98.5	96
	Sen (%)	100	88.5	99	98	86
	Spe (%)	100	94.5	97.5	99	98.5
Beta Rhythm	Acc (%)	98.5	96	98.25	98.75	97.6
	Sen (%)	99	94.5	100	98	94
	Spe (%)	98	97.5	96.5	99.5	98.5
Gamma Rhythm	Acc (%)	96.75	88.25	94.5	96	92.3
	Sen (%)	94.5	98	98	97.5	91
	Spe (%)	99	78.5	91	94.5	92.62

All three classifiers have led to perfect accuracy (100%) with data sample A-E. The wavelet choice with the latter is quite irrelevant. The best performing classifier was k-NN with an average accuracy of 98.75%, followed by SVM with an average accuracy of 98.65%, then ANN with an average accuracy of 98.29%. Again, it is worth noting that only one wavelet was tested with the ANN classifier for each data sample. Therefore, it is highly possible to record better accuracy with different wavelets, and the order is not final. The DWT based method has resulted in the best performance for seizure detection, such that all accuracies, regardless of the data sample and the classifier, were greater than 97%. However, the bior2.2 wavelet achieved the best average accuracy, 98.48% with k-NN. The lowest sensitivity recorded with the best classifier (k-NN) was 92% when using the whole database. Whereas, the specificity did not drop below 98.5%. For all three methods, it is safe to generalize that for the negative class, using set A instead of set B (healthy sets) and set C instead of set D (epileptic interictal sets) during the training leads to higher accuracy in seizure detection.

As it was done in the previous experiment, Epilepsy detection, we tested the DWT based method with the separate EEG rhythms to see if we can narrow down the input to only one rhythm instead of the whole signal. The wavelet used is db7, and again, there is no guarantee that this is the best choice. Two classifiers were tested, SVM and k-NN. The former has the best performance with all rhythms except Gamma. Table 24 refer to the results obtained with the SVM classifier.

Generally, the overall performance was good with all five rhythms. The highest average accuracies were achieved with the Beta and Theta rhythms, 97.82% and 97.78% respectively, followed by Alpha with an average accuracy of 96.85%. Then, Delta and Gamma rhythms with average accuracies 95.86% and 93.56%, respectively. The detection of epileptic seizures is higher in the frequency band 4 Hz to 30 Hz. The main difference between the results recorded with Theta and Beta rhythms is that higher accuracies (≥99%) were achieved with the Theta rhythm with the healthy sets (A and B) whereas, the results achieved with the interictal sets (C and D) were better with the Beta rhythm (≥98.25%). Also, the latter has the best average sensitivity, 97.1%, which is 1.3% higher than the sensitivity recorded with the Theta rhythm. However, the average specificity of the latter, 98.8% is 0.8% greater than the recorded average specificity with the Beta rhythm. The results achieved with the Beta rhythm are very close to those achieved with the whole signal. There is a loss of 0.93% in accuracy, a gain of 0.1% in sensitivity, and a loss of 1.6% in specificity. Here again, the drawn conclusions concern only this method. The fact that epileptic seizures were best detected with the Beta rhythm cannot be generalized to other researches with different methods.

3.2.6 Discussions

In these experiments, we presented three methods for two types of problems concerning Epilepsy. The first one is the detection of the disease during seizure-free intervals from EEG data. The second is the identification of the epileptic seizures from the same data. The difference between the presented methods lies in the features extraction stage. In the first method, we directly used the original signal to extract 8 statistical features. In the second and third methods, an extra step is added. In the former, we first obtained the DCT coefficients then summarized the relevant information in 2 features, whereas in the latter, we used the DWT transformation on the signal then we extracted 16 features. We preferred to perform the classification with more than one model. Hence, we used three classifiers k-NN, SVM, and ANN. Several data samples were tested.

Table 25: The average accuracies obtained for Epilepsy detection using different feature extraction methods and classifiers.

	K-NN	SVM	ANN
Statistical Parameters	73.36%	77.72%	73.7%.
DCT	90.30%	89.72%	90.16%,
DWT	93.88%	92.68%.	94.04%

Table 26: The average accuracies obtained for Seizure detection using different feature extraction methods and classifiers.

	K-NN	SVM	ANN
Statistical Parameters	97.29%	96.66%	97.11%
DCT	97.57%	97.53%,	97.8%
DWT	98.75%	98.65%,	98.29%.

For Epilepsy detection, the first method was proved to be the worst with an average accuracy of 77.72% using SVM as shown in table 25. The best accuracy achieved was 82.75%, for the A-C data sample. But, in most cases, the accuracy was less than 80%. The second method, based on the DCT, performed better. The accuracy was greater than 90% for three data samples and greater than 80% for the remaining three. The best overall performance was achieved with the last method based on the DWT with average accuracy 94.04% using ANN as shown in table 25.

For all data samples, with the k-NN classifier, the minimum accuracy recorded was 92.25%, the minimum sensitivity was 86.25%, and the minimum specificity was 97.5%. For seizure detection, all three methods had a decent performance. Although, the order was the same as in the first experiment. The least average accuracy recorded was 97.29% using the first method (with the k-NN classifier) as shown in table 26. The DCT didn't improve significantly the performance, since we noticed an average gain of only 0.51% in the accuracy (with the ANN classifier).

The best performance was recorded with the DWT based method, where the average accuracy was 98.75%, the average sensitivity was 97%, and the average specificity was 99.6% (with the k-NN classifier).

The last step in both experiments was to test the DWT based method on the five rhythms extracted from the EEG signal. We observed that for Epilepsy detection, almost the same performance could be achieved from only the Delta rhythm. Whereas for seizure detection, very close results to those recorded with the whole signal were achieved from the Beta rhythm.

4 Conclusion

The EEG test gives information about the electrical activity carried out in the brain. It is the most suitable test for Epilepsy diagnosis since epileptic seizures are characterized by the abnormal brain activity and the unnaturally high spikes of voltage recorded during seizure. Many researches were carried out in order to automatize the diagnosis using machine learning. Most of them are based on seizure detection for Epilepsy diagnosis. Our contributions in this study are that we worked on the diagnosis during both ictal (during seizure) and interictal (seizure-free) activities in two different experiments; we have used three techniques for the feature extraction stage and three different classifiers to compare their performances, and we decomposed the EEG signal into five rhythms to deduce the best rhythm for the diagnosis. The first technique is based on the time domain information only, the second on the frequency domain information only and the third is based on both.

Extracting statistical features directly from the time domain signal was the least performing technique especially during interracial intervals. Using the DCT on the signal then extracting statistical features from the coefficients improved considerably the performance compared to the previous technique. As a last method, we used a powerful analysis tool in the feature extraction stage, which is the DWT. The best performance was recorded with this technique. However, the experimental results showed that the choice of the mother wavelet, the order and the level of decomposition might be very difficult and no prior assumption over what is the best choice may be made before carrying out the experiment. In the classification stage, we used three different classifiers with each method, k-NN, SVM and ANN. With the DWT based method, k-NN had a better overall performance than SVM and was more stable to the wavelet, order and level changes.

The last step in our study was to separate the five rhythms from the EEG signals by filtering to see if we could use only one rhythm as input before the feature extraction stage instead of the whole signal. The results showed that the Delta rhythm, which has the lowest frequency band is enough for Epilepsy detection from *interictal* intervals. Whereas, the Beta rhythm had the best performance among the five rhythms for seizure detection. However, these findings do not go beyond the database used which is the *Bonn* database with an augmentation scheme, and the method used which is the DWT based method. In this study and all the previous research carried out about the current topic, the seizures are detected after their occurrence. In the future, it will be interesting to investigate these findings in order to build a forecasting model able to detect the seizures before their occurrence.

References

[1] Mayo clinic (n.d.) Epilepsy. Available at: https://www.mayoclinic.org/diseases-conditions/epilepsy/diagnosis-treatment/drc-20350098

[2] Gandhi, T., Panigrahi, B. K., & Anand, S. *"A comparative study of wavelet families for EEG signal classification"*. Neurocomputing, Vol 74(17), pp. 3051-3057, 2011. https://doi.org/10.1016/j.neucom.2011.04.029

[3] Ullah, I., Hussain, M., Qazi, E. and Aboalsamh, H. *"An Automated System for Epilepsy Detection using EEG Brain Signals based on Deep Learning Approach"*. Expert Systems with Applications, Vol.107, pp.61–71, Octobre 2018. http://dx.doi.org/10.1016/j.eswa.2018.04.021.

[4] Nicolaou, N. and Georgia, J. *"Detection of epileptic electroencephalogram based on permutation entropy and support vector machines"*. Expert Systems with Applications', Vol. 39(1), pp. 202-209, 2012. https://doi.org/10.1016%2Fj.eswa.2011.07.008

[5] Parvez, M.Z. and Paul, M. *"Features Extraction and Classification for Ictal and Interictal EEG Signals using EMD and DCT"*.15th International Conference on Computer and Information Technology (ICCIT), Chittagong, pp.132-137, Dec., 2012. http://dx.doi.org/10.1109/iccitechn.2012.6509719

[6] Bajaj, V. and Pachori, R.B. *"Epileptic seizure detection based on the instantaneous area of analytic intrinsic mode functions of EEG signals"*. Biomedical Engineering Letters, Vol.3(1), pp.17–21, 2013. https://doi.org/10.1109/iciinfs.2014.7036624

[7] Sharaf, Ahmed I., Mohamed, A. and El-Henawy, Ibrahim M. "*An Automated Approach for Epilepsy Detection Based on Tunable Q-Wavelet and Firefly Feature Selection Algorithm"*. International Journal of Biomedical Imaging. Vol.2018, pp.1–12, Sep.2018. https://doi.org/10.1155/2018/5812872

[8] Martis, R.J., Acharya, U.R., Tan, J.H. et al. *"Application of intrinsic time-scale decomposition (ITD) to EEG signals for automated seizure prediction"*. International Journal of Neural Systems. Vol.23(5), pp. 1557–1565, 2013. https://doi.org/10.1142/s0129065713500238

[9] Ahammad, N., Fathima, T., & Joseph, P. *"Detection of Epileptic Seizure Event and Onset Using EEG"*. BioMed Research International, pp.1–7, 2014. http://dx.doi.org/10.1155/2014/450573

[10] Juárez-Guerra, E., Alarcon-Aquino, V. & Gómez-Gil, P. *"Epilepsy Seizure Detection in EEG Signals Using Wavelet Transforms and Neural Networks"*. New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering, pp.261–269, 2014. http://dx.doi.org/10.1007/978-3-319-06764-3_33.

[11] Lasefr, Z., Ayyalasomayajula, S. and Elleithy, K. "*Epilepsy Seizure Detection Using EEG signals"*. IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), pp. 162-167, October 2017. https://doi.org/10.1109/uemcon.2017.8249018

[12] Peachap, A.B. and Tchiotsop D. (2019). *"Epileptic seizures detection based on some new Laguerre polynomial wavelets, artificial neural networks and support vector machines"*. Informatics in Medicine Unlocked, Vol. 16, pp. 100209, 2019. http://dx.doi.org/10.1016/j.imu.2019.100209

[13] Preprocessing - NeurotechEDU. Available at: http://learn.neurotechedu.com/preprocessing/

[14] Salomon, D. Data Compression. Springer Berlin Heidelberg.;2000. http://dx.doi.org/10.1007/978-3-642-86092-8

[15] Sripathi, D. *"Efficient Implementations of Discrete Wavelet Transforms Using FPGAs"*. Master thesis, Florida State University, Florida, 2003.

[16] Corinthios,M. *"Signals, Systems, Transforms, and Digital Signal Processing with MATLAB"*. CRC Press, September 3, 2018. https://doi.org/10.1201/9781315218533.

[17] Expert System. What is Machine Learning? A definition. Available at: https://expert system.com/machine-learning-definition.

[18] Cherifi D., Afoun L., Iloul Z., Boukerma B., Adjerid C., Boubchir L. and Nait-Ali A., *"Multi-class EEG Signal Classification for Epileptic Seizure Diagnosis"*. In: Artificial Intelligence; Renewables Towards an Energy Transition. ICAIRES 2020. Lecture Notes in Networks and Systems, vol 174, pp 635-645. Springer, Cham, 2021. https://link.springer.com/chapter/10.1007/978-3-030-63846-7_60#citeas

[19] Alaliyat, S. *"Video -based Fall Detection in Elderly's Houses"*. Master thesis, Gjøvik University College, 2008. Available at: https://www.researchgate.net/publication/26795394.

[20] Zhang, Zhongheng. *"Introduction to machine learning: k-nearest neighbors"*. Annals of translational medicine. Vol. 4, N11, June 2016. https://doi.org/10.21037/atm.2016.03.37.

[21] Kantardzic, M. *"Data Mining: Concepts, Models, Methods, and Algorithms"*. 3rd edition. New Jersey: IEEE Press, John Wiley & Sons, 672 Pages, November 2019.

[22] Pupale, R. *"Support Vector Machines(SVM)-An Overview"*. Available at: https://tow ardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.

[23] *Artificial neural network.*.Wikipedia, the free encyclopedia., Available at: https://en. wikipedia.org/wiki/Artificial_neural_network.

[24] Cherifi Dalila, El affifi Omar Badis, Boushaba Saddek and Nait-Ali Amine. *"Feature Level Fusion of Face and Voice Biometrics Systems Using Artificial Neural Network for Personal Recognition"*. An International Journal of Computing and Informatics (Informatica). Vol.44(1):85–96, March 2020. https://doi.org/10.31449/inf.v44i1.2596.

[25] Satapathy, Sandeep Kumar, Alok Kumar Jagadev, and Satchidananda Dehuri. *"An empirical analysis of different machine learning techniques for classification of EEG signal to detect epileptic seizure"*. An International Journal of Computing and Informatics (Informatica). Vol41(1): 99–110, 2017. https://www.informatica.si/index.php/informatica/article/view/838

[26] *Healthy-Epileptic EEG Data*, Department of Epileptology, University of Bonn, Germany. Available at: http://epileptologie-bonn.de/cms/front_content.php?idcat=193&lang=3&chang elang=3.

[27] Siuly S., Li Y., Zhang Y. *"EEG Signal Analysis and Classification, Techniques and Applications"*. Book Series title: Health Information Science. Publisher Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-47653-7.

https://doi.org/10.31449/inf.v46i4.3565 Informatica **46** (2022) 507–522 **507**

Parametrized MTree Clusterer for Weka

Marian Cristian Mihăescu, Marius Andrei Ciurez
Department of Computer Science and Information Technologies, Faculty of Automatics, Computers and Electronics, University of Craiova, Craiova, Romania
E-mail: cristian.mihaescu@edu.ucv.ro, mariusandrei.ciurez@gmail.com

Keywords: Clustering, MTrees, metric spaces

Received: May 25, 2021

In the area of clustering, proposing or improving new algorithms represents a challenging task due to an already existing well-established list of algorithms and various implementations that allow rapid evaluation against tasks on publicly available datasets. In this work, we present an improved version of the MTree clustering algorithm implemented within Weka workbench. The algorithmic approach starts from classical metric spaces and integrates parametrised business logic for finding the optimal number of clusters, choosing the division policy and other characteristics. The result is a versatile data structure that may be used in clustering to find the optimal number of clusters, but mainly for loading data sets that already have a known structure. Experimental results show that MTree finds the proper structure in two clustering tasks, although other algorithms fail in various ways. A discussion of further improvements and experiments on real data sets and functions is included.

Povzetek: Opisana je nova metoda algoritma MTree za gručenje.

1 Introduction

As an unsupervised learning technique, clustering is continuously getting attention within the machine learning area due to many available algorithms and a wide range of application domains for which practical solutions are continually being designed and implemented.

The general problem of building clusters of objects is narrowed down in [41] by clearly stating that the first thing that needs to be taken into consideration in the context of the problem. Therefore, building a general-purpose clustering algorithm that may work with any objects and solve any task is not a realistic option. Thus, the objects we need to define and provide a proper task description accompanied by particular distance and evaluation metrics or various algorithmic approaches need to be carefully taken care of in building a clustering data analysis workflows.

In [1], authors recently raised the problem of clusterability of a dataset. Having an available dataset does not necessarily imply that we may meaningfully run a clustering process to solve a particular task. Thus, defining clusterability becomes a critical issue. Checking if a dataset is clusterable becomes an inherently tricky problem, especially when dealing with actual data and solving a particular practical task.

The wide range of approaches used in available implemented clustering algorithms has found various application domains to provide efficient solutions to many practical tackled problems. From the many application domains, we mention medical image processing (i.e., pattern recognition and image segmentation) [39] [24], general and natural language processing knowledge discovery [20] [33] [34], navigation of robots [14] [22] and in many other contexts.

In the area of unsupervised learning, there are several general classes of clustering algorithms (i.e., flat, hierarchical and density-based) that all share two common problems: finding the optimal number of clusters and quickly and efficiently finding the correct clusters taking into consideration specific distance measures appropriate for the objects (i.e., pixels, points, persons, books, etc.) that are being grouped. Further, once the clusters are being correctly determined, there may be later used to query for the nearest neighbours or run specific range queries. Depending upon the inner data structures used for managing the clusters when tree data structures are being used efficiently, searching or traversing may be accomplished efficiently

The objective of this work is to present an improved version of the metric trees (MTree) algorithm that has been firstly proposed by [7] and later by [40] and [43].The first proposal of using MTrees in the context of clustering has been made in [31] and later implemented as Weka [16] package in [30]. This paper presents an improved version of the works from [30] and [31] that has been tested in a comparative benchmark with k-MS morphological reconstruction clustering algorithm [37] along with classical algorithms (i.e., simple k-means, Cobweb, Farther First and Canopy) in [9].

This paper presents a further improved sparametrised version of the MTree algorithm in terms of managed items, used distance, the method for finding and setting K (i.e., the number of clusters), division policy and validation metrics. The critical improvement of the new parametrised version of the MTree clusterer is that it is now suitable to address a broader range of problems. The parametrised version is

not ideal for solving any clustering problem by dynamically choosing the suitable parameters. Instead, a wide range of clustering problems may be addressed to the newly proposed MTree cluster by properly setting its parameters depending on the available data objects and the tackled task.

In general, there are two types of clustering problems. One regards finding patterns in a dataset that we do not know if they have a specific structure or if a certain number of clusters exist. The second type of task regards correctly building a data clustering model that may later be queried many times for getting specific information about managed data. In this scenario, the cluster model is constructed only once such that it may be regarded as a preprocessing step. Later, few insertions and updates may occur at runtime, while most calls are range queries or kNN queries that need to be solved correctly and efficiently. The proposed approach is suited for both tasks, but the second one is more appropriate. Range queries determine items whose distance from a specific query item is smaller than a particular value.

The paper is organised as follows. In Section 2, we perform a literature review with regards to similar libraries other application domain usages of metric trees for clustering such as time series analysis of cytometry data, recommender systems, spatial clustering data, automatic computing the number of clusters for colour or greyscale image segmentation and clustering quality metrics. Section 3 describes the proposed approach with a detailed presentation of how each parameter may be set and how it influences the business logic of the MTree. Section 4 presents the experimental results on two publicly available data sets with runs on several parameter settings. Finally, section 5 contains the conclusions of this work, summarises the main contributions and discusses potential improvements and applications.

2 Related work, limitations and approaches

Data clustering (also known as unsupervised learning) represents a subarea of machine learning. Other areas of machine learning are supervised learning, reinforcement learning or deep learning. A particular sub-domain is represented by age clustering and segmentation [13] which presents the use of subtractive clustering along with classical K-means algorithm to preprocess the data for optimal centroid initialisation. The experimental results were obtained on medical images representing infected blood cells with malaria. The classical images used for segmentation bring better results than k-means taking into account root mean square error (RMSE) and peak signal-to-noise ratio (PSNR) metrics.

Another approach for image clustering was proposed by Chang et al. in [5]. They propose a Deep Adaptive Clustering (DAC) approach that reduces to a classification problem in which similarity is determined by cosine distance and learned labelled features tend to be one-hot vectors obtaining good results on popular and accessible datasets like MNIST, CIFAR-10 and STL-10.

One of the critical practical usages of clustering algorithms is for image segmentation. Including spatial information along with taking outlier points at a later stage in the clustering algorithm has been proposed in [42]. From this perspective, outliers are data points with almost equal distance to their adjacent clusters and therefore should be taken into consideration later. This approach raises two issues. One regards the fact that the order in which data points are given to the algorithm has significant importance. Thus, if a clustering algorithm is highly sensitive to the order in which data points are provided with a custom preprocessing may be necessary. The second issue regards the very nature of the data set from the clusterability perspective. The critical verification that is also highly recommended is to always check for clusterability before starting to the clustering analysis.

Another application of clustering is grey scale image segmentation [44]. As compared with the clustering of colour images or with images that incorporate spatial information, the task is to highly decrease the time complexity of the algorithm by using affinity propagation (AP) clustering algorithm. As always, when real life grey scale images are being clustered the problem of correctly determining the number of clusters or segments is a critical one.

More elaborate applications regard indexing and retrieval of similar images from an image database (CBIR - Content-Based Image Retrieval) which represent a challenging task that has been addressed in [29] and [27]. The first approach uses features, colour and texture. It employs K-means and hierarchical clustering for finding the most similar images. The second approach uses colour, texture and shape as features and K-means as business logic for determining four classes of images: dinosaurs, flowers, buses and elephants. The obtained experimental results are promising in terms of excellent precision and recall values.

Clustering has also been used successfully in recommender systems that were based on collaborative filtering in [11]. A novel K-medoids clustering recommendation algorithm has been proposed, which introduced an improved Kullback-Leibler divergence for computing item similarity. The final task was to improve the effectiveness of the developed recommendation system.

Lately, in the application domain of immunology has been used clustering algorithms - *ChronoClust*, a new density-based clustering algorithm - on time series cytometry data [26]. The task was to characterise the immune response to disease by tracking temporal evolution.

Very recent works [1] put a high emphasis on defining clusterability and checking if a dataset is clusterable as a preprocessing step before any other data analysis is further performed. Therefore, before applying the algorithm for solving a task that requires clustering a *sanity check* may be required, in the way that we should verify the clusterability of the dataset. In other words, clustering may not work

on datasets which do not exhibit any internal structure, irrespective of any particular algorithm that may be employed. In our case, clusterability becomes a prerequisite that the dataset needs to meet before being loaded into the MTree structure. The new proposed clustering algorithm should have as main scenario working with data that is known to be clusterable, for which we know it has a well-defined structure with a known number of clusters. A dataset has a well-defined structure when there exists an assignment of items into clusters that is validated by a domain specialist for real world datasets. In the case of synthetic datasets, the function that creates the instances is designed in such a way that clusters are well-defined and represent the gold labeling for any clustering algorithm.

If the number of clusters is not known than the results highly depend on the particular practical context of the clustering problem. The context is defined by the problem (clustered objects and clustering task) and parameters of the MTree: object type, distance function, splitting policy and validation metrics. If the dataset is not clusterable, we argue that the MTree algorithm - as well as any other clustering algorithms - will exhibit undefined behaviour.

A more complex context occurs when the image source is unknown or when the ground truth for the training dataset is also unknown [4] [10]. In this particular situation, finding the optimal K represents am an inherently difficult task. From an experimental algorithmic perspective, using proper distance metrics and loss function in this optimisation problem represents one of the key ingredients towards successful results. Such approaches propose fancy solutions such as hierarchical clustering or clustering ensembles based graph partitioning methods, cluster-based Similarity Partitioning Algorithm (CSPA), HyperGraph Partitioning Algorithm (HGPA), and Meta Clustering Algorithm (MCLA).

Among the most well-known issues in unsupervised learning consists in determining the actual number of clusters from a dataset. Unfortunately, scenarios in which the value of K is known to occur only in a few practical systems. In general, an application that performs an image processing task does not have any information regarding the actual number of clusters from the target image. This situation may occur when dealing with data streams [25] or with very high-dimensional datasets [17]. In general, one of the most suitable approaches tries to reduce to automatic determination of K that may be based on dynamic clustering [17] or joint tracking segmentation [32].

A general-purpose algorithm for finding the optimal number of clusters has been proposed by [6] and implemented in an R package in NbClust. The main idea of this approach is that it may use up to 30 indices for voting the number of clusters. The package has implemented a function to run a clustering algorithm (i.e., k-means or hierarchical clustering) using various distance measures and aggregation methods. The main limitation of the approach is that it is general and practical usage for particular datasets needs to be parametrised by the appropriate clustering algorithm, subset of indices, distance measure and aggregation method.

One particular usage of clustering regards automatic computing of the number of clusters for colour image segmentation [21]. This approach uses fuzzy c-means algorithms for extracting chromaticity features of colours and trains a Neural-Networks with obtained chromatic data of colours. The trained model may be further used on new colour images to predict the number of clusters in colour images.

Among many clustering libraries, we mention LEAC [36]. It is an open-source library with source code publicly available in the GitHub repository. Thus, once the experiments are also performed on publicly available datasets, the results may be reproducible and also used in other setups for further improvements. Inclusion of 23 state-of-the-art Evolutionary Algorithms for partial clustering within a library that allows easy and fast development and integration of new clustering algorithm represents the solution that we also target when improving the initial version of the MTree clustering algorithm within Weka package.

Another usage of metric trees has been reported recently in [12]. The task is to quickly and efficiently scale-up the problem of shadow rendering for 70 million objects (i.e., triangles) in real-time. The proposed metric tree uses as splitting policy binary space partitioning (BSP) and ternary object partitioning (TOP) for grouping triangles into clusters as precomputed bounding capsules (line-swept spheres).

Finally, the whole clustering process needs validation, and many quality metrics may accomplish this for a wide range of algorithms [18]. Depending on the structure of the dataset, various clustering quality frameworks [23], [38] have been proposed. The key issue that always arises regards choosing the proper similarity and quality metrics [38].

3 Proposed approach

The proposed MTree parametrised clusterer has been firstly proposed in [31] in an attempt to define a new clustering algorithm that has as main business logic the metric trees initially presented in [7] and later in [40] and [43]. The initial C++ implementation approach was designed to cluster students who were defined by three of their obtained grades during one semester. The main shortcoming of the proposed structure was that it as designed only for managing student objects and had several hard-coded parameters needed for building the tree. The most important one is *nrKeys*, which represents the maximum number of students contained in a node (i.e., a cluster). This approach has a critical limitation in the fact that the *splitNote()* method was called based only when a note was full. Other limitations regard the lack of parametrised division policy, distance metrics or other features needed for flexibly running a clustering process.

Later, the initially proposed MTree clustering algorithm has been contributed as an official Weka package [30]. This newly Java-based approach had the goal to be functionally available under Weka workbench as any other clusterer, such that it may be further used in various practical situations. The MTree clusterer from Weka has been used in [37] in a comparative analysis on publicly available images. The results of MTree were inferior such that the limitations were addressed in [9]. As critical improvements, the MTree version from [9] uses off-line dataset preprocessing for finding the optimal number of clusters and adjusts the business logic of the clusterer in terms of division policy and distance metric between instances.

The current proposed version of the MTree cluster represents a flexible parametrised version of the former one in terms of division policy, used distance, the method for finding and setting the number of clusters.

3.1 Definitions and context

The metric space $M = (D,d)$ on a data domain D with the distance function $d : D \times D \to \mathbb{R}$ postulates:

$$Non\ \ negativity : \forall x,y \in D, d(x,y) \geq 0 \tag{1a}$$

$$Symmetry : \forall x,y \in D, d(x,y) = d(y,x) \tag{1b}$$

$$Identity : \forall x,y \in D, x = y \Leftrightarrow d(y,x) = 0 \tag{1c}$$

$$Triangle\ \ inequality : \forall x,y,z \in D, d(x,z) \leq d(x,y) + d(y,z) \tag{1d}$$

The conditions specified above are satisfied by most discrete or continuous distance (or similarity) metrics (or measures): Euclidean, Minkowski, Manhattan, quadratic form distance (i.e., colour histograms, weighted Euclidean distance), edit distance, Jaccard's coefficient or Hausdorff distance. Building clusters of objects in metric spaces relies on the partitioning method and shape of the decision boundary that lays between two adjacent clusters. The partitioning method regards how a set of objects is split into two or more clusters taking into account specific optimisation criteria such as the sum of squared errors (SSE) in case of simple k-means algorithm. Regarding the decision boundary, the two most common options are ball partitioning as in metric spaces and hyperplane partitioning.

As current implementation of the MTree algorithm represents a two-level ball decomposition generalisation of the approach from [40]. The limitations from [40] regard choosing an arbitrary object as the pivot, using only binary splits around the median object, which implies previously sorting the objects and multilayered approach due to recursive construction. From the practical perspective of running a clustering process, these are substantial limitations because ordering may not always be possible, binary splitting may not be useful when dataset consists of many clusters and the notion of the cluster become unclear in a multilayered approach.

Definition 1. The newly designed MTree data structure is a two-level perfectly balanced multiway tree that indexes a set of objects into its leaves which reside only on the second level. After building the tree, the root contains the set of centroids and their corresponding covered radius. On the second level, the leaves represent the clusters containing objects whose distance is smaller or equal to the radius assigned of the corresponding centroid within the root.

The key features for the parametrised MTree implementation are:

1. The possibility to process various object data types provided as input (i.e., image, document, etc.) that is represented as a multidimensional vector.

2. The possibility to set up a particular distance function between objects by direct usage of distance functions that are already available in Weka or by using a newly defined custom function.

3. The possibility to set up a particular division policy that will be used internally used as needed. Practically, the logic of the division policy is managed by the clustering algorithm. This feature needs to be accompanied by a parameter that controls the number of clusters into which full leaf may be splitted. Available options are binary object partitioning (BOP), ternary object partitioning (TOP) or multi-object partitioning (POS).

4. The possibility to set up a specific number of clusters in which the entire dataset will be partitioned, if the number of clusters is known. If the number of clusters is not known, the MTree will find the optimal number of clusters considering the other parameters that have been set.

5. The possibility to compute at request various clustering quality metrics that will give a general idea on the quality of the clustering process. This feature is critical in benchmarking the MTree clustering results against results produced by other clustering algorithms.

The MTree uses only one node data structure for the roots and leaves. In the root node, the instances are represented by centroids, and each element from the *radix* vector represents the covered radius for the corresponding centroid. In the same way, each element from the route vector is an address of the corresponding leaf node. Their vector position accomplishes the correspondence between centroids from the root node and covered radius and leaves. For example, the first element of the instances vector of the root represents the first centroid with a radius defined in the first element of the *radix* vector. The first element of the route vector represents the address of the leaf node that contains objects whose distance to the first centroid is smaller or equal to the covered radius. In the case of leaf nodes, the instances represent the data objects themselves. The *isLeaf* flag is set by the business logic to value *TRUE* such that radix and route vectors are set to null.

Table 1: The structure of MTree node

Field name	Description
nrKeys	The number of objects actually stored in the node.
isLeaf	This flag represents the node type: root or leaf.
radixes	The vector of distances covered by a centroid.
routes	The vector of node addresses to leaf nodes.
instances	The objects from the node (parent or leaf).
parent	The address of the parent node.

Before starting any computation needed for inserting a new object in the MTree, the algorithm checks if we are in a leaf and if the leaf has objects in it. If any of these conditions do not hold than insertion may not take place because insertion may be performed only in a leaf and further splitting may be taken into consideration only if the leaf has objects in it. Further, the leaf is evaluated for splitting parametrised by *evaluatorOfK*, which is a function that determines the optimal number of clusters. This helper function takes as parameter the objects from *treeNode* and outputs *splitEval* as an evaluation of the splitting. If this also is valid, then the leaf node is split into the optimal number of clusters by using a parametrised *divisionPolicy*. The *insertNonFull* function is called to append a new object in the leaf when no splitting is necessary.

The two main ingredients of *mTreeInsert* function are the evaluation of the optimal number of clusters from a leaf and the division policy used for splitting. The current implementation uses a function called *voteK* that computes the optimal number of clusters in a similar way as *NbClust* [6] package in R. The main difference is that *NbClust* uses 30 indices while *voteK* currently integrates only 8 indices: Davies-Bouldin, Dunn, Xi-Beni, Banfeld-Raftery, McClain-Rao, Ray-Turi, Calinski-Harabasz and PBM indices. The architectural design of the package allows the easy call of any method that given an input dataset can compute the optimal number of clusters.

The *nrKeys* represents the number of items (i.e., centroids or items) contained in a node (i.e., root or leaf). For the root node, the items are centroids, and for the leaf node, the items are the sample points. Depending upon the implementation, the centroids may be items from the dataset or computed instances. The *isLeaf* field from the data structure represents a clarification regarding what represent the items from a node: centroids or instances.

As far as our M-tree is concerned, we pay extra attention to the nodes because it is essential whether they are leaves (i.e., terminal nodes containing instances) or internal nodes (i.e., having only centroids) information is critical for the algorithm design.

The instances vector contains the centroid points or items, depending upon the node is either root or leaf. If the node is the root, vector radixes contain the distance covered by each centroid, while the routes vector contains the leaves' addresses. On the other hand, if the node is a leaf, then the field parent includes the root address while the radixes and routes vectors are empty.

The second key ingredient of the *mTreeInsert* function is the division policy. The default option for this parameter is the simple k-means algorithm, but any other strategy may be called as needed. Other options, besides calling particular clustering algorithms as a multi-object partitioning (MOP) strategy, is using binary space partitioning (BSP) or ternary object partitioning (TOP) by simply setting proper parametrised values. Suppose there is no need for node splitting, then insertion reduces to appending the new object into that leaf.

In that case, insertion reduces to appending the new object into that leaf, which is accomplished by calling *insertNonFull* function. Intuitively, when a leaf is not full, we insert a new object; otherwise, we split the leaf. The most crucial difference from former versions or other approaches is that splitting is not called when the number of stored objects reaches a specific value, but when current objects exhibit the property that they are properly clustered, and therefore a split is compulsory.

Algorithm 1 summarizes the business logic for inserting a new object into an existing MTree leaf node.

Algorithm 1 summarises the business logic for inserting a new object into an existing MTree leaf node. The second critical procedure is the one that performs the split of a leaf node as specified in the insertion algorithm. The main ingredient is represented by the *clusteringEvaluator* setup, which allows breaking the leaf into clusters according to with specific division policy and a predetermined optimal number of clusters. The newly obtained clusters are added into a *splitedClusters* vector and returned as output. The main improvement in the split procedure is avoiding splitting a leaf into a hard-coded number of clusters by using a pre-determined optimal number of clusters and parametrising for the division policy for running the effective split. Algorithm 2 summarises the business logic for splitting. The input consists of a tree node (i.e., a cluster) and an evaluator, and the returned value consists of a vector of clusters, called splittedClusters. The evaluator has the task of deciding if the node should remain as a single cluster or be split in two or more clusters. If the evaluator considers more than one cluster in the note, the node will split, and the tree will change its structure.

The bounding volumes of the MTree leaves are circles if objects are 2-dimensional or spheres if objects are 3-dimensional and Euclidean distance is being used. For

Algorithm 1 mTreeInsert (MTree treeNode, Instance newObject)

```
1: if treeNode is leaf and has objects then
2:    Set clusteringEvaluator = getOptimalNrOfClusters (treeNode, evaluatorOfK);
3: end if
4: if splitEval is valid then
5:    nrOfClusters = clusterEval.getNumberOfClusters ();
6:    if nrOfClusters > 1 then
7:       splitNode(treeNode, clusteringEvaluator);
8:    else
9:       insertNonFull(treeNode, newObject);
10:   end if
11: end if
```

Algorithm 2 mTreeSplitNode (MTree treeNode, ClusterEvaluation clusteringEvaluator)

```
1: clusters = getClusters (treeNode, clusteringEvaluator);
2: if size of clusters > 0 then
3:    for each cluster in clusters do
4:       centroidInstance = chooseCenter(cluster);
5:       splitedCluster.addCentroid(centroidInstance.getCentroid());
6:       splitedCluster.setRadix(centroidInstance.getRadix());
7:       splitedClusters.add(splitedCluster);
8:    end for
9: end if
10: Return splitedClusters;
```

n-dimensional spaces, the bounding volumes are hyperspheres, a generalisation for a set of points equally distanced to a given point. This generalisation is valid only for Euclidean distance, as for other distances, the bounding volume is the representative of a sphere for that particular vector space.

3.2 MTree algorithms description

The MTree implementation is aimed for usage by client code in practical scenarios. Although the primary usage scenario addresses the situation for which we know the actual number of clusters, the clusterer may also be used in the case when the data analyst specifies a particular number of clusters depending on the tackled task or in the situation when the number of clusters is not known or does not exist.

In any of the situations mentioned above, the parameters that need to be set are the input data-set, the number of clusters, the distance metric, the evaluator of the number of clusters and the division policy algorithm. The main practical goal of the current version of the MTree clusterer is to verify that it correctly loads the data given specific parameters and creates a data model that validates the ground-truth model from two publicly available data sets. Finally, the clustering validation metrics are verified against other clustering algorithms implemented in Weka workbench.

One key aspect for correctly building the MTree from data regards the order in which the data objects are provided as input. As a general rule, any clustering algorithm is highly sensitive to the order in which data objects are considered in the clustering process. The seed mechanism is the general solution to clustering algorithms and is also used as a standard option in the MTree. For our case, the seed mechanism represents a random shuffle of the order in which the instances are added to the MTree.

The pseudocode of mTreeInsert function presents the structure and logic of operations when inserting a new instance into the tree. The function's parameters are the address of the tree and the item that should be inserted. Since inserting takes place only in leaf nodes, the algorithm first checks that we are in a leaf node and then determines the optimal number of clusters from that leaf. If more than one cluster is found in the leaf, the node is split, and the instance is placed into the appropriate cluster.

The pseudocode of mTreeSplitNode function actually performs the split of a leaf. Along the treeNode, an evaluator is given as parameter, which gathers the parameters needed to determine the clusters. Once the clusters are determined, the centroids and corresponding radixes are placed in the root, the addresses of the new clusters are placed in the routes and instances themselves are placed in the appropriate clusters.

A particular situation occurs when we do not know the correct number of clusters. In this case, the adequate solution is to preprocess the data-set to check clusterability and determine the valid number of clusters if such a number exists. The situation in which the data-set has not a clear well-defined structure may be interpreted in two ways: either the data-set has several possible options as the actual number of clusters, or we do not know the correct number of

clusters. In the first situation, a domain knowledge person should consider the specific data analysis task and choose the correct number of clusters that fit the practical problem. More extensive work needs to be done as a preprocessing step in the latter situation.

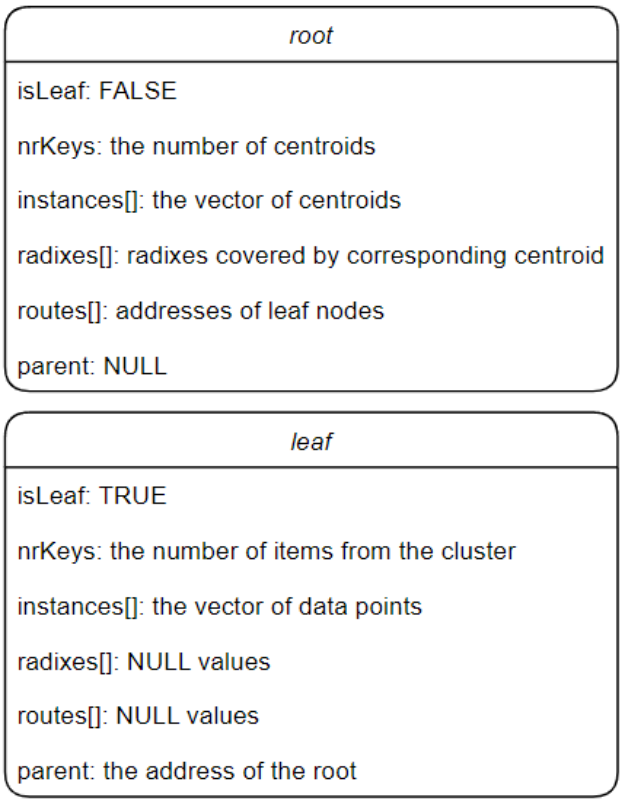


Figure 1: Nodes structure

As a general rule, when the data-set has no structure, the first preprocessing steps should define the number of clusters as a task-dependent value. Then, the centroids should also be defined as representatives for each cluster considered by the domain knowledge person. Finally, the objects from the data-set are added to the pertaining cluster only if a maximal threshold distance from the centroid is not exceeded. The objects that have almost equal distance from centroids are considered outliers and are not added to any cluster. In this way, the data analyst may obtain a clusterable data set with a specific number of clusters. Having a clusterable data-set is a prerequisite for the MTree clustering algorithm and any other algorithm. Therefore, if the data-set does not meet clusterability, building an MTree clusterer or any other clusterer will exhibit undefined behaviour.

Finding the correct number of clusters by using MTree may be performed by running with various parameter settings in terms of the number of clusters and division policy in an attempt to obtain a clusterer whose validation parameters show that the correct patterns have been discovered. In this use case, the MTree data structure is built for finding whatever clusters are to be found.

3.3 Complexity analysis

The complexity analysis of building the MTree from data depends on the number of objects, the number of clusters, the method of finding the optimal number of clusters and the number of distance computations. The number of clusters from the data-set represents the number of splits that need to be performed while building the tree. The most critical operation is finding the optimal number of clusters, which is called after each object insertion. The number of distance computations is related to the number of clusters since distances from the newly inserted object to all the centroids from the root need to be computed to determine the suitable leaf where the insertion should occur. The most time-consuming function is the *getOptimalNrOfClusters* function that is called whenever a new object is to be inserted. We have observed that for a reasonably small number of clusters and a large number of objects, that method *getOptimalNrOfClusters* is used to trigger a split fewer times than the number of clusters. For example, once the number of leaves from MTree has reached the true number of clusters, then looking for the optimal number of clusters becomes useless. Further insertions will be performed in constant time just by determining the proper leaf where the new object needs to be inserted. As stated in [15] the performance of building an MTree with n objects is analogous to that of k-d trees, that is $O(n \log n)$ for worst-case scenario. Depending on the split method the time may increase to $O(n \log^2 n)$ or $O(kn \log n)$ for k dimensions. Still, the currently proposed method is highly sensitive to the order in which objects are being inserted, the seed selection and the particular parameters setup as well as all other clustering algorithms.

The critical property of the MTree is that after correctly building the clusters, the operations of inserting, removing and querying may be performed in $O(\log n)$ time. These aspects are not tested by the current works and need to be further experimentally investigated in practical clustering tasks.

4 Experimental results

4.1 Data-sets description

Experimental results have been performed on two synthetic publicly available data sets from the clustering basic benchmark [19]: *Unbalance* [35] and *Dim2* [28]. *Unbalance* is a synthetic 2-d data-set with 6500 points and 8 Gaussian clusters for which ground truth centroids and partitions are known. Dim2 is also a synthetic 2-d data-set with 1351 points and 9 Gaussian clusters.. Figures 4 and 5 present a plot of the raw input data.

We have chosen two synthetic datasets for which the ground truth centroids and partitions are known because they are suitable for comparing clustering results obtained by MTree algorithm against other ones implemented in Weka workbench.

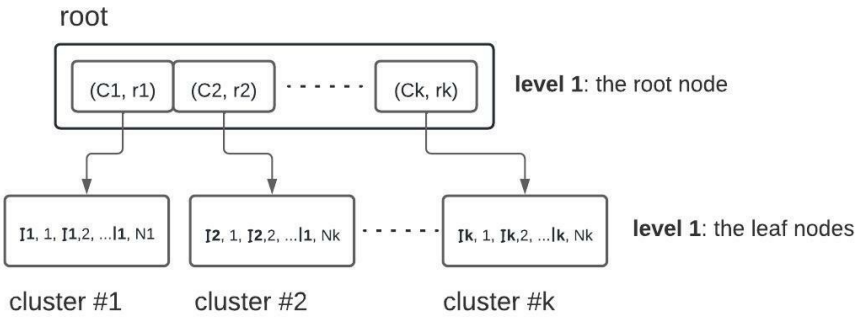


Figure 2: Sample MTree

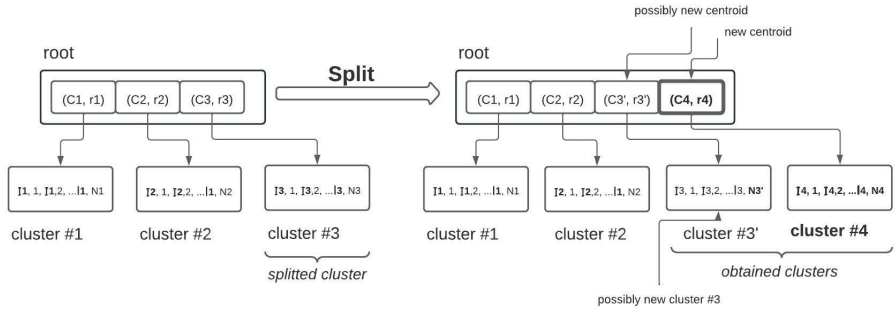


Figure 3: Sample MTree split node

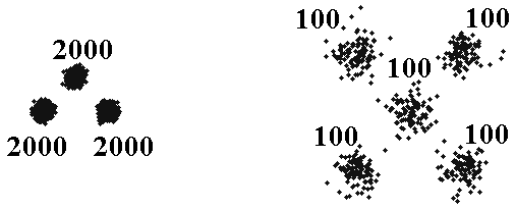


Figure 4: Unbalance dataset

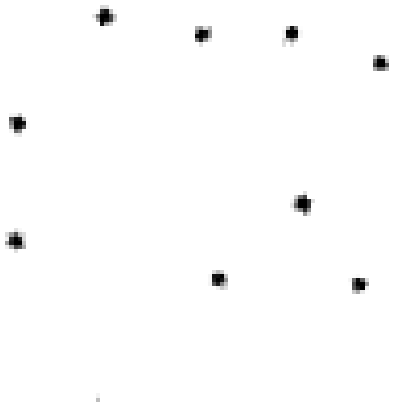


Figure 5: Dim2 dataset

Finally, we have tested the MTree on *Wine* and *Iris* classical datasets from UCI Machine Learning Repository [3]. These real-world datasets were chosen because they may be successfully used in clustering tasks as they are labelled, and classical unsupervised training may correctly determine the classes.

4.2 MTree package description

MTree is implemented in Java and is aimed to be executed within Weka 3.8 workbench by using the *Package manager* tool. The MTree package is based on three classes *Node*, *MTreeBean* and *MTree* along with other three helper classes. Figure 6 shows the software architecture the MTree package as an UML class diagram .

The class *Node* represents a blueprint for a cluster of instances and the *MTreeBean* class contains the root of the MTree. The most important class is *MTree*, which extends *RandomizableClusterer*, which is an abstract class whose direct subclasses are Canopy, Cobweb, FarthestFirst and SimpleKMeans. Further, by implementing the *NumberOfClustersRequestable* and other interfaces, the MTree gets the possibility to be parametrised similarly as other clustering algorithms are in Weka.

The main goal was to obtain a clusterer that may be parameterisable in the same way as already existing clusterers based on interfaces that are already defined in Weka but also offering the possibility of defining new interfaces specific for parameters needed by MTree algorithm. In this way, the

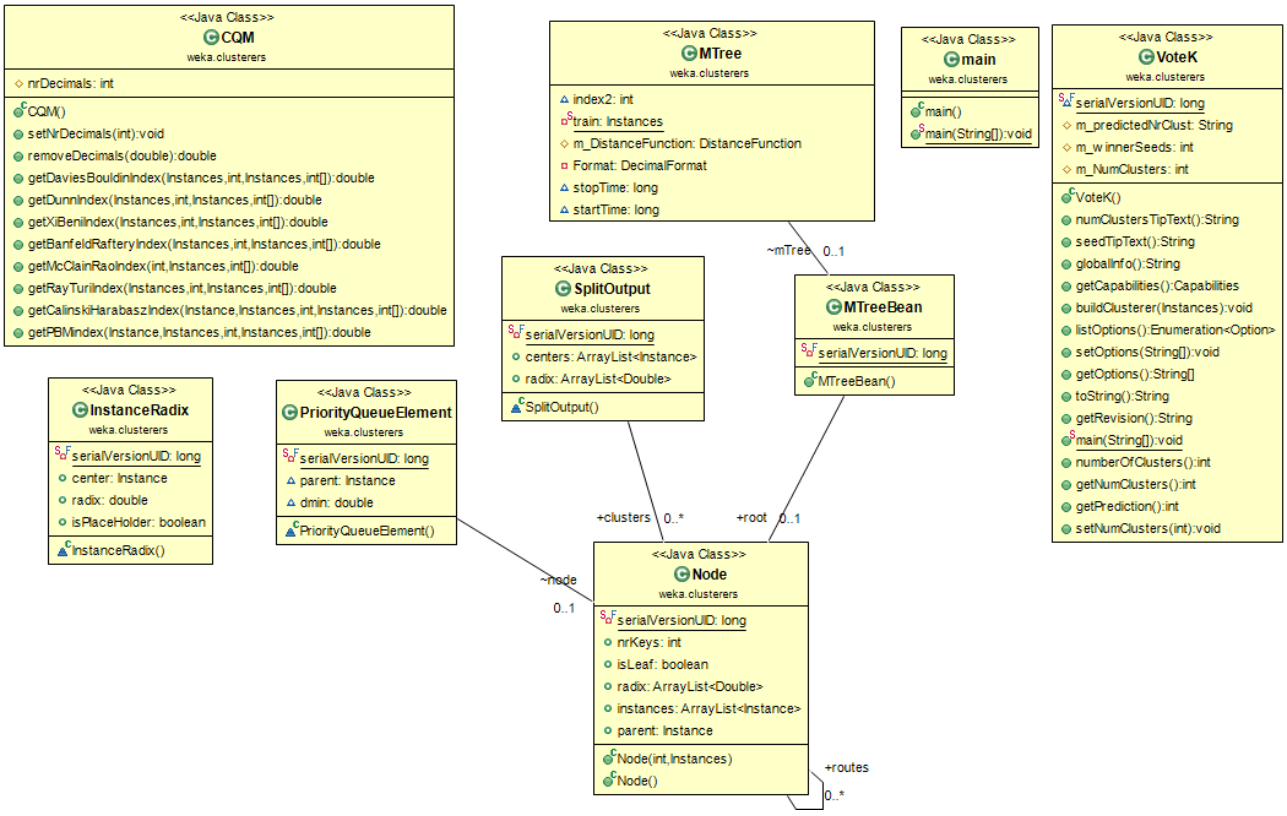


Figure 6: UML class diagram for MTree package

newly obtained MTree can be easily parametrised in the usage of the command-line interface or Weka GUI interface.

4.3 Sample MTree usage

The MTree can be used in three ways. The current implementation provides flexibility for running full experimental runs and benchmarking the performance of an MTree parameter configuration against already existing Weka clustering algorithms.

– *Basic mode, through command line*. This mode allows executing the MTree on any machine that has JVM 1.8 and *weka.jar* version 3.8.3. Figure **??** presents a sample command line execution of the MTree algorithm. This approach is commonly used when batch execution is needed only once for building the clusters and serialising the obtained model (i.e., the distribution of objects into clusters) to persistent storage (i.e., csv, xml, etc.) for later usage is rather tricky.

  The available options when running MTree in command line interface are further presented. ***N*** represents the number of clusters. If we want MTree to decide for itself the number of clusters this option must be set to value -1. ***init*** option may be used for setting the initialization method. ***I*** option is for setting the maximum number of iterations, ***O*** for preserving the order of instances, ***S*** for setting the number of seeds, ***d*** for setting the distance metric, ***findN*** for setting the method for finding the optimal number of clusters and ***splitPolicy*** for setting the method used as splitting policy. Current implementation may use as splitting policy Canopy, Simple k-means, CobWeb, FarthestFirst or HierarchicalClusterer clusterers from Weka.

– *Using the Weka GUI*. This mode is the most user-friendly as the MTree may be used from Weka GUI as any other clustering algorithm. As the MTree package is in the list of official packages, it needs to be installed before usage. Installing the MTree package in Weka is a straightforward procedure as the *MTree.zip* archive is publicly available in SourceForge [30] and the link to the package is available in the list of official packages within the Weka package manager tool.

– *Programmatic way*. The most versatile usage of the MTree is programmatically. This approach allows setting up the parameters at runtime as well as having a ready-to-use in memory MTree object that is ready for querying. This approach allows the usage of the cluster as business logic on a server side in complex applications where client code is performing various queries. Sample code for building the MTree from data is publicly available in [8].

```
java weka.clusterers.MTree -t unbalance.arff \
- N 8 -init 1 -I 1000 -O -S 100 -d 1 -findN weka.clusterers.voteK -splitPolicy 0
```

Figure 7: Command line execution of MTree

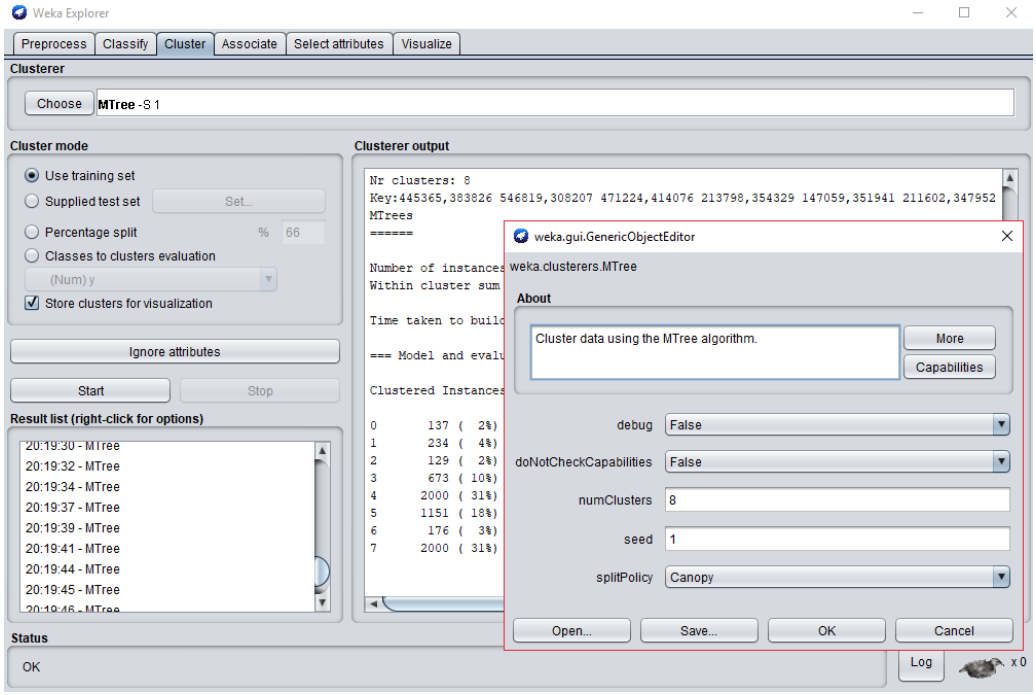


Figure 8: Execution of MTree in Weka GUI

4.4 Sample runs on Unbalance and Dim2 synthetic data-sets

The newly released parametrised MTree implementation has been tested against two publicly available synthetic data sets: *Unbalance* and *Dim2*. The client code that calls the MTree package is publicly available at [8], as further presented experimental results were obtained by programmatically running the MTree implementation.

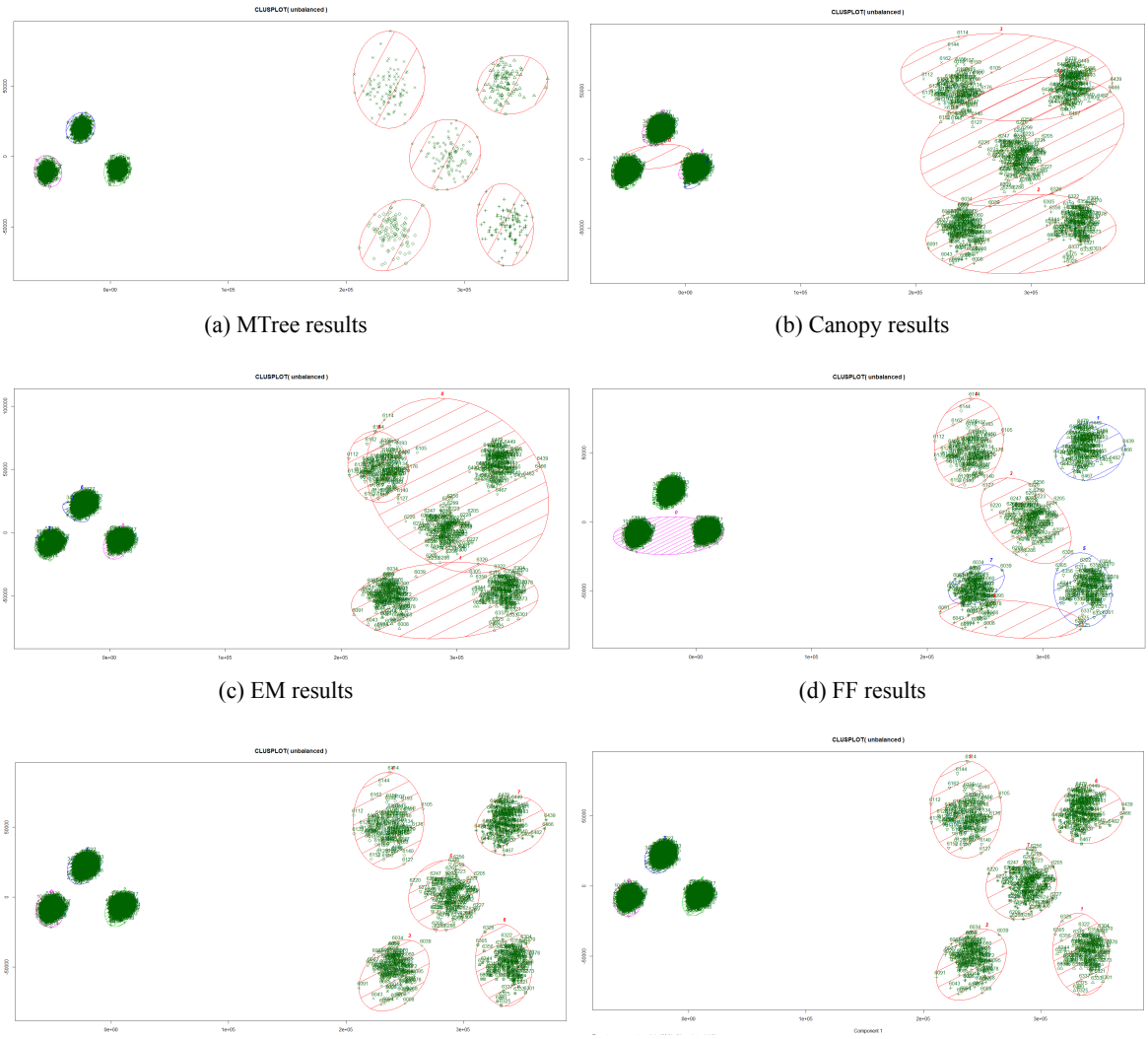
Figures 9 and 10 present the experimental results of running the MTree clusters along other five clusterers implemented in Weka workbench: Canopy, EM (expectation-maximization), FF (Farthest First), HC (Hierarchical Clustering) and SKM (Simple KMeans).

As figure 9 clearly shows, the MTree correctly determines the clusters by using 100 seeds and Canopy for split policy. As a general rule, the clustering result exhibits undefined behaviour regarding the number of seeds, such that correct results may be obtained sometimes for only 10 seeds and sometimes for 1000 seeds. Here is a summary of the experimental results of the other five algorithms:

- *Canopy* algorithm fails to determine the correct number of clusters and the found distribution into clusters is wrong.
- *EM* algorithm fails to determine the correct number of clusters although in many situations it is used for this task as it does not require the value of K. The obtained clusters are reasonable fine with two exceptions: cluster 0 puts together three real clusters and cluster 1 puts together two real clusters.
- *FF* algorithm correctly determines the number of clusters but misses to determine two of them correctly: cluster 0 puts together two real clusters and cluster 2 is composed of objects belonging to two distinct clusters.
- *HC* algorithm correctly solves the task.
- *SKM* algorithm correctly solves the task after finetuning the parameters: 100 seeds, usage of kmeans++ [2] for seed optimisation and maximum 10,000 iterations.

As figure 10 clearly shows, the MTree correctly determines the clusters on a more difficult task by using only 10 seeds. The other investigated algorithms provide the following results:

- *Canopy* algorithm fails to determine the correct number of clusters and the found distribution into clusters is wrong, as it puts together in a cluster objects from two clusters.
- *EM* algorithm correctly determines the clusters.
- *FF* algorithm fails to determine the correct number of clusters and misses to determine one of them correctly.



(a) MTree results (b) Canopy results

(c) EM results (d) FF results

(e) HC results (f) SKM results

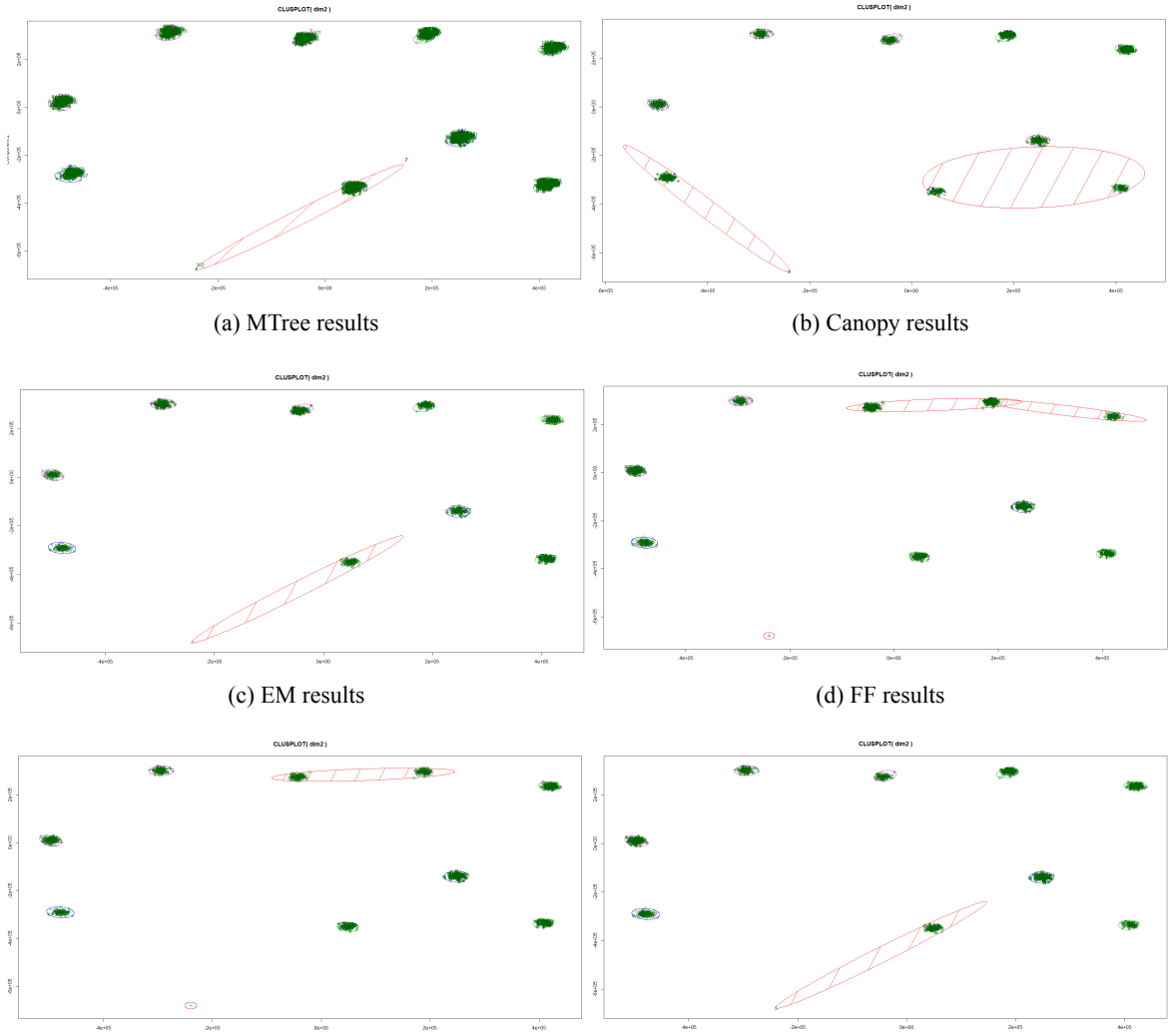
Figure 9: Clustering results on *Unbalance* dataset

Table 2: Running times statistics (measured in seconds)

Algorithm	**Unbalance dataset**	**Dim2 dataset**
MTree	0.3 [per seed]	0.06 [per seed]
Canopy	0.01	0.1
EM	8.21 [per tuned seed]	0.55 [per tuned seed]
FF	0.01 [per seed]	0.01 [per seed]
HC	605.42	4.11
SKM	0.06 [per seed]	0.01 [per seed]

Table 3: Performance results on real world datasets

Algorithm	**Accuracy Wine**	**Accuracy Iris**
MTree+Canopy	0.94382	0.89261
MTree+cKMs	0.92134	0.89261
MTree+FF	0.93258	0.88590
MTree+HC	0.89887	0.89261
KMeans	0.93258	0.88590



(a) MTree results (b) Canopy results

(c) EM results (d) FF results

(e) HC results (f) SKM results

Figure 10: Clustering results on *Dim2* dataset

The result shows that objects from one real cluster are split between two real clusters.

– *HC* fails to determine the correct number of clusters and reports one found cluster as a join between two real clusters.

– *SKM* algorithm correctly solves the task after fine tuning the parameters: 10 seeds, usage of kmeans++ [2] for seed optimisation and maximum 10,000 iterations.

Experimental results show that the first K instances have the most significant impact over the final result, where K is the number of clusters. Thus, the current implementation uses kmeans++ [2] seed optimisation, so the first K instances that are added to the MTree are a rather sparse one from another.

Table 2 summarizes the running time statistics for all six algorithms on both data-sets. In the case of EM algorithm, each tuned seed has been obtained by more iterations, and more k-means runs, a fact that explains the more significant running time. HC and Canopy do not have seeds and Canopy's poor results on both data-sets were obtained regardless of the configuration parameters. The number of seeds for the algorithms that correctly solved the *dim2* data-set has been set to 10.

Finally, the SSE (sum of squared error), as well as the assignment of objects, are correctly computed for MTree as they compute to the same values as SKM of 4.2471 and 0.3367 for *unbalance* and *dim2* datasets, respectively. Therefore, the clusters produced by the MTree are valid and represent the real ones from the datasets and SSE represents a good optimisation metric for these datasets.

4.5 Sample runs on Iris and Wine real-world data-sets

Wine data was normalised and then MTree was used on all 13 features. The algorithm was run on 100 seeds and the best result was targeting to be with the best (smallest) SSE. As can be seen from the table, smaller SSE does not always provide the best accuracy, with a SSE value of 66 against 68 but an accuracy of 89 against 94.This suggests that a different cluster quality metrics may be able to improve the performance of the proposed algorithm. KMeans run on 100 seeds obtains 93% accuracy or 12 wrong predictions.

Iris data was normalised and MTree used on all 4 features. As on the previous data-set, the algorithm was run on 100 seeds and best SSE was targeted. It is interesting to notice that different SSE provide the same accuracy, it seems that the algorithm converged with 16 wrong predictions being its best. MTree+cobweb is not able to cluster the data. On the same data, KMeans run on 100 seeds obtains 88% accuracy or 17 wrong predictions.

Table 3 presents accuracy results of MTree parametrised by various splitting algorithms (i.e., Canopy, KMeans, Farthest First and Hierarchical clustering) against baseline KMeans algorithm. Experimental results show the MTree clustering algorithm correctly finds groups at least as good as simple KMeans algorithm.

5 Conclusions and future work

This paper presents an improved parametrised MTree clusters for Weka workbench. The experimental results show that MTree correctly solves two synthetic datasets for which the correct structure (i.e., number of clusters, centroids and distribution) is known. More, five other clustering algorithms implemented in Weka are outperformed in various situations due to that fact that they do not solve the clustering task correctly or need intensive tuning for solving it.

Still, the current approach is used only for *sanity check* of the clustering capabilities of the MTree implementation, rather than solving a particular clustering task on a real dataset. The implementation is Java-based and is available as open source Weka package. The experimental results are correct and promising such that further development under Weka offers the possibility of proper benchmarking of further clustering tasks that may be taken into consideration.

The main contributions are summarised as follows:

1. An updated parametrised version of the MTree package is presented. The parametrisation mainly regards used distance metric, the method for finding and setting the number of clusters and the division policy. The data structure can load various object types after being properly processed as well as providing validation insights.

2. The proposed software architecture of the MTree enables parameterisation through easy integration of other internal algorithmic strategies that perform key tasks within the business logic.

3. The implementation of the MTreeis available as an open source package in Weka workbench. This approach gives the opportunity for further usage and benchmarking against other clustering algorithms.

4. The experiments demonstrate that the proposed approach outperforms current clustering algorithms on two datasets.

Future works may take into consideration extending the *voteK* algorithm as a Java implementation of the already existing R package NbClust. Extending *voteK* should take into consideration the available clustering quality indices and parametrisation capabilities. In terms of internal business logic, MTree may try different approaches regarding the order in which objects are processed when building the tree. One option is to cluster the outlier objects later in the process.

As the most expensive operations are finding the optimal number of clusters and splitting, one option is trying to predict how the insertion of an object will impact the tree in

terms of triggering a split. Checking for the optimal number of clusters should be performed only when an insert is highly to determine a split, as most inserts do not require a split, especially when the dataset has a well-defined structure.

MTree currently implements range query and kNN query. These implementations should be further tested in practical real data scenarios. Other tasks in which MTree may be also used are outlier detection and finding the correct number of clusters in a dataset. Finally, MTree algorithm may be further tested for finding patterns in data in the situation when internal structure is not known.

Acknowledgement

This work was partially supported by the grant 135C/2021 ”Development of software applications that integrate machine learning algorithms”, financed by the University of Craiova.

References

[1] Andreas Adolfsson, Margareta Ackerman, and Naomi C Brownstein. “To cluster, or not to cluster: An analysis of clusterability methods”. In: *Pattern Recognition* 88 (2019), pp. 13–26. DOI: 10.1016/j.patcog.2018.10.026.

[2] David Arthur and Sergei Vassilvitskii. “k-means++: The advantages of careful seeding”. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[3] Catherine L Blake and Christopher J Merz. *UCI repository of machine learning databases, 1998*. 1998.

[4] Roberto Caldelli et al. “Fast image clustering of unknown source images”. In: Jan. 2011, pp. 1–5. DOI: 10.1109/WIFS.2010.5711454.

[5] Jianlong Chang et al. “Deep adaptive image clustering”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5879–5887. DOI: 10.1109/iccv.2017.626.

[6] Malika Charrad et al. “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set”. In: *Journal of Statistical Software* 61 (Oct. 2014), pp. 1–36. DOI: 10.18637/jss.v061.i06.

[7] Paolo Ciaccia et al. “Indexing metric spaces with m-tree.” In: *SEBD*. Vol. 97. 1997, pp. 67–86.

[8] Marius Andrei Ciurez. *MTree client code*. https://github.com/kyko007/Cordoba/tree/master/MTree. 2019.

[9] Marius Andrei Ciurez and Marian Cristian Mihaescu. “Improved Architectural Redesign of MTree Clusterer in the Context of Image Segmentation”. In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2018, pp. 99–106. DOI: 10.29007/sm6x.

[10] Abhisek Dash et al. “Image Clustering without Ground Truth”. In: *CoRR* (Oct. 2016). DOI: 10.48550/arXiv.1610.07758.

[11] Jiangzhou Deng, Junpeng Guo, and Yong Wang. “A Novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering”. In: *Knowledge-Based Systems* (Mar. 2019). DOI: 10.1016/j.knosys.2019.03.009.

[12] F Deves et al. “Scalable Real-Time Shadows using Clustering and Metric Trees”. In: *Eurographics Symposium on Rendering*. Karlsruhe, Germany, July 2018, pp. 1–12. DOI: 10.2312/sre20181175. URL: https://hal.archives-ouvertes.fr/hal-02089095.

[13] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. “Image segmentation using K-means clustering algorithm and subtractive clustering algorithm”. In: *Procedia Computer Science* 54 (2015), pp. 764–771. DOI: 10.1016/j.procs.2015.06.090.

[14] Gianni A Di Caro, Frederick Ducatelle, and L Gambardella. “A fully distributed communication-based approach for spatial clustering in robotic swarms”. In: *Proceedings of the 2nd Autonomous Robots and Multirobot Systems Workshop (ARMS), affiliated with the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)(Valencia, Spain, June 5)*. Citeseer. 2012, pp. 153–171.

[15] Herbert Edelsbrunner. *Algorithms in combinatorial geometry*. Vol. 10. Springer Science & Business Media, 2012. DOI: 10.1007/978-3-642-61568-9.

[16] Frank Eibe, MA Hall, and IH Witten. “The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques”. In: *Morgan Kaufmann* (2016). DOI: 10.1016/B978-0-12-804291-5.00024-6.

[17] Ahmed Ali Abdalla Esmin, Rodrigo A. Coelho, and Stan Matwin. “A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data”. In: *Artif. Intell. Rev.* 44.1 (2015), pp. 23–45. DOI: 10.1007/s10462-013-9400-4.

[18] Adil Fahad et al. “A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis”. In: *IEEE Trans. Emerging Topics Comput.* 2.3 (2014), pp. 267–279. DOI: 10.1109/tetc.2014.2330519.

[19] Pasi Fränti and Sami Sieranoja. "K-means properties on six clustering benchmark datasets". In: *Applied Intelligence* 48.12 (2018), pp. 4743–4759. DOI: 10.1007/s10489-018-1238-7.

[20] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, 2007. DOI: 10.1137/1.9780898718348.

[21] Farid Garcia-Lamont et al. "Automatic computing of number of clusters for color image segmentation employing fuzzy c-means by extracting chromaticity features of colors". In: *Pattern Analysis and Applications* 23 (2020), pp. 59–84. DOI: 10.1007/s10044-018-0729-9.

[22] Melvin Gauci et al. "Clustering objects with robots that do not compute". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2014, pp. 421–428.

[23] Ángel Castellanos Gonzáles, Juan Manuel Cigarrán, and Ana García-Serrano. "Formal concept analysis for topic detection: A clustering quality experimental analysis". In: *Information Systems* 66 (2017), pp. 24–42. ISSN: 0306-4379. DOI: 10.1016/j.is.2017.01.008.

[24] Suchita Goswami and Lalit Kumar P Bhaiya. "Brain tumour detection using unsupervised learning based neural network". In: *2013 International Conference on Communication Systems and Network Technologies*. IEEE. 2013, pp. 573–577. DOI: 10.1109/csnt.2013.123.

[25] Sudipto Guha and Nina Mishra. "Clustering Data Streams". In: *Data Stream Management - Processing High-Speed Data Streams*. Ed. by Minos N. Garofalakis, Johannes Gehrke, and Rajeev Rastogi. Springer, 2016, pp. 169–187. DOI: 10.1049/iet-smt.2018.5389.

[26] Givanna H. Putri et al. "ChronoClust: Density-based clustering and cluster tracking in high-dimensional time-series data". In: *Knowledge-Based Systems* 174 (Feb. 2019). DOI: 10.1016/j.knosys.2019.02.018.

[27] K. Anil Jain and Aditya Vailaya. "Image retrieval using color and shape". In: *Pattern Recognition* 29.8 (1996), pp. 1233–1244. DOI: 10.1016/0031-3203(95)00160-3.

[28] Ismo Kärkkäinen and Pasi Fränti. "Gradual model generator for single-pass clustering". In: *Pattern Recognition* 40.3 (2007), pp. 784–795. DOI: 10.1016/j.patcog.2006.06.023.

[29] Manish Maheshwari, Sanjay Silakari, and Mahesh Motwani. "Image clustering using color and texture". In: *Computational Intelligence, Communication Systems and Networks*. IEEE, 2009, pp. 403–408. DOI: 10.1109/CICSYN.2009.69.

[30] Marian Cristian Mihaescu. *MTree Clusterer*. Accessed: 2019-05-30. URL: http://weka.sourceforge.net/packageMetaData/MTreeClusterer/index.html.

[31] Marian Cristian Mihaescu and Dumitru Dan Burdescu. "Using M tree data structure as unsupervised classification method". In: *Informatica* 36.2 (2012).

[32] Anton Milan et al. "Joint tracking and segmentation of multiple targets". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 2015, pp. 5397–5406. DOI: 10.1109/cvpr.2015.7299178.

[33] Jose A Miñarro-Giménez, Markus Kreuzthaler, and Stefan Schulz. "Knowledge Extraction from MEDLINE by Combining Clustering with Natural Language Processing". In: *AMIA Annual Symposium Proceedings*. Vol. 2015. American Medical Informatics Association. 2015, p. 915.

[34] Traian Rebedea and Ştefan Trăuşan-Matu. "Autonomous News Clustering and Classification for an Intelligent Web Portal". In: *Foundations of Intelligent Systems*. Springer Berlin Heidelberg, 2008, pp. 477–486. DOI: 10.1007/978-3-540-68123-6_52.

[35] Mohammad Rezaei and Pasi Fränti. "Set matching measures for external cluster validity". In: *IEEE Transactions on Knowledge and Data Engineering* 28.8 (2016), pp. 2173–2186. DOI: 10.1109/TKDE.2016.2551240.

[36] Hermes Robles et al. "LEAC: An efficient library for clustering with evolutionary algorithms". In: *Knowledge-Based Systems* (May 2019). DOI: 10.1016/j.knosys.2019.05.008.

[37] Érick Oliveira Rodrigues et al. "K-MS: a novel clustering algorithm based on morphological reconstruction". In: *Pattern Recognition* 66 (2017), pp. 392–403. DOI: 10.1016/j.patcog.2016.12.027.

[38] Tiago Rodrigues Lopes dos Santos and Luis E. Zárate. "Categorical data clustering: What similarity measure to recommend?" In: *Expert Syst. Appl.* 42.3 (2015), pp. 1247–1260. DOI: 10.1016/j.eswa.2014.09.012.

[39] Lincoln F Silva et al. "Hybrid analysis for indicating patients with breast cancer using temperature time series". In: *Computer methods and programs in biomedicine* 130 (2016), pp. 142–153. DOI: 10.1016/j.cmpb.2016.03.002.

[40] Jeffrey K Uhlmann. “Satisfying general proximity/similarity queries with metric trees”. In: *Information processing letters* 40.4 (1991), pp. 175–179. DOI: 10.1016/0020-0190(91)90074-R.

[41] Ulrike Von Luxburg, Robert C Williamson, and Isabelle Guyon. “Clustering: Science or art?” In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. 2012, pp. 65–79.

[42] Zhimin Wang et al. “Adaptive spatial information-theoretic clustering for image segmentation”. In: *Pattern Recognition* (Sept. 2009), pp. 2029–2044. DOI: 10.1016/j.patcog.2009.01.023.

[43] Pavel Zezula et al. *Similarity search: the metric space approach*. Vol. 32. Springer Science & Business Media, 2006. DOI: 10.1007/0-387-29151-2.

[44] Shibing Zhou and Zhenyuan Xu. “Automatic grayscale image segmentation based on Affinity Propagation clustering”. In: *Pattern Analysis and Applications* (Feb. 2019). DOI: 10.1007/s10044-019-00785-4.

OMCOKE: A Machine Learning Outlier-based Overlapping Clustering Technique for Multi-Label Data Analysis

Said Baadel[1], Fadi Thabtah[2], Joan Lu[3], Saida Harguem[4]
[1]Mt. Royal University, Calgary, Canada, [2]ASDTests, Auckland, New Zealand, [3]University of Huddersfield, Huddersfield, UK, [4]Canadian University Dubai, Dubai, UAE
Email: sbaadel@mtroyal.ca, fadi@asdtests.com, j.lu@hud.ac.uk, saida.harguem@cud.ac.ae

Keywords: K-means; Machine Learning; OMCOKE; Overlapping Clustering; Unsupervised Learning

Received: March 3, 2021

Clustering is one of the challenging machine learning techniques due to its unsupervised learning nature. While many clustering algorithms constrain objects to single clusters, K-means overlapping partitioning clustering methods assign objects to multiple clusters by relaxing the constraints and allowing objects to belong to more than one cluster to better fit hidden structures in the data. However, when datasets contain outliers, they can significantly influence the mean distance of the data objects to their respective clusters, which is a drawback. Therefore, most researchers address this problem by simply removing the outliers. This can be problematic especially in applications such as fraud detection or cybersecurity attacks risk analysis. In this study, an alternative solution to this problem is proposed that captures outliers and stores them on-the-fly within a new cluster, instead of discarding. The new algorithm is named Outlier-based Multi-Cluster Overlapping K-Means Extension (OMCOKE). Empirical results on real-life multi-label datasets were derived to compare OMCOKE's performance with other common overlapping clustering techniques. The results show that OMCOKE produced a better precision rate compared to the considered clustering algorithms. This method can benefit various stakeholders as these outliers could have real-life applications in cybersecurity, fraud detection, and the anti-phishing of websites.

Povzetek: V tej študiji je predlagana alternativna rešitev (OMCOKE), ki zajame izstope in jih sproti shrani v novo gručo, namesto da bi jih odstranila.

1 Introduction

Clustering is an unsupervised learning process that involves grouping a set of data objects into subsets, each of which has its own label based on a predefined similarity metric [2] [5]. Prior to learning, each resulting subset will contain data objects usually exhibiting similar traits but dissimilar from data objects in the other subsets [25]. In clustering, some structural characteristics are not known a priori unless some sort of domain knowledge is presented in advance (i.e. there are no labels attached to the data patterns as in the case of supervised classification), thus deeming clustering a difficult problem due to this unsupervised nature [23] [32]. Various clustering techniques such as probabilistic, distance-based, and grid-based have been explored in machine learning with the distance-based proving to be popular [1] [24].

Undoubtedly, the K-means [30], and its generic extensions and adaptations, is one of the most widely used distance-based partition-clustering algorithms [23] [26] [28]. There are many reasons attributed to this, such as it is easy to implement, its versatility allows any part to be easily modified, and its guaranteed nature to converge at a quadratic rate [16]. Thus, the K-means algorithm has been primarily utilized to deal with non-overlapping clustering problems that limit each data object to a single cluster. However, one of the main challenges of K-means and its successors is sensitivity to exceptional data (outliers). K-means often derives clusters by optimizing the mean Sum of Squared Error (SSE) (Equation 1) by calculating the Euclidean distance between the data objects and the clusters 'computed centroids.

$$SSE = \sum_{k=1}^{K} \sum_{xi \in Ck} ||xi - ck||^2 \quad (1)$$

Where Ck is the kth cluster, xi is a point in Ck, and ck is the mean of the kth cluster.

In cases when the input dataset contains few outliers, this may significantly influence the mean distance (the outlier will skew the mean and variance) of the data objects to their respective clusters, and thus K-means tends to discard outliers [6] [15] [34]. Existing algorithms that extend the K-means and allow objects to overlap include Kernel Overlapping K-means (KOKM) [9][11], Overlapping K-means (OKM) [17-18], Parametrized R-OKM [9] and Multi-Cluster Overlapping K-Means Extension (MCOKE) [3][4]. Detecting these outliers is advantageous for decision makers as these outliers could be used for fraudulent activities such as in the case of cybersecurity or a fraud insurance claim. Therefore, it will

be more useful to store these outliers (as opposed to discarding them) in a separate cluster for potential usage as they represent exceptional patterns.

This research addresses the above issue by detecting outliers during the clustering process and then storing them, making it different from the other overlapping clustering algorithms. Since the user identifies the number of k clusters a priori when running clustering algorithms, our algorithm is able to adjust to this and accommodate the outliers by adding a new cluster during the learning process called an Outlier cluster (k+1). The proposed algorithm is called Outlier-based Multi-Cluster Overlapping K-Means Extension (OMCOKE); uses the outlier cluster for later analysis by decision makers.

The current study approach adds immense value to the learning process as we save these data objects to investigate and understand their characteristics. These data objects could potentially be a result of an imbalanced data set with high cardinality (i.e. natural overlaps) and perhaps the k number of clusters, which is defined a priori, can be revised to accommodate the data and allow the algorithm to better fit the clusters.

Outliers could also indicate suspicious data objects with malicious intent. Therefore, an outlier cluster that can be investigated has profound real-life implications such as in e-banking, website phishing, cyber security, or medical screening. For example, in cyber security, historical data can reveal statistical acceptable trends through the data patterns and how they are clustered together. Any outlier objects outside the regular clustered trends will automatically raise red flags. Such red flags can be used in data analytics to alert the user of a potential security threat or an intrusion attempt.

Experimental results using real datasets indicate that OMCOKE is able to detect outliers and to produce clusters with higher precision and accuracy when compared to existing algorithms such as OKM, KOKM, and R-OKM among others.

The rest of the paper is structured as follows: Section 2 reviews the literature concerning overlapping clustering. Section 3 discusses OMCOKE and datasets used in the empirical experiments. Section 4 provides the results and analysis with a comparison of different ML clustering techniques. Lastly, we provide conclusions and further research in Section 5.

2 Literature review

Outliers are data objects or points that do not conform to the normal behaviour or model of the dataset, hence are deemed inconsistent or grossly different [12]. This data can be erroneous, but could also be classified as suspicious data in fraudulent activity; that could be useful for fraud detection, intrusion detection marketing, website phishing sites, etc.

Outlier detection is considered a task in itself; research in the data mining domain has focused on an efficient and optimal way to detect distance-based outliers. Outlier detection surveys such as by Chandola, et al. [15], Bay and Schwabacher [7], and Kadam and Pund [27] discussed several approaches used to tackle anomalies and noise data. In [8] and [31] the authors provide methods that would efficiently mine outliers in large datasets. Other recent studies have devised methods in clustering analysis that will prune or screen out outliers from the dataset such as Liu, et al. [29], Barai and Dey[6], Gan and Nk [21] , Danganan et al. [19] and Chagas et al.[14] . For example, Yu, et al. [35] proposed an outlier detection method to identify and eliminate outliers in the dataset forming an outlier-eliminated dataset (OED). The authors then applied the K-means algorithm on the OED, thereby improving the accuracy of the clustering.

Similarly, the Barai & Dey [6] approach is to divide their algorithm into two steps. The first step calculates the threshold value used in detecting outliers by taking the average of the maximum and minimum values of pairwise distance of all data. Each data point is then reiterated and compared to the threshold. Those that have a distance value greater than the threshold are deemed as outliers and are subsequently tossed out of the dataset. The second step then runs the K-means algorithm without outliers, thus improving the clustering process.

Liu et al., [29] also propose a two-phased approach for their clustering with the outlier removal (COR) algorithm. In the first phase, their method runs the K-means algorithm to generate basic partitions and discover outliers. The outliers here are identified as objects with large distances to their nearest centroid. The second phase removes the identified outlier objects and the remainder are partitioned into k clusters.

Similarly, Danganan et al [19] proposed a modification of MCOKE [3] by incorporating a median absolute deviation (MAD) that measures any potential outliers in the dataset. The authors proposed a three-phased approach in which the objects are ranked in ascending order and the distance of each object is calculated against MAD which is multiplied to a certain constant number determined by the user to obtain a decision value. If the distance of an object is greater than the decision value, that object is deemed an outlier and is pruned from the dataset.

While many studies focus on pruning and discarding the outliers to improve the classification process, rarely do we find algorithms that detect outliers simultaneously while performing clustering [19]. The K-means with outlier removal (KMOR) algorithm is similar to the standard K-means algorithm but introduces an outlier cluster (k+1) that takes into account objects that don't fit in the k defined clusters. The algorithm identifies outliers as objects that are above a calculated threshold which is defined by the average distance multiplied by a certain parameter greater or equal to 0. The average distance is calculated during the clustering phase. The KMOR algorithm requires three parameters such as the k number of clusters, the maximum number of outliers n0 (to control the number of objects being assigned as outliers), and finally, a third parameter to classify outliers and those that

are not. Two additional parameters are used to help terminate the algorithm.

All the studies mentioned above utilize the K-means partition algorithm that eventually constrains objects to single clusters. Overlapping partitioning clustering methods tends to relax or remove the constraints allowing overlaps between clusters; this better fits any hidden structures in the data and assign data objects to one or more clusters building a non-disjoint partition of the data [4] [5].

The present study focuses on overlapping partitioning methods which have several applications in real-life such as dynamic system identification, document categorization (a document belonging to different clusters), data compression, bioinformatics, image recognition, model construction, etc. [1] [21].

Extensions of the K-means that allow overlaps include Kernel Overlapping K-means (KOKM) [9,11], Overlapping K-means (OKM) [17,18], Parametrized R-OKM [10] and Multi-Cluster Overlapping K-Means Extension (MCOKE) [3][4].

The OKM algorithm is an extension of K-means that allows overlaps by using a heuristic that discovers a combinatorial set of possible assignments of the data points. For each observation, the heuristic sorts the clusters from the closest to the farthest; it then assigns the objects to those centroids in the defined order while minimizing the distance between the centroid and the observed object.

The KOKM algorithm is a variant of OKM that utilizes the use of kernel methods for overlapping clustering. The authors use two variants in their method; one is a kernelization of the Euclidean metric, similar to the one used in OKM, that calculates the distances between the objects and the clusters in a high dimensional mapping space; the second variant performs all the clustering steps where data is implicitly mapped.

The Parameterized R-OKM algorithm is another variant of OKM that lets users regulate the overlaps via a parameter. As the size of the parameter increases, the algorithm builds clusters with reduced overlaps, and vice-versa when the size of the parameter approaches zero. The PR-OKM algorithm is reduced to OKM when this parameter is set to exactly zero.

Unlike other algorithms that prune the outliers and discard them, the proposed algorithm saves them on a newly created outlier cluster during the iteration process. The present study considers the same idea as the KMOR algorithm and introduces an outlier cluster k+1 that stores the anomalies or outlier objects separately from the normal instances. As noted above, the KMOR algorithm requires users to define the maximum number of outliers, including a parameter to classify the outliers and those that are not. This is impractical in real-life scenarios in unsupervised datasets where no prior knowledge of the data is given. Also, their method requires additional parameters to help terminate the algorithm. This is not an easy feat to be determined by novice users. However, in this study we do not require users to enter parameters to terminate the algorithm or to identify the maximum number of outliers in the dataset; this makes it more practical in machine learning. None of the overlapping K-means algorithms above have the capability to detect outliers and store them for additional scrutiny. Thus, we provide additional value to the literature by introducing this new overlapping clustering method.

This study considers the key classification evaluation measures of Precision and F-measure. We evaluate and compare the results to highlight the significance of excluding the outliers in the dataset when clustering and how that improves the precision of the algorithm.

The following section discusses the proposed clustering algorithm and the dataset used for evaluation.

3 The proposed OMCOKE algorithm and experimental dataset

The proposed method is an enhancement of the MCOKE algorithm [3] that allows objects to overlap and belong to more than one cluster based on their distance comparison to the maxdis variable. Maxdist calculates the largest distance of any object assigned to any centroid during the partitioning phase for it to belong to a particular cluster. That distance is used as an outer radius of similarity threshold and as the benchmark to allow objects to belong to other clusters that were not initially assigned to them, allowing them to overlap. However, K-means, being a greedy algorithm, guarantees all objects to be assigned to a cluster including any outliers, hence the maxdist radius benchmark could easily be influenced by outliers.

The present study introduces another variable that calculates the average distance (averdist) between the object and the centroid for all clusters. Averdist acts as a new threshold for the inner radius between the object and the centroid.

$$averdist = \frac{1}{n_i}\sum_{x_i \in C_k} ||x_i - C_k||^2 \quad i = 1, 2, \dots K \qquad (2)$$

Where Ck is the kth cluster, xi is a point in Ck.

It is assumed that most objects being clustered will fall close to the inner radius threshold (i.e. close to their cluster centroid) that is based on the average distance of all objects belonging to the cluster centroids. Anomalies or outliers therefore tend to be further away from their closest cluster centroid. Objects that have a distance greater than the inner radius but less or equal to the outer radius (maxdist) are subject to further scrutiny and are flagged to ensure they are not outliers on the border of the clusters. Therefore, the maxdistThreshold defines the radius distance to be considered from the outer boundary, for example, 0.98 will mean the area covered inside the

outer boundary for objects is not to be considered an anomaly. This logic is based on the assumptions that:

a) Anomalies tend to be in sparse clusters, whereas normal instances usually belong to dense clusters

b) Anomalies tend to be far from the closest cluster centroid, whereas normal instances tend to be near their closest cluster centroid.

In cases where some knowledge of the data is known beforehand, this value can also be adjusted by the user prior to running the algorithm.

This modification logic is summarized in the pseudocode provided below.

Outlier Detection Pseudocode

```
1. For each x_i ϵ C_k
2. Do
3. If (dist (x_i, centroid C_k) ≤ averdist)
4.         Cluster ← x_i
5. Else
6.     If (dist (x_i, centroid C_k) ≥ maxdist * maxdistThreshold)
7.         Outlier_Cluster ← x_i
8.     Else
9.         Cluster ← x_i
10.    End if
11. End if
```

In Step 6 of the code above, the area covered by the maxdistThreshold is multiplied by the maxdist, calculated as a percentage of the overall maximum distance for any object belonging. This acts as the cut-off point and any object that has a distance value greater than the upper percentile of this value is deemed an outlier. Upon identification of at least one outlier, the k number of clusters entered by the user prior to running the method is incremented by 1 on the fly; the outlier object is assigned to this newly created cluster. All other identified outliers, a subset S from the initial population, are assigned to belong to this newly created cluster. Once an outlier is detected, the algorithm adds *k+1* clusters as the new output vector with the outlier cluster indexes listed as part of the output. This allows for further investigation of those data points as opposed to discarding them as is usual. When no outliers are detected, the algorithm will simply cluster with overlaps without incrementing the number of *k* clusters.

3.1 Experimental dataset

Different datasets from the Mulan: A Java Library for Multi-Label Learning repository [34] are used to evaluate the proposed algorithm's performance. The data repository hosts more than 25 different datasets in the domains of text, audio, video, music, images, and biology to mention only a few. Items of multi-label datasets can be members of multi-groups which are true for real world problems and, as a result, ideal for the study of overlapping clustering. In our empirical experiment, three different domain datasets that have been used, along with their specifications and descriptive statistics, are displayed in Table 3 and Table 4 respectively.

Data Set	Instances	# of Labels	Attribute	Cardinality
Emotions	593	6	72	1.869
Yeast	2417	14	103	4.237
Scene	2407	6	294	1.074

Table 3: Statistics of used Benchmarks

Data Set	Min	Max	Mean	StdDev
Emotions	0.01	0.195	0.069	0.031
Yeast	0.371	0.52	0.001	0.097
Scene	0.0	1.0	0.659	0.214

Table 4: Descriptive Statistics of used Benchmarks

3.2 Description of the Overlapping Datasets

This study conducted experiments on real-life overlapping datasets to measure the effectiveness of the methods used to identify such overlapping groups. The three datasets have a wide diversity in their dataset making them a suitable combination for use as benchmarks. For example, their sizes vary from 593 (Emotions) to 2417 (Yeast), their dimension (attributes) from 72 (Emotions) to 294 (Scene), cardinality (i.e. overlap rates) from 1.074 (Scene) to 4.237 (Yeast). Their application domain also varies considerably i.e. music, biology, and images.

The following is a brief description of the three datasets (Emotion, Yeast, and Scene).

3.2.1 Emotion dataset

Analyzing music signals is used in the detection of emotion in music. In this case, music can be classified into several categories at the same time since they are not usually disjointed i.e. it can make you feel both "sad" and "angry". The dataset contains sound clips that can be described by 72 attributes which were annotated by three male music experts into six emotional clusters. Only the songs that had all three experts unanimously agree on their label were kept, resulting in a total of 593 songs being selected for the dataset.

3.2.2 Yeast dataset

The Yeast dataset is classified into 14 gene groups or classes. A gene can belong to several different classes at the same time thus making this a multilabel dataset. For example, the gene YAL014W may belong in the following four groups: {Cell Growth, Cell Division}, {Cellular Organization}, {Cellular Communication, Signal

Transduction} and {Transposable elements, Viral and Plasmid Proteins}.

3.2.3 Scene dataset

The dataset contains 2407 natural scene images. The images were classed into six categories. In this case, the images can be classified into different categories at the same time since they are not usually disjointed i.e. they become multilabelled and can belong to more than one category such as field + mountain or fall foliage + mountain.

4 Experimental results

4.1 Experimental settings

To exhibit the performance of our algorithm, with respect to different measures when contrasted with a wide range of ML, the current study selected clustering algorithms using the following criteria:

a) Algorithms that utilize the partitioning method that extends the K-means algorithm
b) The algorithms use the Euclidian distance to calculate the similarities between the sets of observations
c) All algorithms work on numeric attributes only
d) All are known algorithms that have been evaluated by previous researchers in ML.

All experiments have been run on an Intel Core i7 computer with a 3.4 GHz processor and 8.0 GB RAM running on a 64-bit, Windows 10 Operating System.

We used the pair-based Precision-Recall measure that is calculated over pairs of observations. The precision-recall is computed as follows:

Where TP is a true positive decision, FP is a false positive decision (two dissimilar objects assigned to the same cluster), and FN is a false negative (two similar objects assigned to different clusters).

$$\Pr ecision = \frac{|TP|}{|TP + FP|} \tag{3}$$

$$\Re call = \frac{|TP|}{|TP + FN|} \tag{4}$$

$$F - measure = \frac{|2 * \Pr ecision * \Re call|}{|\Pr ecision + \Re call|} \tag{5}$$

Where TP is a true positive decision, FP is a false positive decision (two dissimilar objects assigned to the same cluster), and FN is a false negative (two similar objects assigned to different clusters).

4.2 Empirical results and analysis

For fair comparisons, datasets with different sizes and from different domains have been chosen and are compared to well-known algorithms that have been evaluated by previous researchers. Through experimental study, we evaluated and compared the performance of OMCOKE with three existing methods namely: Kernel Overlapping K-means (KOKM), Overlapping K-means (OKM), and Parametrized R-OKM as shown in Table 5 below.

For each experiment, we set the parameters for KOKM, OKM, and P-ROKM as follows:

- Maximum iterations = 10
- Number of clusters = 3
- Number of labels = Emotions (6), Yeast (14), and Scene (6).
- Minimal improvement = 0.01
- Alpha = 1 and 0.1 for P-ROKM algorithms.

In addition to the number of iterations and clusters set as above, the following parameters were also set in OMCOKE:

- maxdistThreshold = 0.99
- useMeasures = True

Overlapping methods will have an overlap that is greater than 1 since the objects belong to more than one cluster. The size of the overlaps affects the value of Precision i.e., there will be low value of Precision because the observations are assigned to more than one cluster.

Method	Emotion		Yeast		Scene	
	P.	***F.***	***P.***	***F.***	***P.***	***F.***
KOKMII	0.471	**0.641**	0.785	**0.878**	0.193	0.324
OKM	0.467	0.586	0.234	0.376	0.234	0.376
P-ROKM (α=1)	0.474	0.524	0.919	0.565	0.379	**0.506**
P-ROKM (α=0.1)	0.468	0.578	0.802	0.654	0.288	0.439
OMCOKE	**0.565**	0.419	**0.972**	0.496	**0.706**	0.453

Table 5: Comparison of Performace

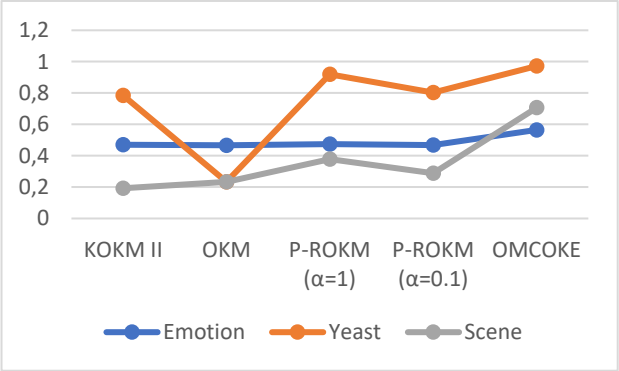


Fig. 3: Precision Accuracy of the Benchmark Datasets

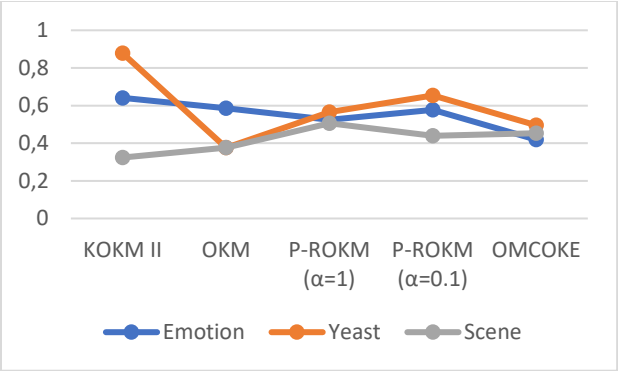


Fig. 4: F-Measure of the Benchmark Datasets

The pair-based Precision-Recall method used in the empirical results is calculated over pairs of observations. This allows for the evaluations of clusters independently and compares their partitions with different numbers of clusters in the dataset. It measures whether the predicted pair is correctly assigned in the same cluster as indicated in the true class datasets. However, the Recall measure uses a binary function to compute the relationship between pairs of observations, and not considering that those pairs of observations could also feature in multiple clusters in the overlap. This results in a biased Recall measure, especially when the cardinality in the dataset is large. Thus, we chose not to use the Recall in our experiment as a measure of OMCOKE.

It is evident from the above empirical results that the OMCOKE algorithm has a high precision rate and outperforms all the other overlapping algorithms in the study as shown in Figure 3 above. This can be attributed to the algorithm's ability to separate outliers from the rest of the data objects when assigning them to clusters. For the Emotion, Yeast, and Scene datasets, OMCOKE precision was 0.565, 0.972, and 0.706 followed by P-ROKM (α=1) at 0.474, 0.919, and 0.379 respectively.

High values of F-Measures are generally induced by the high values of Recalls as opposed to non-overlapping algorithms whose high values of F-measures are generally as a result of the Precision. When compared to the other algorithms, OMCOKE performs relatively well in the F-Measure as shown in Figure 4 below, scoring second behind P-ROKM (with α=1) in the Scene dataset; the P-ROKM method with the alpha value of 1 yielded an overlap of exactly 1 and dataset had a cardinality of 1.07.

The F-Measure values are higher for clustering methods whose overlap rates are closer to the actual cardinality of the dataset. The cardinality shown in Table 3 is the natural overlaps in the dataset i.e., the average number of categories each observation can belong to. The analysis shows that the F-Measures and Precision are significantly affected by the overlap rate in the actual dataset. Algorithms that have partitions with smaller overlaps fared well in their F-Measure meaning that they produced non-disjointed partitions that fit the data better compared to others. OMCOKE performed reasonably well in the Scene and Emotion datasets since the cardinalities of the datasets are low (1.074 and 1.869 respectively) nearing 1 but did poorly in the Yeast dataset that had an overlap of over 4. Our algorithm detected several outliers in the dataset. These are listed in Table 6 below.

Dataset	Number of Outliers	Identified Outlier Instances
Emotion	1	27
Scene	2	304; 1502
Yeast	1	1819

Table 6: Outliers Detected in the Three Datasets

As indicated, an input dataset containing a few outliers significantly influences the mean distance (the outlier will skew the mean and variance) of the data objects to their respective clusters. This explains why OMCOKE outperformed the other methods in all datasets in terms of Precision rate. This also shows that by separating the outliers from the rest of the data, the OMCOKE was able to build its model relatively closer and more acceptable to the actual overlaps in each of the datasets; this is as compared to the other methods for the precision to be higher than the rest.

5 Conclusions and future work

In this paper, some different K-means variants of overlapping clustering methods were discussed.

The proposed algorithm, with the capability of detecting outliers and treating them as a separate cluster, was evaluated and compared with three existing overlapping clustering methods namely: Kernel Overlapping K-means (KOKM), Overlapping K-means (OKM), and Parametrized R-OKM. We used real-life multi-label datasets for our experiments. The empirical results showed that the F-Measures and Precision were significantly affected by the overlap rate in the actual dataset. OMCOKE did well in the Scene dataset since the cardinality of the dataset is very low and did poorly in the

Yeast dataset that had a significant high overlap rate of over 4. However, when it came to Precision, OMCOKE outperformed the other overlapping algorithms in all datasets indicating that our method had a better detection rate of clusters and for assigning observations with a better precision after it segregated the outliers in the dataset.

The proposed algorithm detects and stores outliers during the clustering process making it different from the other overlapping clustering algorithms, thus adding value in this domain. As opposed to discarding anomalies and outliers, our method can provide tremendous benefit to cyber security experts, medical practitioners, IT administrators, data mining researchers, and other stakeholders as these outliers could have real-life applications such as fraudulent activities as in the case of cybersecurity, fraud insurance claims in the banking domain, or to help raise flags in the medical field especially in the screening process.

In future, we plan to extend the method to increment k cluster to more than 1 to cater for other dispersed objects that may not necessarily be deemed anomalies but could form dispersed clusters that have common characteristics that are somehow dissimilar from the rest of the data objects. These newly created clusters can then be fused and merged based on their similarity weights to minimize the number of clusters produced in large datasets.

References

[1] Aggarwal, C., & Reddy, C. K. (2014). *Data clustering: Algorithms and applications*. CRC Press.

[2] Arabie, L. J., Hubert, G., & DeSoete, P. (1999*). Clustering and classification.* World Scientific.

[3] Baadel, S., Thabtah, F., & Lu, J. (2015). MCOKE: Multi-Cluster Overlapping K-Means Extension Algorithm. International Journal of Computer, Control, Quantum and Information Engineering 9(2). Pp. 374-377.

[4] Baadel, S., Thabtah, F., & Lu, J. (2016). *Overlapping clustering: A review*. IEEE SAI Computing Conference, London, UK. Pp 233-237. https://doi.org/10.1109/sai.2016.7555988

[5] Baadel, S. (2021). Big Data Analytics: A Tutorial of Some Clustering Techniques. International Journal of Management and Data Analytics, 1(2). Pp 38-46.

[6] Barai, A., & Dey, L. (2017). Outlier detection and removal algorithm in K-means and hierarchical clustering. *World Journal of Computer Application and Technology, 5*(2). 24-29. https://doi.org/10.13189/wjcat.2017.050202

[7] Bay, S., & Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In KDD. https://doi.org/10.1145/956750.956758

[8] Beltran, B., Vilarino, D., Martinez-Trinidad, J., Carrasco-Ochoa, J.A. (2020). K-means based method for overlapping document clustering. Journal of Intelligent and Fuzzy Systems, 39 (2). Pp. 2127-2135. https://doi.org/10.3233/jifs-179878

[9] BenN'Cir, C., & Essoussi, N. (2012). Overlapping patterns recognition with linear and non-linear separations using positive definite kernels. International Journal of Computer Applications (IJCA), pp 1–8. https://doi.org/10.5120/8916-2981

[10] BenN'Cir, C., Cleuziou, G., & Essoussi, N. (2013). Identification of non-disjoint clusters with small and parameterizable overlaps. In IEEE International Conference on Computer Applications Technology (ICCAT), pages 1–6. https://doi.org/10.1109/iccat.2013.6522010

[11] BenN'Cir, C., Essoussi, N., & Bertrand, P. (2010). Kernel overlapping k-means for clustering in feature space. In International Conference on Knowledge discovery and Information Retrieval (KDIR), pp 250–256. https://doi.org/10.5220/0003095102500256

[12] Berkhin P. (2006) A survey of clustering data mining techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-28349-8_2

[13] Boundaillier, E., & Hebrail, G. (1988). Interactive interpretation of hierarchical clustering. Intelligent Data Analysis.

[14] Chagas, G. O., Lorena, A., Dos Santos, R. (2019). A hybrid Heuristic for the overlapping Clustering problem. Applied Soft Computing. 81(105482), 1-48. https://doi.org/10.1016/j.asoc.2019.105482

[15] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-72. https://doi.org/10.1145/1541880.1541882

[16] Celebi, M., Kingravi, H., & Vela, P. (2013). A comparative study of efficient initialization methods for the K-means clustering algorithm. Expert Systems with Applications. 40 (1). 200-210. https://doi.org/10.1016/j.eswa.2012.07.021

[17] Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In International Conference on Pattern Recognition ICPR, pp 1–4. https://doi.org/10.1109/icpr.2008.4761079

[18] Cleuziou, G. (2009). Two variants of the okm for overlapping clustering. Advances in Knowledge Discovery and Management. pp 149–166. https://doi.org/10.1007/978-3-642-00580-0_9

[19] Danganan, A., Sison, A., Medina, R. (2019). OCA: Overlapping Clustering application unsupervised approach for data analysis. Indonesian Journal of Electrical Engineering and Computer Science, 14 (3) pp. 1473-1478. https://doi.org/10.11591/ijeecs.v14.i3.pp1471-1478

[20] Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. In T.G. Dietterich, S. Becker, and Z. Ghahramani, (eds), Advances in Neural Information Processing Systems.

[21] Gan, G., & Ng, M. K. (2017). K-means clustering with outlier removal. Pattern Recognition Letters,

90,8-14. https://doi.org/10.1016/j.patrec.2017.03.008

[22] Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). Fuzzy cluster analysis: Methods for classification, data analysis and image recognition. Wiley.

[23] Hrushka, E. R., Campello, R., Freitas, A., & Carvalho, A. (2009). A survey of evolutionary algorithms for clustering. IEEE Transactions on Systems, Man, and cybernetics, Part C. (Applications and Reviews), 39(2), 133-155. https://doi.org/10.1109/tsmcc.2008.2007252

[24] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31(8) 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[25] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data, Prentice Hall.

[26] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys 31(3) 264–323. https://doi.org/10.1145/331499.331504

[27] Kadam, N. V., & Pund, M. A. (2013). Joint approach for outlier detection. International Journal of Computer Science Application, 6 (2), 445–448.

[28] Lam, D., & Wunsch, D. (2014). Clustering. Academic Press Library in Signal Processing, Signal Processing Theory and Machine Learning, (1). https://doi.org/10.1016/b978-0-12-396502-8.00020-6

[29] Liu, H., Li, J., Wu, Y., & Fu, Y. (2018). Clustering with outlier ermoval. Proceedings of ACM Sig on Knowledge Discovery and Data Mining (KDD). ACM, New York, NY, USA.

[30] McQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations, In: Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297.

[31] Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In SIGMOD. https://doi.org/10.1145/335191.335437

[32] Saxena, A., Prasad, M., … Gupta, A. (2017). A review of clustering techniques and developments. International Journal of Neurocomputing. 267. Pp 664-681. https://doi.org/10.1016/j.neucom.2017.06.053

[33] Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multilabel classification of music into emotions. Proceeding of the 2008 International Conference on Music Information Retrieval (ISMIR 2008), pp. 325-330, Philadelphia, PA, USA. https://doi.org/10.1186/1687-4722-2011-426793

[34] Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data, data mining and knowledge discovery handbook, O. Maimon, L. Rokach (Ed.), Springer, 2nd ed., 2010. https://doi.org/10.1007/978-0-387-09823-4_34

[35] Yu, Q., Luo, Y., Chen, & C., Ding, X. (2016). Outlier-eliminated k-means clustering algorithm based on differential privacy preservation. Applied Intelligence, 45 (4). 1179–1191. https://doi.org/10.1007/s10489-016-0813-z

[36] Zhang, J. S., & Leung, Y. (2003). Robust clustering by pruning outliers. IEEE Trans. on Systems, Man, and Cybernetics – Part B 33 (6) 983–999. https://doi.org/10.1109/tsmcb.2003.816993

Focus Web Crawler on Drug Herbs Interaction Patterns

Fatini Nadhirah Mohd Nain[1], Nurul Hashimah Ahamed Hassain Malim*[1], J. Joshua Thomas[2] and Mei Lan Tan[3]
Email: fatininadhirahnain@student.usm.my, nurulhashimah@usm.my, jjoshua@kdupg.edu.my, tanml@usm.my
* Corresponding author
[1]School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Pulau Pinang, Malaysia.
[2]Department of Computing, UOW Malaysia KDU Penang University College, Georgetown 10400, Pulau Pinang, Malaysia.
[3]School of Pharmaceutical Sciences, Universiti Sains Malaysia, Gelugor 11800, Pulau Pinang, Malaysia.

Keywords: Drug-herb interactions, Focus Web crawler, Breadth-First Search (BFS), PageRank.

Received: March 14, 2021

The types of pharmaceutical products include cosmetics and drugs. Some of the pharmaceutical products comprise a mix of drugs and herbs without considering their interaction effects. Drug-herb interactions (DHIs) refer to the interactions between conventional drugs and herb medicines. However, the available information on DHIs is scattered because it has heterogeneous databases and website resources, apart from some of the paid or subscribed databases. Easy access to information on DHIs would allow researchers to explore more. Therefore, this study proposes improvements in the focus web crawler to collect DHIs information from the heterogeneous resources on the Internet, present priority levels of a resource link through anchor text and URLs, and traversing the link with the aid of depth. The improved focused crawler was tested on two algorithms namely the Breadth-First Search (BFS) and PageRank. Information of DHIs crawled 4,744 herbals from the focus web crawler. The accuracy values for Chinese Med Digital Projects and MedlinePlus were 98% for PageRank and 71% for BFS. Additionally, a focused web crawler may gather more relevant web pages in the same amount of time as a wide crawler. Hence, the proposed crawler may successfully gather DHIs on the web in response to the user queries.

Povzetek: Razvit je nov algoritem za preiskovanje spleta za iskanje vzorcev medsebojne odvisnosti zdravil.

1 Introduction

Despite the advancements in modern medicine, most people still use herbals to cure their illnesses. In the 17th century, many countries practiced herbal medicine based on their traditional knowledge of a plant that was used by the local communities and was passed down from one generation to another. Now, many of the products have mixed conventional drugs with herbals. Therefore, it will lead to a large gap in increasing the number of chemicals consisting of primary and secondary metabolites of the active substance using single pharmacology that contributes to the effects of either moderate, resisting, etc.[1].

Contrary to the popular believe, the side effects of herbal medicines are greater compared to conventional drugs, regardless of the generalization of 'natural means safe' due to the lack of appropriate quality control, inadequate labeling, and lack of appropriate patient information. [2,3]. Of lately, various plant-derived products are being incorporated into cosmetics and natural products. These products contain active phytochemicals in a range of unstandardized preparations (i.e. tablets, capsules, sachets, or pills). Some of the sports drinks, supplements and energy bars contain ingredients that have been mixed with herbs and medicines. The effect of a mixture of homemade medicines used where the patient begins to manage it on their own without supervision and advice from a doctor, therefore, offer an increase to the rate of drug-herbal interactions (DHIs) [4]. DHIs refer to the interactions between conventional drugs and herbal medicines [5]. DHIs commonly occur during the pharmacokinetic and pharmacodynamic interactions in prescribed drugs, dietary supplements, or a small portion of food items [6]. In oncology studies, pharmacokinetics interactions can metabolize enzymes like cytochrome P450 (CYP) and P-glycoprotein (P-gp), while pharmacodynamics interactions refer to drugs that influence each other's effects directly. However, excessive DHIs can lead to unexpected Adverse Drug Reactions (ADRs). For instance, a herb that interacts with cisplatin to cure cancer is the Black Cohosh [7].

Information on DHIs can be obtained from the World Wide Web (WWW). However, medical professionals like doctors, pharmacists, medical researchers, and others require an automatic solution to gather the information from articles, databases, and other websites. Therefore, a web crawler is proposed as a solution to accumulate all the information. Thus, the focus web crawler seeks pages that satisfy the relevant information related to the search topics [8,9]. The focused crawler retrieves the maximum number of relevant pages simultaneously and transverse the minimum number of irrelevant pages on the website [10–12]. The focus web crawler also indexes the website entry where the users can

send the index via query and provide the results of the website that matches with the query.

Information on DHIs are scattered since many databases are available to store the information, including Medical Literature Analysis and Retrieval System Online (MEDLINE), and PubMed [13]. A majority of the healthcare professionals prefer to search for research and case reports on DHIs in databases like MEDLINE, PUBMED, EMBASE, and COCHRANE libraries using the following search terms or combinations thereof: "drug–herb interaction," "herb-drug interaction," "interaction," "cytochrome P450," "plant," "extract," "medicinal," "concurrent administration," and "herbal and orthodox medicines." Appropriate search terms were used to represent numerous medicinal herbs used in Africa, America, Asia, Europe, and Australia. This study searched and compiled interaction reports between orthodox medications and their mechanisms of action. The searches were not restricted by publication date or location, but only considered publications in the English language [14,15]. PubMed and MEDLINE contain journals and articles on experiments and studies conducted by medical professions, while the other resource websites such as WebMD, HerbMed, and Natural Medicines Comprehensive Database provided information related to DHIs. Meanwhile, some journal articles like PubMed and MEDLINE require an account subscription to be able to download and read the papers. Therefore, medical professionals face limited time and access to information on DHIs from various websites, as some websites require a purchased subscription.

There are various types of supplements and pharmaceutical companies in Malaysia that manufacture supplements by mixing drugs and herbs regardless of the interactions and ADRs, prescribed and approved by the Drug Control Authority (DCA) and National Pharmaceutical Regulatory Agency (NPRA). Whereas, some people consume drugs and herbs as alternative treatments without consulting their doctors. People are unaware that they can obtain information on DHIs from websites and databases. Hence, a web crawler is proposed in this study to help extract relevant information on DHIs. This study allows readers and researchers from different backgrounds to explore more on the DHIs and web crawlers. Moreover, the web crawler can also download and extract information efficiently and faster. Therefore, a web crawler is the best solution to be implemented in the medical field. This study aims to perform web crawling from several herbal medicine websites related to DHIs and to evaluate web crawler algorithms for DHIs.

The most crucial part of a focused crawler is the selection of the URLs. The primary goal of an effectively focused crawler is to locate relevant web pages and guide them to those pages. Here, classification is widely accepted as the most common method for determining relevant and irrelevant pages. However, classification is not used when DHI websites and databases mostly contain DHI related information and even related URLs also require an indexing algorithm to sort the most preferred websites and databases. Therefore, a focus web crawler approach and indexing algorithms were proposed in this study to identify the most important websites and databases. Page URLs were divided into two categories by indexing algorithms namely primary websites and hyperlinks. To improve the indexing algorithms, this study set different depths for various main pages based on their content ability. The greater the number of pages and the higher weights for the main website can be achieved with a higher number of depths. Meanwhile, newly improved pseudocode was developed by indexing algorithms to improve the algorithm convergence. The performance of focused crawling is directly influenced by the method used to select URLs. This strategy allows the crawlers to find relevant web pages. This study picked sites from the unvisited list, and sorted them in an ascending manner relevant to the given topic of the page being visited. When determining the link weights during crawling, the current page's anchor text, context, and URL string are all considered [16–18]. The most frequently accessed links' features indicate the user's current location to assess their trends and patterns towards a site. After dividing the web page into sections, we evaluated each section as a single content block. Previously unvisited URLs were also extracted and added to the frontier where applicable, with the weight assigned based on its importance. Then, all of the content block links were removed.

This study specifically seeks to make three key contributions. Firstly, the study assessed the topic of ADRs, DHIs and Web Crawler, while prior studies on web crawling largely focused on its benefits for healthcare professionals. It is very important to study the associated costs thoroughly before initiating research. This study focused on the conflicts and pressures created by DHIs which could impact public health. It also focused on the adaptation of web crawlers in reducing costs and improving the efficiency of the web crawlers in DHIs. Therefore, issues and phenomena unique to DHIs are focused on in this study along with their interactions between each other and the outcomes. It also assessed the existence of side effects or ADRs, the implementation of web crawlers in DHIs, sorting of the heterogenous websites and databases, highlighting our extant understanding of DHIs, web crawler processes and outcomes derived from the ADRs and web crawler literature. Secondly, this study adds to the understanding of role theory especially regarding focused web crawling because it is the latest technology adopted by researchers. Thirdly, we contributed to the growing evidence of selecting the best indexing algorithms by comparing their performances. Although previous literature investigated other moderators for DHIs, they often focused on DHIs' research methods by obtaining random information manually. Instead, this study focused on DHIs, an aspect of self-regulation by medical professionals that is easily accessible, therefore could reduce the cost and time spent for obtaining information about DHIs.

The rest of the study is organized as follows: Section 2 describes related studies on ADRs, DHIs, and web crawler methods. Section 3 elaborates the research methodology in terms of the implementation and

experimental design, while Section 4 discusses the analyses outcomes. Lastly, Section 5 concludes this study.

2 Background and Related Work

ADRs originate from the term Adverse Events. Adverse Events are defined by National Care Institute as unexpected symptom that occurs during treatment or therapy (Figure 1) [19]. Adverse Events that occur when a patient consumes a drug excessively, might be defined as ADRs. ADRs refer to the negative or harmful responses due to medication [20]. ADRs which occur due to excessive consumption of conventional medicine with conventional medicine or herbal medicine with conventional medicine is also known as overdose. ADRs could affect adults, children and infants too. The World Health Organization (WHO) defines medication errors as failed treatments that could harm the patients [21,22]. Medication errors could occur during the medication process, choosing a medicine, errors in writing the prescriptions, using the wrong formula by the manufacturer, etc. [23–26].

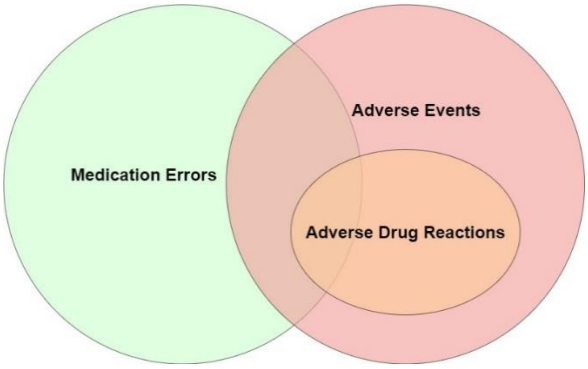


Figure 1: Relationship between medication errors, adverse events, and ADRs.

Several recent studies on ADRs were published by researchers from the United States, Malaysia, and other developed countries. Most of the ADR reports were obtained from MEDLINE, PubMed, etc. [27]. However, the studies revealed that ADRs do not occur only due to the overdose of drugs but also due to discontinuing drug therapy. Another study assessed the pervasiveness of ADEs and reported that 5.1% of the clinic affirmations were expected to be ADRs [28]. They reported that 5.3% of the admissions were caused by ADRs [28]. The percentage of patients who got admitted to the hospital increased due to ADRs [29].

In Malaysia, ADR reports are submitted to the Malaysian Adverse Drug Reactions Advisory Committee (MADRAC) [30,31]. The study also grouped causality into five, namely Certain, Probable, Possible, Unlikely, and Unclassifiable, to classify the ADR reports from MADRAC.

For some medications, the DHIs could also lead to ADRs, especially with herbal products. According to a study, 39 out of 46 herbal medical products interacted with conventional drugs [32]. For instance, Perforate St John's-wort is known as Hypericum perforatum, a prevalent natural item highlighted for its administration as an anti-depression have also been broadly considered for pharmacokinetic DHIs. Hypericum perforatum can be purchased over the counter and is consumed by patients with different pathologies. Hence, it is discouraged. In clinical preliminaries, Hypericum perforatum is known to better stimulate movement compared to fake treatment with a lower dose. Even though the concentrates of Hypericum perforatum (a leafy herb that come from the Hypericaceae family) contains a few phytochemicals and hyperforin, the dynamic energizer specialist is involved in interceding DHIs [33]. Compared to different phytochemicals, hyperforin is bioavailable in humans, where the amount should be halved every 12 hours. It should also take into account the aggregation that deserves attention in many parts of the human body. Hyperforin enacts the Pregnane X Receptor (PXR), an atomic receptor found solely in the liver and digestive tract. Having demonstrated an EC50 estimation of 23 nM with a 380 nM top plasma focus, and a 200 nM enduring state plasma levels reachable in people ingesting the standard routine of 900 mg hypericum perforatum (typically in three separated dosages), hyperforin activates PXR under the fixation to be achieved in human plasma. Similarly, different preclinical investigations have revealed the capacity of *Ginkgo biloba* concentrates to repress human metabolic proteins including CYP1A2, CYP2C9, CYP2E1, and CYP3A4 [34,35]. Most modern medicines contain expected effects including their toxicity when interacting with herbs. Drug toxicity can occur when someone consumes multiple drugs at a time. It specifically occurs when the dose consumed by the individual exceeds the prescribed dose (intentionally or accidentally).

Table 1: Effects of toxicity when interacting with DHIs

No.	Effects of toxicity when interacting with DHIs.
1.	Dizziness
2.	Low/high blood pressure
3.	Low/high sugar level
4.	Eczema
5.	Bruise

Table 1 lists some examples of toxic effects that may occur through the interaction of drugs and herbs. Patients who are in the age range of 33 to 78 years might experience bleeding while consuming certain drugs (i.e. aspirin, warfarin, acetaminophen, and ergotamine-caffeine) with the *Ginkgo biloba*. Most of them experienced major or minor bleeding, while some died due to massive cerebral haemorrhage. Meanwhile, patients who consume products containing Vitamin E with excessive *Ginkgo biloba* could experience an increased rate of platelet activity. The related side effects include blurred vision, headache, and dizziness.

Focus web crawler which is also known as topical web crawler is a technique that only collects web pages that satisfy specific properties. Focus web crawler can analyze the crawler's boundary to determine the most relevant links and avoid unnecessary or irrelevant regions

of the web. Focus web crawler aims to find pages that satisfy the relevant information related to particular topics [8,9]. The focused crawler retrieves a maximum number of relevant pages simultaneously and excludes the minimum number of irrelevant pages on the web [10,11]. Previous studies on focus web crawlers used keyword matching or regular expression matching. The focus web crawler was introduced using the Fish Search algorithm. Fish Search algorithm is an algorithm that stimulates crawling using a group of fish that migrates with the web [36]. Each of the crawled URLs is compared to a fish because survivability relies on visited page pertinence and remote server speed. Page importance is assessed utilizing a double order by employing straightforward catchphrase or customary articulation matches. The fish dies if it navigates through a specific number of unrelated pages. However, studies have tried to improve oriented crawlers to effectively collect related information. The depth parameter could restrict the crawler to not visit a site, which is not important for the searching fish. The relationship between URLs were given importance or priority value based on the similarity found on Shark Search. The priority value is estimated based on the degree of similarity. Although most studies attempt to improve targeted crawlers to gather related information effectively, the parameter depth will limit the visitor to not visit a site that is not meant for searching fish. Based on the similarities identified on Shark Search, the relationships are assigned through the significance or priority status. The target interest is determined by the degree of similarity. Hence, a keyword-focused crawler was proposed by Agre et al. [37]. The study implemented an approach that dealt with domain ontology to find the most relevant pages according to the user requirements. Domain ontology is used to filter out the repository information. The advantages of keyword web crawler over traditional web crawler are that it works intelligently and efficiently without requiring relevant feedback. Consequently, the crawler workload is reduced. On the other hand, a focus crawler system for automation webpage classification was proposed by Goyal [38]. This study aims to determine whether the web pages consist of information of Indian original faculty working in foreign universities. They introduced the automation webpage of the Indian faculties through methods of URL filtering using feature extraction, a genetic algorithm (GA)-based classification. A mutation algorithm was employed to calculate the number of feature extraction. NetBeans IDE 6.9.0 was chosen as the software to execute the implementations. The URLs were selected from the faculties and university websites using keywords. The tags and terms used as the feature extraction were named as chromosomes. The genetic algorithm-based classifier that was implemented used six steps, which are coding, generation of the initial population, evaluation of initial population, selection, crossover, and mutation. Their performance was analyzed in terms of document matrix (ranging from 0 to 1). A precision score higher than 0.8 considers a page to be relevant. An efficient focused web crawler searches for medical plants and relevant diseases using several algorithms such as Naïve Bayes Classifier Algorithm, Decision Tree Algorithm and Multilayer Perceptron [12]. Naïve Bayes classifier was employed to determine whether the current web page was relevant or is not related to the medicinal plant information. This method was proven to perform better. The three types of lexical features, which are title-feature, meta-description, and anchor text, were also implemented for the Naïve Bayes classification. Moreover, a simple decision tree algorithm was developed to determine the relevancy of the medicinal plant URLs. Three different techniques were applied and analyzed ("yes" represents related medicinal plant and "no" represents not related medicinal plant) for each of the medicinal plant URL. The accuracy of Naïve Bayes produced 90% accuracy compared to the other algorithms.

Meanwhile, another study proposed a technique called keyword focused web crawler [39]. This study aimed to improve the performance of web crawler by exploring in depth of the relevant web to the topic. This approach was proposed to extract keywords based on URLs or criteria regarding Indonesian recipes to obtain the best search. This scheme only downloaded URLs which contained Indonesian recipes from the searched keywords. They also used some metrics such as link analysis algorithm including Breadth First Search (BFS) and other URL prioritizing techniques to rank the URLs. The technique did not find the relevant web pages through any other branches as the parent node was related to "milk tea recipe" and "tilapia recipe". The results indicated that the resultant data was much higher than the information path which contained the word "fried chicken recipe".

Another study proposed the essentials of the pre-processing task in social network user behavior [40]. This study aimed to analyze the user's structural behaviour by implementing network link-based properties only. This study employed the BFS algorithm to traverse a range of nodes of the entire social network, where the vertices were put into a specific database format. The result indicated that the BFS successfully sorted the links according to their priority. Meanwhile, a different study employed a novel edge-based parallel algorithm based on the Shiloach-Vishkin (SV) approach for distributed memory systems [41]. This study aimed to reduce the data volume and balance the load as the iterations progress. They implemented a Hybrid Approach, consisting of Parallel SV and Parallel BFS. The graph nodes of parallel BFS were grouped using power-law degree distribution before being implemented into the parallel SV algorithm. This study also implemented Edison, a Cray XC30 machine. The speedup of the machine was measured. The results indicated that the speedup achieved a maximum speed up by 8 times higher than the default value of 16. In another study, a measurement graph of Swarm social, an Online Social Networks (OSN) application on the mobile phone was proposed [42]. The purpose of this study was to provide a comprehensive view of a mainstream OSN that consists of tens of millions of nodes. They created the social graph for Swarm, calculated the key graph metrics, clustering coefficient, assortative, PageRank, connected components and communities. They also implemented BFS in this study to queue the selected user and their

profile information. Moreover, they also implemented Metropolis-Hastings Random Walk (MHRW) and BFS with different subgraphs. The tool used for the implementation was the C++ programming language. Case studies were also performed by sampling the 1%, 5% and 10% of all nodes to calculate the mean and variance. The results indicated that the number mean and variance for BFS was higher than that of MHRW.

Authors in a different study improved the network models design by implementing a local PageRank algorithm [43]. They aimed to measure the influence of a single article regardless of the specialities of the field. They modified the PageRank algorithm by defining the value of $\lambda = 0.1, 0.15, 0.2$, and 0.25. This algorithm was used to rank the co-citation graphs in scientometrics. In co-citation networks, nodes represent articles and the edges represent the citation of articles. Then, the score of the PageRank is computed for each of the nodes. A score of more than 0.8 indicated that the articles ranked by PageRank were relevant. Similarly, another study reviewed a PageRank algorithm [44]. In the study, PageRank searched the web pages based on inbound and outbound hyperlinks.

Another study assessed the improvements of weighted PageRank [45]. This study aimed to determine the ranking popularity of the pages based on the user's usage trends and browsing behavior. They implemented the weighted PageRank based on the links visited. This algorithm assigned a higher-ranking value to the outbound links for the most node visits due to higher popularity compared to inbound links. Hence, they introduced three methods, weight calculator, relevance calculator, and weighted PageRank algorithm based on content and link visits (WPRCLV) calculator in the study. The result demonstrated the efficiency of PageRank in ranking the relevant pages. Whereas, a novel approach to finding topical authorities on Twitter named FAME was proposed by another group of researchers [46]. This study implemented personalized PageRank to deploy a variant query-dependent on FAME. This study chose a suitable feature such as Twitter users' relationships to perform PageRank. The experiment exhibited the improvement of FAME for the authorities from Twitter.

As such, many studies attempted to improve oriented crawlers for the effective collection of related information. However, the depth parameter may restrict the crawler to not visit a site that is not important for searching fish. The URLs can be given the importance or priority value based on the similarity found on Shark Search. The priority value can be estimated based on the degree of similarity.

Based on the literature, the focus web crawler, BFS and PageRank are active research topics in gathering crawling the text information. Since DHIs can lead to pros and cons to our body and health, people need to know relevant information regarding the DHIs. The Internet is a very powerful tool because it is the main element of information growth and dissemination. It is used by billions of people every day from around the world for opportunities and information by people from all walks of life, especially researchers. There are various databases and websites on DHIs available on WWW. This study employed the focus web crawler as the suitable type of web crawler because it only crawls information regarding a specific topic. Tin line with a previous study [39], this study also implemented the addition of an extension to the web-crawling algorithm called PageRank. PageRank was chosen for this study because it was described to be able to rank relevant information, produce more accurate results, and take less time to execute the program.

3 Research method

The focus web crawler was enhanced to search for information on DHIs. This study aimed to ease the algorithms to index the URLs. As such, BFS and PageRank indexing methods were employed to index the URLs according to their priority. Figure 2 illustrates the proposed methodology of the efficient focused web crawler.

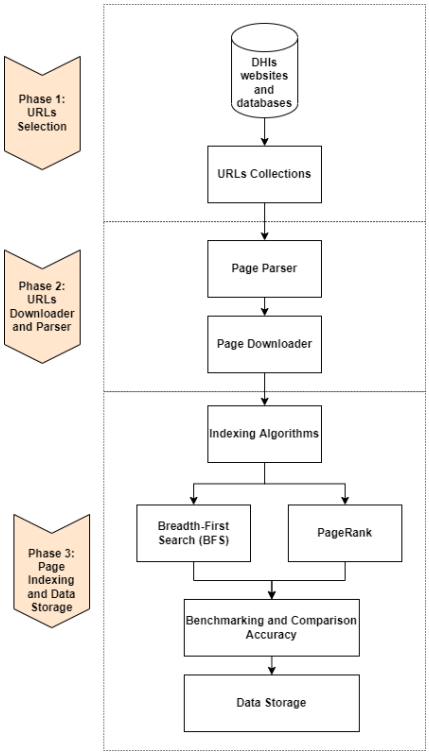


Figure 2: Proposed methodology of focus web crawler.

3.1 URLs selection and collection

There are various websites and databases holding information on DHIs. In phase 1, 5 different published websites or databases on the internet were randomly checked one by one whether or not they contained information on DHIs through the Google search engine (Table 2). Medical professionals commonly use most of these websites to retrieve information on DHIs. Some of the websites listed in Table 2 have a list of databases related to DHI and not related to DHIs. URLs related to DHIs will be collected from the main websites and databases.

Table 2: Lists of websites and databases.

No.	Name of Websites and Databases

1.	Medline Plus Database
2.	Western Botanical Medicine
3.	Global Information Hub on Integrated Medicine (GlobinMed)
4.	Chinese Med Digital Projects
5.	Countway Library of Medicine

3.2 Page parsers

Upon completing page collection, the HTML tags parsing the URLs were fetched from the page collections, followed by a page downloader to extract relevant information before storing the contents of those pages into the disk. DHI information is submitted to crawler downloader that containing the name, type of interaction etc. by a collection of pages used for data download, indexing and storage processes. The page parsers also submitted the information to determine the BFS and PageRank of the last crawled page. Page parsers information index the URL to check whether the URLs have enough crawl information based on its priority.

3.3 Page downloader

The page downloader fetches the URLs and puts them in the URLs queue to download the corresponding relevant pages from the web. The page downloader contains a domain to download the relevant pages. This domain is used to send the domain request and proceed with downloaders. The domain needs to set a timeout to ensure that it does not take too much time to read large pages or wait for the response of web servers. Robot Exclusion Protocol is an important step that needs to be considered for crawl page files because it provides a mechanism to the webserver. Thus, the webserver administrator can determine which pages cannot be accessed by the web crawlers. Meanwhile, the crawler used to exclude robots from a server is called robot Exclusion Protocol. This method creates a file on the server, where, this created page file must be accessible via local URLs or "robots.txt". A crawler can only check whether the pages can be downloaded or not with the approval of robots.txt. Figure 3 displays the examples of crawler.txt implemented in a web crawler. This example indicated that crawlers of other pages and public files are allowed to specify the address of the folder:/other/public/folder to facilitate the crawler's search to crawl relevant information. The page files contain cache to increase the efficiency of crawling. Therefore, it can avoid re-duplicating the page files when downloading the main pages from the same server.

```
User-agent: *
Disallow: /private/
Disallow: /confidential/
Disallow: /other/
Allow: /other/public/
```

Figure 3: Examples of crawler.txt

3.4 Page indexing and data storage

In this phase, the URLs are indexed by implementing two algorithms (BFS and PageRank) and data storage from the page extraction stage.

3.4.1 Indexing algorithms

3.4.1.1 BFS

BFS uses the frontier as a First in First Out (FIFO) queue in which the URL collection was arranged in the order they were encountered. When the FIFO queue is full with URL collection, the crawler added only one link from a crawled page. The BFS crawling method is illustrated in Figure 4. The pseudocode of BFS is summarized in Table 3 [47].

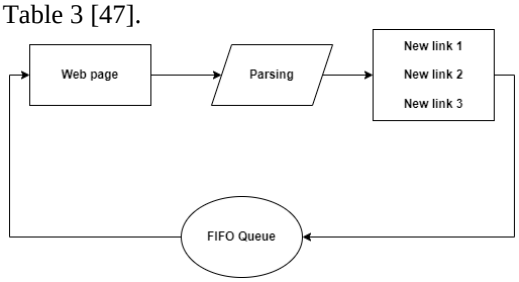


Figure 4: BFS crawling methods.

Table 3: BFS pseudocode.

PSEUDOCODE 1: INDEXING THE URLS USING BFS

```
Input: get_source
Output: links, urls, rank
if depth=4
    return
    print source, depth
    page_source = get_source(source)
    links = Set(findId(page_source))
else
    print 'some error encountered'
    return
end if
Repeat for link in links:
    if link not in urls:
        urls = urls ([link])
    end if
end for
Repeat for link in urls:
    Rank= (link,depth+1)
end for
Print output.
end
```

3.4.1.2 PageRank

Table 4: PageRank pseudocode for focus web crawler

PSEUDOCODE 2: INDEXING THE URLS USING PAGERANK ALGORITHM FOR FOCUS WEB CRAWLER	
	Input: urls, pages
	Output: hyperlinks, rank
1	initialize array of urls
2	initialize pages in integer
3	fetch hyperlinks from urls
4	**if** pages >0
5	Fetch hyperlinks from urls
6	**Repeat for** hyperlinks in urls
7	**if** hyperlinks exists
8	put hyperlinks in queue for rank
9	rank the number of hyperlinks
10	**end if**
11	**end for**
12	**end if**
13	print output
14	**end**

Besides BFS, PageRank algorithm is popularly used to index websites and to determine a page's relevance. The PageRank algorithm was introduced by the founders of Google, Brin and Page. PageRank uses probability distribution to represent the user's behavior [48]. PageRank can be calculated for collections of URL pages of different sizes. PageRank needs several passes or also known as "iterations" via the collection to adjust the approximate PageRank values to reflect the accurate theoretical value. The PageRank in the web crawler is calculated as a sum of the PageRank of all the pages that are linked to each other divided by the number of links on each of those pages. Equation 1 represents the formula for PageRank [48]. The pseudocode for PageRank is highlighted in Table 4. [47].

$$PR(A) = (1-d) + d(\frac{PR(T1)}{C(T1)}) + \cdots + PR(\frac{Tn}{C(Tn)})$$

Equation 1:PageRank equation.

Where

- PR(A) is the PageRank of page A (main page).
- PR(T1) is the PageRank of pages T1 which is linked to page A (child page).
- C(T1) is the number of outside links from page T1.
- d is a damping factor with the range 0 < d < 1, and is usually set to 0.85.

3.4.2 Data storage

Data storage is also an important process for search engines for future use. There are two types of data storage for crawled data namely memory-based and disk-based storage. Once the page extraction and page parsers are conducted, the data storage is performed to store all the extracted information from the websites on the disk. There are several ways to store the data including the non-relational databases depending on the structures of data, JavaScript Object Notation (JSON) files, Comma-separated values (CSV) files or Extensible Mark-up Language (XML) files. While data that consists of DHIs information (herbal name, herbal URLs, herbal description, herbal interactions, and the levelness of DHIs) are stored on disk in JSON file format for future references and usages. Each herbal name in the databases and websites contain the levelness of interaction, for instance, American ginseng interacts majorly with warfarin, hence, needs to be consumed with caution. This American ginseng data will be collected and stored in a database. The data containing URLs along with the rank numbers computed by BFS and PageRank are stored in SQLite for visualization purposes.

3.4.3 Evaluation

The quality of the system was evaluated based on certain criteria, that is accuracy. Accuracy is measured between two indexing algorithms that contain information of DHIs based on their priority and effectiveness. Accuracy determines the best algorithm for this study. The number of pages that contain related information about the interaction of drugs and herbs along with the number of pages that contain relevant and irrelevant DHIs information is required to calculate the accuracy of the indexing algorithms that has been indexed. Accuracy for each indexing algorithm indicates the number of downloaded DHI websites and databases apart from the number of relevant websites for each keyword. Then, the accuracy of all crawlers were compared to each other for all keywords. Accuracy indicates the efficiency of the web crawler in crawling the pages. The calculation of the number of pages retrieved from the main website and the number of pages retrieved from hyperlinks must be done first before calculating the accuracy of the indexing algorithms. Equation 2 represents the formula for accuracy.

$$Accuracy = \frac{number\ of\ related\ pages}{number\ of\ retrieved\ pages}$$

Equation 2: Accuracy equation.

4 Experimental analysis and results

In the proposed method, the DHIs thesaurus information was used for query expansion to induce more web content associated with the user query. The herbal dataset was extracted from an authorized website, downloaded, indexed, and stored in a structured format.

This medicinal plant database consisted of information on 4,744 herbals (Figure 5) from 24 websites and databases (Table 5). The effectiveness of the focus web crawler for DHIs was compared using two algorithms, BFS and PageRank.

Table 5: Herbals extracted from the DHIs databases.

No.	Names of DHIs databases	Number of herbals extracted that are related to DHIs
1.	MedlinePlus.	167
2.	National Center for Complementary and Integrative Health.	52
3.	Chinese Herbal Medicine Database.	408
4.	Western Herbs.	79
5.	Medicinal Herbs & Plant Database (Consumers)	79
6.	South Africa Herbs.	14
7.	Ayurveda Herbs.	25
8.	Native American Herbs.	31
9.	Essential Oils.	74
10.	Alternative Nature Online Herbal.	68
11.	Chinese Medicine Specimen Database.	859
12.	Medicinal Plant Images Database.	1159
13.	Chinese Medicinal Material Images Database.	420
14.	A Modern Herbs.	44
15.	The Raintree Tropical Plant Database.	180
16.	Longwood Herbal Tasks Force.	71
17.	Memorial Sloan Kettering Cancer Centre.	274
18.	HerbMed Pro	130
19.	HerbClip Online.	6
20.	Healthy Ingredients.	107
21.	The Commission E Monographs	120
22.	Herbal Medicine: Expanded Commission E.	107
23.	South Central America Herbs	100
24.	Medicinal Herbs and Plant Database.	200
	Total:	4,744

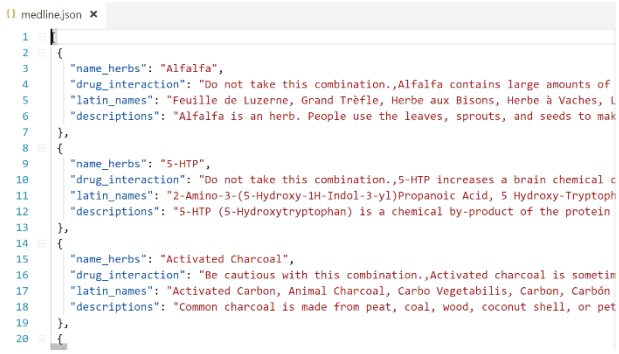


Figure 5: Sample of medicinal plant database consisting of herbal information.

Table 6: Lists of keywords consisting of main element biological transport and biological action.

List of Main Keywords (Keywords 1)
Drug-herb interactions
Drug-food interactions
Botanical medicine
Herbal medicine
Plant medicine
Traditional medicine
Alternative medicine

List of Biological Transport Keywords (Keywords 2)
P450 Cytochromes
Organic anionic transporters
Organic cationic transporters
P-glycoprotein
Drug transporters
Organic anion transporting polypeptide
ABC: ATP binding cassette transporter superfamily
SLC: solute-linked carrier transporter family
MDR1: multi-drug resistance
BCRP: breast cancer resistance protein

List of Action Keywords (Keywords 3)
Substrate, inhibitor, inducer
extracts, bioactive compounds

Keywords 1 contained main elements to perform user query, while keywords 2 consisted of biological transporter and keywords 3 contained action keywords to perform as a 'bridge' to link keywords 1 and keywords 2. For instance, Drug-herb interaction inhibitor P450 Cytochromes as tabulated in Table 6. Each keyword underwent a stemming process, to ease the root words and synonym words, for instance, Drug-herb interaction inhibits P450 Cytochrome. Logical rules, AND and OR were also implemented to crawl the exact DHIs, for instance, Drug-herb interaction AND inhibit AND P450 Cytochrome.

The working of BFS is very simple. It operates on the first come first serve basis. It starts with the https://medlineplus.gov/druginfo/herb_All.html

hyperlinks and is printed as 0 as it is the parent node. https://www.nlm.nih.gov/m, https://nccih.nih.gov/health/echinacea/ataglance.htm, and other hyperlinks are the child nodes. BFS queues all the hyperlinks and continues to update the hyperlinks until there are no more hyperlinks inside the website.

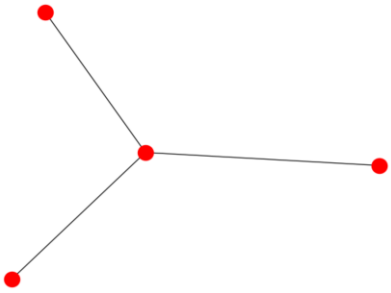


Figure 6: BFS visualization.

Based on Figure 6, BFS is visualized based on the results of the rank. For instance, https://medlineplus.gov/druginfo/herb_All.html is the parent node (right side). The hyperlinks in https://medlineplus.gov/druginfo/herb_All.html are the child nodes (middle and left side). The only limitation of BFS is that it cannot draw nodes if there are more than one similar ranking when the hyperlinks are different.

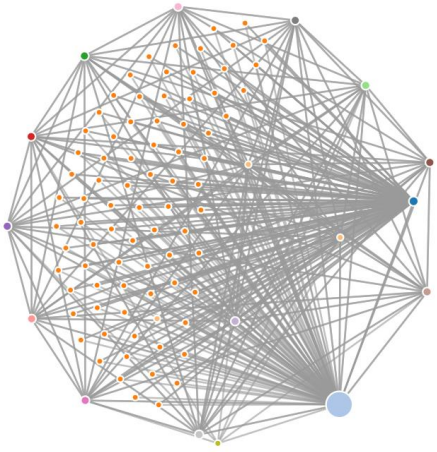


Figure 7: PageRank visualization.

Figure 7 illustrates the graph of PageRank. The biggest circle in light blue color represents the parent node. The other circles in orange color are the child nodes consisting of hyperlinks of the DHIs information. The size of the circle (big or small) depends on the computation page weightage for each of the hyperlinks. The page weightage is calculated based on the priority of hyperlinks. The bigger the size of the circle, the higher the priorities of the hyperlinks. PageRank also has some limitations. Firstly, the graph will become messier if the number of hyperlinks increases. Secondly, PageRank can only draw one big graph for one website or database. So, it cannot calculate the page weightage of more than one URL of a website or database. All the extracted herbs information is stored in JSON file format for future references and usage. All the ranking webpages are stored in SQLite.

Based on

Figure 8, the accuracy of each DHIs main website was different due to the different number of hyperlinks in the websites. Chinese Med Digital Projects and MedlinePlus recorded the same highest accuracy for both the indexing algorithms with 98% for PageRank and 71% for BFS. This is due to the number of hyperlinks of Chinese Med Digital Projects and MedlinePlus are higher than the other main websites. Whereas, Western Botanical Medicine website recorded the lowest accuracy, with 88% for PageRank and 58% for BFS. The lowest accuracy was achieved because of the missing hyperlinks that map into the other websites due to the removal of the domain name or unavailable domain name for the public. Based on Figure 9, PageRank for the GlobinMed website took the shortest time to execute the program compared to BFS. Meanwhile, the American Botanical Council website took the longest time for BFS compared to PageRank. In general, PageRank is faster than BFS as the number of hyperlinks in main websites increases. Based on the results, it can be concluded that PageRank has higher accuracy and is faster compared to BFS due to a few factors. 1) PageRank is generated using the entire internet graph, rather than a small set, it is less susceptible to localised linkage than other ranking systems, 2) a single indicator of a page's quality at the time of the crawl is coupled with a standard information retrieval score for the period, and 3) ranking is based on a page's popularity, therefore, it delivers the most relevant results. Therefore, PageRank is a better performer than BFS.

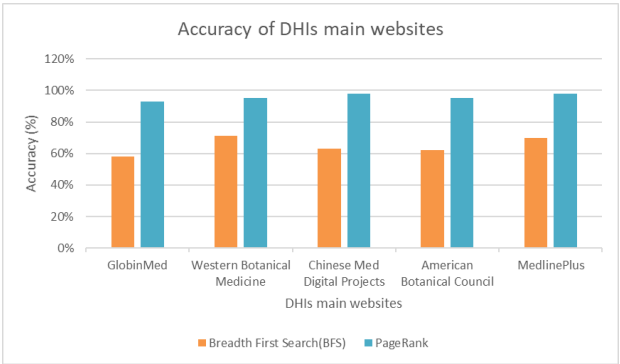


Figure 8: Accuracy of BFS and PageRank for DHIs main websites.

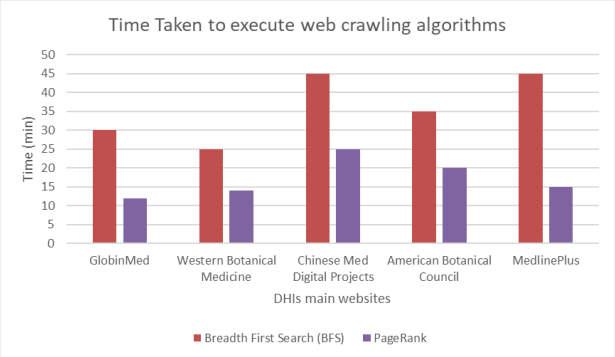


Figure 9: Time taken for web crawling algorithms to execute the program.

5 Conclusion and future work

This study proposed a framework for an efficient focus web crawler to organize and manage a large number of different herbal information and their interactions with drugs. In this proposed framework, the higher the accuracy of the indexing method, the higher the priority of relevant information of DHIs collected from databases and websites. In this study, BFS and PageRank were utilized on the dataset, where accuracy was also calculated. PageRank was more accurate for all the main websites. This focused web crawler provided accurate information while gathering relevant information on DHIs.

Based on the outcome, the indexing algorithms, time is taken for a crawler, and the visualization graph for indexing algorithm can be improved in the future. Firstly, more indexing algorithms could be used as this study only utilized two algorithms. Secondly, the BFS algorithm needs to be modified in terms of increasing the indexing depth of websites and databases. Next, the time taken to execute has to be reduced especially for modified BFS compared to modified PageRank to suit focus web crawler. Besides that, BFS consumes a lot of memories when more URLs are added into the queue as it also ranks some of the irrelevant information of DHIs. Hence, to overcome this limitation, the number of URLs of the DHIs webpages needs to be limited to increase efficiency, produce a better visualization, and rank results.

6 Acknowledgement

This project was supported by Smart Challenge Fund (SMART Fund) - SR0917Q1027 Ministry of Science, Technology and Innovation Malaysia (MOSTI).

7 References

[1] A. Fugh-Berman, E. Ernst, Herb-drug interactions: Review and assessment of report reliability, Br. J. Clin. Pharmacol. 52 (2001). https://doi.org/10.1046/j.0306-5251.2001.01469.x.

[2] R. Hooda, Herbal drug interactions - a major safety concern, Res. Rev. J. Pharmacogn. Phytochem. 4 (2016).

[3] B. Li, B. Zhao, Y. Liu, M. Tang, B. Lüe, Z. Luo, H. Zhai, Herb-drug enzyme-mediated interactions and the associated experimental methods: a review, J. Tradit. Chin. Med. 36 (2016). https://doi.org/10.1016/s0254-6272(16)30054-1.

[4] J.J. Bruno, J.J. Ellis, Herbal use among US elderly: 2002 National Health Interview Survey, Ann. Pharmacother. 39 (2005). https://doi.org/10.1345/aph.1E460.

[5] I. Meijerman, J.H. Beijnen, J.H.M. Schellens, Herb-Drug Interactions in Oncology: Focus on Mechanisms of Induction, Oncologist. 11 (2006). https://doi.org/10.1634/theoncologist.11-7-742.

[6] I. Cascorbi, Drug interactions - Principles, examples and clinical consequences, Dtsch. Arztebl. Int. 109 (2012). https://doi.org/10.3238/arztebl.2012.0546.

[7] N.C. for C. and I. Health, Herb-drug interactions, 355 (2015). https://doi.org/10.1016/S0140-6736(99)06457-0.

[8] M. Diligentit, F.M. Coetzee, S. Lawrence, C.L. Giles, M. Gori, Focused crawling using context graphs, in: Proc. 26th Int. Conf. Very Large Data Bases, VLDB'00, 2000.

[9] B. Novak, a Survey of Focused Web Crawling Algorithms, Proc. SIKDD. 5558 (2004).

[10] C. De Groc, Babouk: Focused web crawling for corpus compilation and automatic terminology extraction, in: Proc. - 2011 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2011, 2011. https://doi.org/10.1109/WI-IAT.2011.253.

[11] R. Gaur, D.K. Sharma, Review of ontology based focused crawling approaches, in: ICSCTET 2014 - Int. Conf. Soft Comput. Tech. Eng. Technol., 2016. https://doi.org/10.1109/ICSCTET.2015.7371191.

[12] N. Pawar, K. Rajeswari, A. Joshi, Implementation of an Efficient web crawler to search medicinal plants and relevant diseases, in: Proc. - 2nd Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2016, 2017. https://doi.org/10.1109/ICCUBEA.2016.7860006.

[13] Y. Qian, X. Ye, W. Du, J. Ren, Y. Sun, H. Wang, B. Luo, Q. Gao, M. Wu, J. He, A computerized system for detecting signals due to drug-drug interactions in spontaneous reporting systems, Br. J. Clin. Pharmacol. 69 (2010). https://doi.org/10.1111/j.1365-2125.2009.03557.x.

[14] H. Ibrahim, A. Saad, A. Abdo, A. Sharaf Eldin, Mining association patterns of drug-interactions using post marketing FDA's spontaneous reporting data, J. Biomed. Inform. 60 (2016). https://doi.org/10.1016/j.jbi.2016.02.009.

[15] M.L. Rethlefsen, MEDLINE: A Guide to Effective Searching in PubMed and Other Interfaces, J. Med. Libr. Assoc. 95 (2007). https://doi.org/10.3163/1536-5050.95.2.212.

[16] Review of various web page ranking algorithms in web structure mining, Int. J. Adv. Eng. Res. Dev.

3 (2015). https://doi.org/10.21090/ijaerd.ncrretcs20.

[17] C.J. Luh, S.A. Yang, T.L.D. Huang, Estimating Google's search engine ranking function from a search engine optimization perspective, Online Inf. Rev. 40 (2016). https://doi.org/10.1108/OIR-04-2015-0112.

[18] A.E. Wibowo, K.M. Lhaksmana, M. Isd, Perbandingan Peformansi Terhadap Algoritma Breadth First Search (BFS) & Depth First Search (DFS) Pada Web Crawler, E-Procceding Eng. 6 (2019) 9905–9914.

[19] NCI, NCI Dictionary of Cancer Terms, (n.d.). https://www.cancer.gov/publications/dictionaries/cancer-terms/def/adverse-event (accessed October 19, 2018).

[20] S. Scott, J. Thompson, Adverse drug reactions, Anaesth. Intensive Care Med. 15 (2014). https://doi.org/10.1016/j.mpaic.2014.02.008.

[21] J.K. Aronson, Medication errors: Definitions and classification, Br. J. Clin. Pharmacol. 67 (2009). https://doi.org/10.1111/j.1365-2125.2009.03415.x.

[22] WHO, Electronic Tools: Technical Series on Safer Primary Care., WHO Press. (2016) 1–21.

[23] J.K. Aronson, Medication errors: What they are, how they happen, and how to avoid them, QJM. 102 (2009). https://doi.org/10.1093/qjmed/hcp052.

[24] I.R. Edwards, J.K. Aronson, Adverse drug reactions: Definitions, diagnosis, and management, Lancet. 356 (2000). https://doi.org/10.1016/S0140-6736(00)02799-9.

[25] A.M. Mayo, D. Duncan, Nurse perceptions of medication errors what we need to know for patient safety, J. Nurs. Care Qual. 19 (2004). https://doi.org/10.1097/00001786-200407000-00007.

[26] M. Shamna, C. Dilip, M. Ajmal, P. Linu Mohan, C. Shinu, C.P. Jafer, Y. Mohammed, A prospective study on Adverse Drug Reactions of antibiotics in a tertiary care hospital, Saudi Pharm. J. 22 (2014). https://doi.org/10.1016/j.jsps.2013.06.004.

[27] J.R. Nebeker, P. Barach, M.H. Samore, Clarifying Adverse Drug Events: A Clinician's Guide to Terminology, Documentation, and Reporting, Ann. Intern. Med. 140 (2004). https://doi.org/10.7326/0003-4819-140-10-200405180-00009.

[28] S. V Taché, A. Sönnichsen, D.M. Ashcroft, Prevalence of Adverse Drug Events in Ambulatory Care: A Systematic Review, Ann. Pharmacother. 45 (2011). https://doi.org/10.1345/aph.1p627.

[29] A. Krähenbühl-Melcher, R. Schlienger, M. Lampert, M. Haschke, J. Drewe, S. Krähenbühl, Drug-related problems in hospitals: A review of the recent literature, Drug Saf. 30 (2007). https://doi.org/10.2165/00002018-200730050-00003.

[30] USER929, Introduction to MADRAC, (n.d.). https://www.npra.gov.my/index.php/en/about/malaysian-adverse-drug-reactions-advisory-committee-madrac/madrac-introduction (accessed October 19, 2018).

[31] H. See Lei, A.A. Fatah Rahman, A.M. Haq Syed Haq, Adverse drug reaction reports in Malaysia: Comparison of casuality assesmentds, Malaysian J. Pharm. Sci. 5 (2007).

[32] P. Posadzki, L. Watson, E. Ernst, Herb-drug interactions: An overview of systematic reviews, Br. J. Clin. Pharmacol. 75 (2013). https://doi.org/10.1111/j.1365-2125.2012.04350.x.

[33] A. Nahrstedt, V. Butterweck, Lessons learned from herbal medicinal products: The example of St. John's wort, J. Nat. Prod. 73 (2010). https://doi.org/10.1021/np1000329.

[34] P.C. Chan, Q. Xia, P.P. Fu, Ginkgo biloba leave extract: Biological, medicinal, and toxicological effects, J. Environ. Sci. Heal. - Part C Environ. Carcinog. Ecotoxicol. Rev. 25 (2007). https://doi.org/10.1080/10590500701569414.

[35] C. Gaudineau, R. Beckerman, S. Welbourn, K. Auclair, Inhibition of human P450 enzymes by multiple constituents of the Ginkgo biloba extract, Biochem. Biophys. Res. Commun. 318 (2004). https://doi.org/10.1016/j.bbrc.2004.04.139.

[36] P. de Bra, G.-J. Houben, Y. Kornatzky, R. Post, Information Retrieval in Distributed Hypertexts, RIAO. (1994).

[37] G.H. Agre, N. V. Mahajan, Keyword focused web crawler, in: 2nd Int. Conf. Electron. Commun. Syst. ICECS 2015, 2015. https://doi.org/10.1109/ECS.2015.7124749.

[38] N. Goyal, R. Bhatia, M. Kumar, A genetic algorithm based focused web crawler for automatic webpage classification, in: IET Conf. Publ., 2016. https://doi.org/10.1049/cp.2016.1546.

[39] G.A.F. Alfarisy, F.A. Bachtiar, Focused web crawler for Indonesian recipes, in: Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017, 2018. https://doi.org/10.1109/SIET.2017.8304134.

[40] K. Das, S.K. Sinha, Essential pre-processing tasks involved in data preparation for social network user behaviour analysis, in: Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2017, 2018. https://doi.org/10.1109/ISS1.2017.8389423.

[41] C. Jain, P. Flick, T. Pan, O. Green, S. Aluru, An Adaptive Parallel Algorithm for Computing Connected Components, IEEE Trans. Parallel Distrib. Syst. 28 (2017). https://doi.org/10.1109/TPDS.2017.2672739.

[42] Y. Chen, J. Hu, H. Zhao, Y. Xiao, P. Hui, Measurement and Analysis of the Swarm Social Network with Tens of Millions of Nodes, IEEE Access. 6 (2018). https://doi.org/10.1109/ACCESS.2018.2789915.

[43] A. London, T. Németh, A. Pluhár, T. Csendes, A

local PageRank algorithm for evaluating the importance of scientific articles, Ann. Math. Informaticae. 44 (2015).

[44] A. Vishwakarma, R. Saxena, M. Awasthi, M. Yamuna, Comparative analysis of PageRank and hits: A review, Int. J. Pharm. Technol. 8 (2016).

[45] R. Prajapati, S. Kumar, Enhanced weighted PageRank algorithm based on contents and link visits, in: Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016, 2016.

[46] P. Lahoti, G. De Francisci Morales, A. Gionis, Finding topical experts in twitter via query-dependent personalized PageRank, in: Proc. 2017 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2017, 2017. https://doi.org/10.1145/3110025.3110044.

[47] D. Shestakov, Intelligent Web Crawling, IEEE Intell. Informatics Bull. 14 (2013) 5–7. http://www.comp.hkbu.edu.hk/~iib/2013/Dec/article1/iib_vol14no1_article1.pdf (accessed August 1, 2019).

[48] I. Rogers., The Google PageRank Algorithm and How It Works, (n.d.).

An Improved Bagging Ensemble in Predicting Mental Disorder Using Hybridized Random Forest - Artificial Neural Network Model

Oluwashola David Adeniji[1], Samuel Oladele Adeyemi[2], Sunday Adeola Ajagbe[3,4*]
[1]Computer Science Department, University of Ibadan, Ibadan, Nigeria
[2] Clinical Nursing Department, University College Hospital, Ibadan, Oyo State
[3] Computer Engineering Department, Ladoke Akintola University of Technology, LAUTECH, Ogbomoso, Nigeria
[4] Computer and Industrial Production Engineering, First Technical University, Ibadan.
E-mail: od.adeniji@ui.edu.ng, samadeyemi10@yahoo.com, saajagbe@pgschool.lautech.edu.ng (Corresponding author's email)

Keywords: Hybrid model, Machine learning (ML), Mental disorder, Mental health, RF-ANN

Received: January 16, 2022

Machine Learning majorly provides the process of collecting, identifying, pre-processing, training, validating and visualization of data. This study identifies the problem of late detection of mental disorders in IT employees. There are many cases of mental disorders that are not apparent, notable or diagnosed until they become critical. This affects the productivity of the employees not only in the information technology (IT) industry. The objective of the study is to develop a Hybrid Random Forest (RF) and Artificial Neural Network (ANN) model to predict mental health disorders among employees in the IT industry. The experiment applied a hybrid Random Forest and Artificial Neural Network (RF-ANN) model in predicting the chances of IT employees developing mental disorders. To measure the performance of the model, RF and ANN algorithms were separately developed, their results were recorded and compared with the results of the hybrid model. In the hybrid model using "Bagging Ensemble," the prediction of an IT employee developing a Mental Disorder shows the weighted average performance of 84.5% for precision, recall, and accuracy and precision is 82.5% using the hybridized RF and ANN models on "Bagging Ensemble". This result obtained from the hybrid model correctly shows a significant improvement in its performance over individual performances of the RF model and ANN models. There was a marginal improvement in the performance of the hybrid model when compared with the result of the parameter-tuned RF. This suggests that by applying the RF-ANN model an improved dataset could be investigated and compared with the results obtained in this study.

Povzetek: Članek se ukvarja z napovedovanjem mentalnih težav s pomočjo metod globokih nevronskih mrež, naključnih gozdov in vrečastega ansambla.

1 Introduction

Mental health is defined as a person's psychological, social, and emotional state when they are functioning at an acceptable level of behavioural and emotional adjustment. Mental health can be viewed as a measure of an individual's ability to handle stress and make decisions in all aspects of their life, as it has a significant impact on how such an individual act, thinks and feels. Mental health is an important factor at any stage of life, whether it is adulthood or childhood [1]. According to the World Health Organization, depression is the leading cause of Mental Health Disorders worldwide, affecting individuals as well as communities. It is estimated that more than 350 million people worldwide suffer from depression as of 2020 [2, 3]. Mental health issues have a significant impact on workplace productivity, not only for the individual but also for the organization as a whole. Unfortunately, people generally find it difficult to discuss mental health issues in public, and society does not raise enough awareness. The current evolution of Machine Learning (ML) solutions has resulted in automated models that can predict, classify, and diagnose some of the issues associated with mental health disorders.

The alarming trend of rising mental health problems, combined with the global inability to find effective solutions, was impeding both individual and societal prosperity. There were numerous and significant barriers to accessing mental health care, ranging from socioeconomic inequalities to personal stigmas. This provides an opportunity for technology, particularly artificial intelligence (AI)-based technology, to help alleviate the situation and provide numerous unique benefits. Kolenik & Gams, (2021) [4] provided a brief overview of persuasive technology (PT) for mental health, as well as general, technical, and critical thoughts on implementation and impact in terms of potential benefits and risks. While potential benefits identified in the research include; cost, availability and stigma. Group exclusion and research bias were identified as the PT risk. We believe that such technology can supplement existing mental health care solutions by reducing access inequalities as well as those caused by a lack of it.

In recent years, ML techniques have been adopted in numerous medical researches, especially in biomedicine and neuroscience to gain further insight into mental health disorders [5]. Machine learning, being an area of artificial intelligence involves the process of computers learning from data through the use of heuristic algorithms [6, 7]. ML is divided into two types: supervised ML and unsupervised ML. Supervised ML models are typically used to assign a set of attributes to a target class, which implies classification and regression. Unsupervised ML models are used to describe the relationship or characteristics of a set of attribute data. Unsupervised ML primarily necessitates the processes of feature selection, clustering, and association rule mining [8]. Studies show that employees in IT industries are at high risk of developing mental disorders due to increased stress and pressure to meet targets and deadlines. In many cases, these disorders are not obvious, known or diagnosed until they become life-threatening.

The existing studies in different fields have implemented various machine techniques to predict mental disorders. However, there is a need to address the issue of late detection of mental disorders in IT employees. Developing automated models that can predict, diagnose and classify mental health disorders is now possible with the help of computer-aided systems [9]. By using these developed models, they help in saving manpower, time and other resources, while also removing the possibilities of human bias. A large amount of data is readily available thanks to the advancement in the usage, power and capacity of the latest computer technologies. This has resulted in an increase in the ability to collect, store and manipulate data. Subsequently, knowledge can be extracted from the data by bringing out patterns and relationships through the development of a methodology. Such methodology can be developed from a database of existing tools and methods available for the discovery of knowledge and data mining [10, 11, 12].

A hybrid model of neural network (NN) with a random forest (RF) structure can produce a result with improved generalization ability and accuracy. The ability of this hybrid architecture to reduce the back-propagation algorithm to a more powerful and generalized decision tree structure makes it more effective than random forests. In addition, this model is more efficient to train as the number of training examples usually requires only a small constant factor [12, 13]. Therefore, this study aims to develop a model that can predict the chances of IT employees developing mental disorders using a hybrid of the two best performing models in previous studies consisting of RF and ANN. The developed hybrid model was evaluated using standard metrics in this study area such as precision, recall, and F1-score. The organization of this paper is as follows: the introduction is in section one, review of the literature is contained in section two. Section three is the methodology and the result and discussion were highlighted. Finally, the conclusion is contained in section 5.

2 Review of literature

Mental health disorders, also known as mental illnesses, refer to a variety of mental health conditions that affect a person's thinking, mood, and behavior. Anxiety disorders, depression, addictive behaviors, schizophrenia, eating disorders, and other mental disorders are examples. Many people experience various mental health issues from time to time. However, these mental health issues only become a mental disorder when the ongoing signs and symptoms cause frequent stress and impair a person's ability to function effectively. Loss of pleasure or interest, poor concentration, loss of appetite, disturbed sleep, feelings of guilt, and low energy are all symptoms of mental health disorders. These problems have the tendency to become chronic and recurrent, and thus impair a person's ability to take care of their daily responsibilities [14]. According to [15], more than 30% of people suffering from major mental disorders do not seek treatment, while more than 80% of people battling with some form of mental disorder do not seek to be treated at all. Variations of mental illnesses Depression, bipolar disorder, schizophrenia and other psychoses, dementia, and developmental disorders are all examples of mental illnesses.

Machine Learning and Healthcare: Precision medicine is a way in which healthcare professionals can move to more personalized care by adopting ML in finding patterns and reasons about data [16]. With the large volume of data being collected about patients in the healthcare sector, it is near impossible for humans to analyze. With sufficient data and permission to use, there are numerous ways in which ML can be applied in healthcare. In times past, hard-coded software has been developed based on external studies to provide recommendations and alerts for different medical practices. The limitation to this however is the problem with the accuracy of data due to other factors such as location, environment, population, and so on. With ML, data can be refined to a particular environment, for instance, refining data from a hospital and the surrounding environment in a way that the patient's information is anonymized. Examples of ways in ML can help healthcare providers include: the prediction of a possible outbreak of disease, predicting the possibility of hospital readmission for critically ill patients, prediction of cancer risks in patients, and so on [9].

According to the study by Groves et al. (2013) [17], being able to identify patients that are most liable to the risk of hospital readmission helps healthcare providers to offer better support after discharge. The lives of those at risk are improved when the rate of readmission is lowered, and this can be made possible with the intervention of ML. Implementation of artificial intelligence in healthcare organizations as a response to the needs of doctors to aid the patients in their daily decision-making activities is now on the increase. This hopes to improve decision making and reduce errors. In the long run, it reduces cost and improved workflow and the general well-being of people.

Related works:

The Internet of Things (IoT), which refers to the integration of technology into everyday life and the interconnectivity of omnipresent devices, has stymied a dedicated research venture in the field of mental health. Recognizing that mental health issues are on the risen, affecting individuals and society in increasingly complex ways and that existing human resources are insufficient to address the crisis, decision-makers have turned to technology to see what opportunities it may provide. The role of IoT-enabled technology in this new digital mental health landscape can be divided into two complementary processes: assessment and intervention [18].

Prediction of mental health problems in children using eight ML techniques, three of which, multilayer perceptron, multiclass classifier, and logical analysis of data (LAD) tree, produced more accurate results with only a slight difference between their performances over the full attribute set and the selected attribute set. The study found that by developing a high-performing model, early diagnosis of mental health problems in children will help healthcare professionals to treat it at an earlier stage and subsequently improve the quality of patients' life. Therefore, there comes an urgent need to treat basic mental health problems that persist among children which may lead to complicated problems, if not treated at an early stage [19]. By introducing a genetic algorithm (GA) in developing a system for intelligent data mining and ML for mental health diagnosis, Azar, et al., (2015) [20] were able to extract keywords from the user's symptoms. The research introduced a new approach that was used for a semiautomated system that helps in the preliminary diagnosis of the psychological disorder patient. This was achieved by matching the description of a patient's mental health status with the mental illnesses. The study constructed a semi-automated system based on an integration of the technology of genetic algorithms, classification data mining and ML. The goal was to help psychological analysts make informed, appropriate and intelligent assessments leading to accurate prognoses by ensuring that they are aware of all possible mental health illnesses that could match the patient's symptoms.

The predictive research for mental health disease was proposed in a prototype that used RF classification to determine the mental state of a person based on attributes such as lifestyle, age, education, gender, vision, occupation, sleep, personal income, mobility, diabetes and hypertension [21]. With the amount of data produced by humans daily and with most of this data stored in a semi-structured way, these researchers believed that by using this ML technique, hidden patterns can be found between the different attributes of data. WEKA and RATTLE were used and the result of 83.33 % and 92.85. % accuracy was reported. With these, the system would be able to predict whether a patient was suffering from mental illness or not. A critical review using SVM to identify imaging biomarkers of neurological and psychiatric disease was conducted by [22]. The study provided an overview of the method and reviewed studies that applied SVM in the investigation of schizophrenia, Alzheimer's disease, Parkinson's disease, bipolar disorder, pre-symptomatic Huntington's disease, major depression, and an autistic spectrum disorder. Standard univariate analysis of neuroimaging data revealed a host of neuroanatomical and functional differences between healthy individuals and patients suffering a wide range of psychiatric and neurological disorders.

The mental health evaluation model based on the fuzzy neural network was carried out by [23] by selecting the important factors such as the input vector, the model was used to evaluate the psychological health of college students in China. The combination of neural networks (NN) and fuzzy mathematics improved the accuracy of the mathematical model compared to other traditional models and made it easy to analyze the overall mental health trend of students. Recurrent and linear models to detect depression early were developed [24]. The goal of the study was to achieve early automatic detection of depression from users' posts on the social media site – Reddit. For prediction, both sequential (RNN) and non-sequential (SVM) models were used. The results showed the superiority of sequential models over nonsequential models. The research did not sufficiently explore the broad range of possible features. Different ML techniques such as KNN, SVM, naïve bayes classifier, decision trees, and logistic regression to identify the state of mental health in a target group. The replies to the designed questionnaire from the target group were first exposed to unsupervised learning techniques. The Mean Opinion Score was used to validate the labels obtained by clustering. The cluster labels were then utilized to create classifiers that could predict an individual's mental health. Population from a wide range of groups such as college students, high school students, and working professionals were considered as target groups.

A survey on the analysis of the mental state of social media users to predict depression was conducted by [25]. The survey was done to detect depression and mental illness through the use of social media are surveyed. They found out that there was a very high rate at which depression and mental illness were being diagnosed in recent times. They observed that some symptoms linked to mental illness were detectable on Facebook, Twitter, and web forums. They suggested that using automatic methods would help in locating inactivity and other mental diseases. Various automated detection methods could help to detect depressed people using social media. Mentally ill users were pointed out through the use of screening surveys, their Twitter analysis based on community distribution, or their membership in online forums, and they were detectable through the patterns in their language and online activities. Additionally, they observed that a number of authors experienced that numerous activities on social networking sites could be linked to low self-confidence, especially in young people and adolescents.

A predictive model for the determination of the risk of depression among university students was also developed by [10]. The study extracted knowledge on the factors causing depression among university students. In the study, a predictive model for depression risk with a view to determining the risk of depression among university students was formulated, simulated and validated.

The result of the study identified variables that have strong relevance to developing depression among university students. The simulation results showed that the model without feature selection gave a total of 465 correct classifications out of 507 records with an accuracy of 91.7% while feature selection, gave a total of 475 correct classifications with an accuracy 93.7 %. It has been established that the are research gaps in the area of using ML techniques to provide solutions to some of the issues relating to mental disorders among the IT employee, hence, this study was informed. Table 1 shows the summary of the related works, the goals and their respective contributions

S/N	Ref	Goals	Contribution
1	[18]	Identify approaches in digital mental health	Distinguished technology for mental health diagnosis into two complementary processes: assessment and intervention
2	[21]	RF (ML) classifier was used alongside predictive models to determine mental health diseases.	WEKA and RATTLE were used as predictive models, 83.33 % and 92.85 accuracies were reported.
3	[20]	To introduce GA in developing intelligent data mining and ML system for mental health diagnosis, was able to extract keywords from the user's symptoms.	Matched the description of a patient's mental health status with the mental illnesses. The research introduced a new approach that was used for a semiautomated system that helps in the preliminary diagnosis of the psychological disorder patient.
4	[19]	Prediction of mental health in children	Multilayer perceptron, Multiclass classifier, and LAD tree outperformed other ML models used.
5	[22]	To identify imaging biomarkers of neurological and psychiatric disease using SVM was conducted	Method and reviewed studies that applied SVM in the investigation of schizophrenia, Alzheimer's disease, Parkinson's disease, bipolar disorder, pre-symptomatic Huntington's disease, major depression, and an autistic spectrum disorder were reported.
6	[23]	To evaluate mental health based on the fuzzy neural network	The combination of NNs and fuzzy mathematics improved the accuracy of the mathematical model compared to the existing method.
7	[24]	To achieve early automatic detection of depression from users' posts on the social media site RNN and SVM	The approach achieved early automatic detection of depression and ensured superiority of sequential models over nonsequential models
8	[25]	to conduct a survey on detecting depression and mental illness by the use of social media information.	They found out that there was a very high rate at which depression and mental illness were being diagnosed in recent times. They observed that some symptoms linked to mental illness were detectable on Facebook, Twitter, and web forums.
9	[10]	To depression risk with a view to determining the risk of depression among university students using predictive models	93.7% accuracy was achieved on the ML model used

Table 1: Summary of related works

3 Methodology

The developed model consists of the phases namely: the dataset pre features extraction (training and testing), and the evaluation results. During the preprocessing phase, values of missing data were replaced and even distribution of data was ensured with features scaling. Features of the pre-processed data were split into two with 67% of the data set aside for training and the remaining 33% set aside for testing. Using the bagging ensemble method, the training set was passed through hybridized RF and Artificial Neualgorithms, using RF as the base. The performance of the result was then evaluated using the standard evaluation metric namely: precision, recall, f1-score and accuracy. Figure 1 shows the developed methodology framework in this research, the framework showing the stages followed to achieve the results. The selection gave a total of 475 correct classifications.

3.1 Data collection and implementation

The input data for this model is the dataset provided by OSMI Ltd. The dataset is derived from a survey aimed at measuring the attitudes of people towards mental health in IT workplace and examining the rate of mental disorders set contains 63 variables or columns and 1,433 responses/ observations. The performances of the models

built were evaluated on the basis of their precision, recall, and F1-score. To further understand and measure the performance of the hybrid model, a different set of algorithms was implemented. These models include: RF with default parameters, RF with tuned parameters and the developed model consists of three different phases namely: the dataset pre-processing, features extraction (training and testing), and the processing phase, values of missing data were replaced, even distribution of data was ensured with features processed data were split into two with 67% of the data set aside for training and the remaining 33% set aside for testing. Using the bagging ensemble method, the training set was passed through hybridized RF and ANN.

4 Results and discussions

The performances of these models were observed and compared with the hybrid model. The results obtained from the trained models are discussed below:

Random forest technique*:* The result obtained from the model as shown in Table 2, though the model performance was poor in the classification of IT employee with the status of mental health disorder with 44%, 44%, and 45% precision, f1 accuracy, recall respectively. The default-parameterized RF model performed at a weighted average of 48% precision, 49% F1-score and 50% recall respectively. This poor performance called for the need to tune the parameter in order to obtain improved performance. This outcome attests to the study of [22].

Table 2: Result of the random forest model

	Precision	**F1-Score**	**Recall**	**Support**
0	0.23	0.19	0.17	103
1	0.53	0.53	0.53	183
2	0.56	0.60	0.65	157
Macro average	0.44	0.44	0.45	473
Weighted average	0.48	0.49	0.50	473
Accuracy	-	0.50	-	473

Artificial neural networks (ANN) techniques: The result of ANN model is as shown in Table 3 poorly incorrectly classifying “Employees not sure of their mental health status”. Overall, the model performed at a weighted average of 72% precision, 69% F1 score and 71% recall.

Table 3: Result of the ANN Model

	Precision	**F1-Score**	**Recall**	**Support**
0	0.52	0.36	0.28	103
1	0.90	0.79	0.70	183
2	0.65	0.77	0.95	187
Macro average	0.69	0.64	0.64	473
Weighted average	0.72	0.69	0.71	473
Accuracy	-	0.71	-	473

RF-ANN technique: RF and ANN algorithms were hybridized for the Bagging ensemble issue. The hybrid model was trained on the training set and its performance was evaluated on the testing set. The result of the hybridized model in table 4 showed a significant improvement in the performance of the hybrid model over the performances of the RF model with default parameters and ANN model and a slight improvement over the per parameter-tuned RF ANN model’s weighted average performance was improved by the hybrid model from (72%, 69% and 71%,) to (74%, 72% and 74%) precision, F1-score and recall respectively. The results are similar to the finding in [23-24].

Table 4: Result of the hybrid model

	Precision	**F1-Score**	**Recall**	**Support**
0	0.64	0.48	0.38	103
1	0.84	0.79	0.74	183
2	0.69	0.94	0.94	187
Macro average	0.73	0.69	0.68	473
Weighted average	0.74	0.72	0.74	473
Accuracy	-	0.74	-	473

Back-propagation is the stage in which the weights are adjusted based on the loss in an attempt to find an optimal weight. Training an ANN is a process that entails determining the optimal weight that will minimize loss. In doing so, "categorical cross-entropy loss" was used to find the loss and "adam" optimization was used to find the optimal solution, with "accuracy" as a metric for performance evaluation. As shown, the model was trained with 45 epochs and a batch size of 10. Figures 2 and 3 Depict the training process, with decreasing loss and increasing accuracy as training progresses.

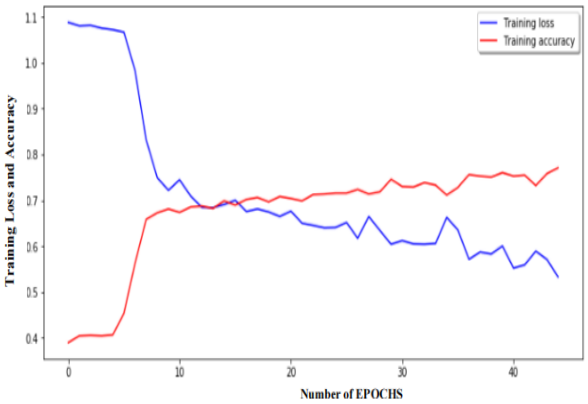


Figure 2: The Training loss and the Training Accuracy Against Number of Epochs

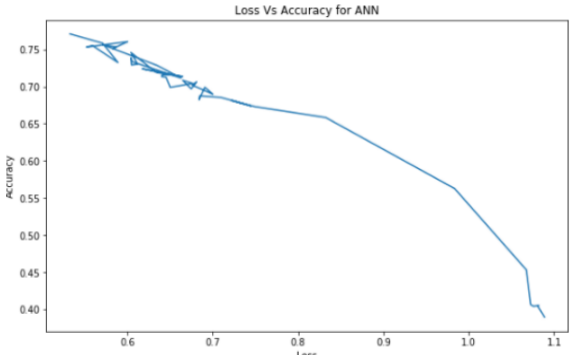


Figure 3: Training Accuracy versus Training Loss Showing Inverse Relationship

Findings: From all the results obtained, Table 5 reveals that the hybrid RF and ANN using "Bagging Ensemble" gave the best performance with the weighted average performance of 84.5% for precision, recall, and accuracy and precision is 82.5%. It can also be observed that the model is able to correctly predict IT employees suffering from Mental Health Disorders with 97% recall. Additional insight has gotten from the results in this experiment reveals that there was only a marginal improvement in the performance of the hybridized model when compared with the result of the parameter-tuned RF. This shows that RF is one of the best classifiers in the predictive algorithm which is consistent with the work of [21] and [25-28].

Table 5: Result Comparison of the 3 Models

Performance metrics	**Recall (%)**	**Prediction accuracy (%)**	**F1 Score (%)**	**Precision**
ANN	81.5	81.5	82.5	82.5
RF	60.5	60.5	58.5	58.5
Developed model (RF-ANN)	**84.5**	**84.5**	**82.5**	**84.5**

A comparison of the existing work with our study is presented in Table 6, showing the comparison of the goals in the existing works with our goals viz-a-viz the respective contributions

5 Conclusion

The study described various approaches to predicting mental disorders. It focused on the development of a hybrid predictive model for determining the risk level of mental disorders among employees in IT industry. Most of the existing models focused on predicting mental disorders using a single ML technique. This study identified the variables measured in IT employees which are relevant to the prediction of mental disorders in the dataset collected. The results obtained showed that developing a hybridized RF-ANN model had the best overall performance, hence, our developed model. The study described various approaches to predicting mental disorders. It focused on the development of a hybridized predictive model for determining the risk level of mental disorders among employees in IT industry. Most of the existing models focused on predicting mental disorders using a single ML technique. However, there is also the need to address the issue of late detection of mental disorders in IT employees. This study identified the variables measured in IT employees which are relevant to the prediction of mental disorders in the dataset collected.

In conclusion, the best performing model to predict mental disorders in employees of IT industry has been identified, the work has been able to develop a predictive model based on the most relevant factors causing mental disorders. From all the results obtained the hybrid RF and ANN using "Bagging Ensemble" gave the best performance in predicting IT employees suffering from Mental Health Disorders with a weighted average performance of 84.5% for precision, recall, and accuracy and precision is 82.5%. meaning there was a marginal improvement in the performance of the hybrid model when compared with the result of the parameter-tuned RF. This suggests that by applying the RF-ANN model an improved dataset could be investigated and compared with the results obtained in this study. In addition, insights gotten from the results in this study reveal that there was only a marginal improvement in the performance of the hybrid model when compared with the result of the parameter-tuned RF. The study is not only contributing to the computing field but also to healthcare delivery.

Recommendations

The following recommendations are made based on the findings of this study: This model can be adopted as an assistant tool by mental health professionals to help them in making an early and more consistent diagnosis of mental disorders. The model can be integrated into an existing employees' Health Information System (HIS) which has clinical data about the employees. It is recommended that variables monitored in IT employees be reviewed on a regular basis in order to increase the amount of information relevant to developing an improved mental health disorder prediction model. In the future, the hybridized algorithm could be used to predict mental disorders in a variety of fields. Similarly, using RF with parameter tuning on a better dataset could be investigated and compared to the findings of this study.

Table 6: Comparison of the existing work with our study

S/N	Ref	Goals	Contribution
1	[18]	Identify approaches in digital mental health	Distinguished technology for mental health diagnosis into two complementary processes: assessment and intervention
2	[21]	RF (ML) classifier was used alongside predictive models to determine mental health diseases.	WEKA and RATTLE were used as predictive models, 83.33 % and 92.85 accuracies were reported.
3	[20]	To introduce GA in developing intelligent data mining and ML system for mental health diagnosis, was able to extract keywords from the user's symptoms.	Matched the description of a patient's mental health status with the mental illnesses. The research introduced a new approach that was used for a semiautomated system that helps in the preliminary diagnosis of the psychological disorder patient.
4	[19]	Prediction of mental health in children	Multilayer perceptron, Multiclass classifier, and LAD tree outperformed other ML models used.
5	[22]	To identify imaging biomarkers of neurological and psychiatric disease using SVM was conducted	Method and reviewed studies that applied SVM in the investigation of schizophrenia, Alzheimer's disease, Parkinson's disease, bipolar disorder, pre-symptomatic Huntington's disease, major depression, and an autistic spectrum disorder were reported.
6	[23]	To evaluate mental health based on the fuzzy neural network	The combination of NNs and fuzzy mathematics improved the accuracy of the mathematical model compared to the existing method.
7	[24]	To achieve early automatic detection of depression from users' posts on the social media site RNN and SVM	The approach achieved early automatic detection of depression and ensured superiority of sequential models over nonsequential models
8	[25]	to conduct a survey on detecting depression and mental illness by the use of social media information.	They found out that there was a very high rate at which depression and mental illness were being diagnosed in recent times. They observed that some symptoms linked to mental illness were detectable on Facebook, Twitter, and web forums.
9	[10]	To depression risk with a view to determining the risk of depression among university students using predictive models	93.7% accuracy was achieved on the ML model used. Only accuracy was measured
10	**Our study**	**to develop a hybrid model to predict mental health disorders among employees in an IT industry**	**An effective hybrid technique was developed that is capable of predicting the mental health of IT industry employees, the model was also evaluated singly before hybridizing and after hybridization, the hybridized model shows improved performance than the single ones**

References

[1] P. Sandhya and K. Mahek, "Prediction of Mental Disorder for employees in IT Industry. *International Journal of Innovative Technology and Exploring Engineering* (IJITEE) 8(6), vol. 8, no. 6, pp. 2278-3075, 2019.

[2] S. S. Lim, T. Vos, A. D. Flaxman, G. Danaei, K. Shibuya, H. Adair-Rohani, M. A. AlMazroa, M. Amann, H. R. Anderson and K. G. Andrews, "A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990 – 2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet,* vol. 380, pp. 2224-2260, 2012.

[3] S. A. Ajagbe and A. O. Adesina, "Design and Development of an Access Control Based Electronic Medical Records (EMR)," *Centrepoint Journal (Science Edition),* vol. 26, no. 1, pp. 98-119, 2020.

[4] T. Kolenik and M. Gams, "Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality?," *IEEE Technology and Society Magazine,* vol. 40, no. 1, pp. 80-86, 2021.

[5] J. F. Dipnall, J. A. Pasco, M. Berk, L. J. Williams, S. Dodd, F. N. Jacka and D. Meyer, "Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression," *PloS one,* vol. 11, no. 2, pp. 148-195, 2016

[6] J. Han, J. Pei and M. Kamber, "Data Mining: Concepts and Techniques," *Elsevier,* 2011.

[7] J. B. Awotunde, S. A. Ajagbe, M. O. Oladipupo, J. A. Awokola, O. S. Afolabi, M. O. Timothy and Y. J. Oguns, "An Improved Machine Learnings Diagnosis Technique for COVID-19 Pandemic Using Chest X-ray Images," in *Applied Informatics. ICAI 2021. Communications in*

Computer and Information Science, Springer, Cham, 2021, p. 14555.
[8] I. R. Idowu, O. D. Adeniji, S. Elelu and T. O. Adefisayo, "Prediction of Breast Cancer Images Classification Using Bidirectional Long Short Term Memory and Two-Dimensional Convolutional Neural network," *Transactions on Networks and Communications*, vol. 9, no. 4, pp. 29-38, 2021.
[9] E. O. Ogunseye, C. A. Adenusi, A. C. Nwanakwaugwu, S. A. Ajagbe and S. O. Akinola, "Predictive Analysis of Mental Health Conditions Using AdaBoost Algorithm," *Paradigmplus,* vol. 3, no. 2, pp. 11-26, 2022.
[10] T. M. Awoyelu, A. R. Iyanda and S. K. Mosaku, "Formulation of a Predictive Model for the Determination of Depression among University Students," in *Proceeding of the 14th International Conference on Smart Nations, Digital Economies and Meaningful Lives*, 2016.
[11] T.P. Olalere O.D. Adeniji "An Artificial Intelligent Video Assistant Invigilator to Curb Examination Malpractice" *International Conference on Innovative Systems for Digital Economy | ISDE'2021, pp* 58-65, 2021.
[12] S. Wang, C. Aggarwal and H. Liu, "RandomForest-Inspired Neural Networks," *ACM Transmission Intelligence Systems Technology, 9, 6(69),* vol. 9, no. 6, pp. 1-25, 2018.
[13] G. Chen, S. Li, H. Long, X. Zeng, P. Kang and H. Zhang, "A Hybrid Algorithm Introducing Cross Mutation and Non-linear Learning Factor for Optimal Allocation of DGs and Minimizing Annual Network Loss in the Distribution Network," *IAENG International Journal of Applied Mathematics*, vol. 51, no. 3, pp. 1-18, 2021.
[14] P. Gilbert, Depression: The Evolution of Powerlessness, Routledge, London, United Kingdom,, 2016.
[15] O. D. Adeniji and J. J. Ukame "A Novel Immune Inspaired Concept with Neural Network for Intrusion Detection in Cybersecurity" *International Journal of Applied Information Systems* vol 12, issue pp. 13-17, 2020.
[16] E.O Alabi, O. D. Adeniji, T. M. Awoyelu and E. O. Fasae, "Hybridization of Machine Learning Techniques in Predicting Mental Disorder," *International Journal on Human Computing Studies*, vol. 3, no. 6, pp. 22-30, 2021, https://journals.researchparks.org/index.php/IJHCS
[17] P. Groves, B. Kayyali, D. Knott and S. V. Kuiken, "The Big Data Revolution in Healthcare," 2013.
[18] T. Kolenik, "Methods in Digital Mental Health: Smartphone-Based Assessment and Intervention for Stress, Anxiety, and Depression," in *Integrating Artificial Intelligence and IoT for Advanced Health Informatics*, Springer, Cham, 2022.
[19] M. R. Sumathi and B. Poorna, "Prediction of Mental Health Problems Among Children Using Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 7, no. 1, 2016.
[20] G. Azar, C. Gloster, N. El-Bathy, S. Yu, R. H. Neela and I. Alothman, "Intelligent Data Mining and Machine Learning for Mental Health Diagnosis using Genetic Algorithm," in *2015 IEEE International Conference on Electro/Information Technology (EIT)*, 2015.
[21] S. Chauhan and A. Garg, "Predictive Research for Mental Health Disease," *Blue Eyes Intelligence Engineering & Sciences Publication*, 2019.
[22] G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori and A. Mechelli, "Using Support Vector Machine to Identify Imaging Biomarkers of Neurological and Psychiatric Disease: A Critical Review," *Neuroscience and Biobehavioral Reviews*, vol. 36, p. 1140–1152, 2012.
[23] L. Jing, "Mental Health Evaluation Model based on Fuzzy Neural Network," in *2016 International Conference on Smart Grid and Electrical*, 2016.
[24] F. Sadeque, D. Xu and S. Bethard, "UArizona at the CLEF eRisk Pilot Task: Linear and Recurrent Models for Early Depression Detection," in *In CEUR workshop proceedings, volume 1866. NIH Public Access*, 2017.
[25] A. Khan, M. H. Husain and A. Khan, "Analysis of Mental State of Users using Social Media to Predict Depression: A Survey," in *Conference Paper at the International Journal of Advanced Research in Computer Science*, 2018.
[26] R. Obiedat and S. A. Toubasi, "A combined approach for prediction employee's productivity based on ensemble machine learning methods," *An International Journal of Computing and informatics*, vol. 46, no. 5, pp. 49-58, 2022. DOI: 10.31449/inf.v46i5.3839
[27] T. P. Olalere and O. D. Adeniji, "An artificial intelligent video assistant invigilator to curb examination malpractice," in *International Conference on Innovative Systems for Digital Economy | ISDE'2021*, 2021
[28] O. D. Adeniji, D. B. Adekeye, S. A. Ajagbe, A. O. Adesina, Y. J. Oguns and M. A. Oladipupo, "Development of DDoS Attack Detection Approach in Software Defined Network Using Support Vector Machine Classifier," in *Pervasive Computing and Social Networking*, 2023.

Estimation of Parameters in Regression Analysis Based on *QR* Decomposition of Rectangular Matrices by Householder Reflections

Oleksandr Dorokhov*[1], Lyudmyla Malyarets[2], Kadri Ukrainski[1] and Dmytro Yevstrat[3]
[1]Chair of Public Economics and Policy, University of Tarty Narva mnt 18, 51009, Tarty, Estonia
[2]Department of Higher Mathematics and Economic and Mathematical Methods, Simon Kuznets Kharkiv National University of Economics Nauka Ave 9A, 61166, Kharkiv, Ukraine
[3]Department of Information Systems, Simon Kuznets Kharkiv National University of Economics Nauka Ave 9A, 61166, Kharkiv, Ukraine
E-mail: oleksandr.dorokhov@ut.ee, malyarets@ukr.net, kadri.ukrainski@ut.ee, dmitryyevstrat@gmail.com
*Corresponding author

Keywords: regression analysis, Householder reflection, *QR* decomposition

Received: February 04, 2022

An approach to eliminate multicollinearity problems in regression analysis using QR decomposition of rectangular matrices by Householder reflection has been proposed. The reliability of this computational procedure has been proved.

Povzetek: V regresijski analizi je narejena ocena parametrov s pomočjo dekompozicije QR.

1 Introduction

In substantiating the decision taken in the management of various socio-economic systems, it is important to choose an analytical tool that analyzes the state of systems and predicts their further development.

In the presence of various kinds of uncertainties and a large amount of various data, in the problems of economic modeling, two main approaches are most often used recently: fuzzy modeling [1-5] and various statistical methods [6-8].

Among the last, most often, multiple regression analysis is used as such a tool [9-13].

Despite its prevalence in economics and duration in applications, many problems still need to be solved, since the existing algorithms for multiple regression analysis are far from perfect [14-17].

To solve many problems of mathematical methods in economics, the so-called *QR* decomposition of rectangular matrices is useful [18-21].

An $n \times m$ matrix X is decomposed into two factors $X = QR$, where Q is an $n \times n$ orthogonal matrix and R is an upper triangular $n \times m$ matrix with zero entries below the main diagonal.

Recall that the determinant of the orthogonal matrix is equal to one $|Q| = 1$; its inverse matrix coincides with the transposed $Q^{-1}=Q'$ (i.e. $QQ' = Q'Q = I$), and orthogonal transformation of vectors a, b does not change their scalar products: $(Qa,Qb) = (a,b)$; $|Qa| = |a| = a$; $|Qb| = |b| = b$.

It can be argued that numerical algorithms with orthogonal transformations do not introduce additional errors into the solution of the problem.

2 *QR* decomposition to solve the problem of multicollinearity

Consider how *QR* decomposition helps to overcome the problem of multicollinearity in regression analysis.

Suppose we need to find the best estimates of the parameters b of a linear three factor regression model from 6 observations ($n = 6$; $m = 3+1 = 4$):

$$Y = Xb + E,$$

which in expanded form reduces to solving the overvalue value of a system of 6 linear equations with respect to 4 unknowns (b_0, b_1, b_2, b_3):

$$\begin{aligned}
y_1 &= b_0 + b_1 \cdot x_{11} + b_2 \cdot x_{21} + b_3 \cdot x_{31} + e_1 \\
y_2 &= b_0 + b_1 \cdot x_{12} + b_2 \cdot x_{22} + b_3 \cdot x_{32} + e_2 \\
y_3 &= b_0 + b_1 \cdot x_{13} + b_2 \cdot x_{23} + b_3 \cdot x_{33} + e_3 \\
y_4 &= b_0 + b_1 \cdot x_{14} + b_2 \cdot x_{24} + b_3 \cdot x_{34} + e_4 \\
y_5 &= b_0 + b_1 \cdot x_{15} + b_2 \cdot x_{25} + b_3 \cdot x_{35} + e_5 \\
y_6 &= b_0 + b_1 \cdot x_{16} + b_2 \cdot x_{26} + b_3 \cdot x_{36} + e_6
\end{aligned}$$

Here E – vector of errors (residuals e_i), which can be found after determining the estimates of the model parameters b_0, b_1, b_2, b_3.

If the *QR* decomposition of the matrix $X = QR$ is known, then the above problem is solved as follows.

Should multiply the matrix equation $Y = QRb + E$ leftward to the orthogonal matrix Q' and we obtain an equivalent equation $Z = Rb + \Xi$, where marked $Z = Q'Y$, $\Xi = Q'E$.

In expanded form we have a system of linear equations with a triangular matrix R:

$$z_1 = b_0 \cdot r_{01} + b_1 \cdot r_{11} + b_2 \cdot r_{21} + b_3 \cdot r_{31} + \xi_1$$
$$z_2 = b_1 \cdot r_{12} + b_2 \cdot r_{22} + b_3 \cdot r_{32} + \xi_2$$
$$z_3 = b_2 \cdot r_{23} + b_3 \cdot r_{33} + \xi_3$$
$$z_4 = b_3 \cdot r_{34} + \xi_4$$
$$z_5 = \xi_5$$
$$z_6 = \xi_6$$

Due to the orthogonality of the matrix transformation, the following relation is preserved $||\Xi|| = ||E||$, that is, the sum of the squares of the converted errors $\sum \xi_i^2$ (vector norm Ξ) is always equal to the sum of the squares of the residuals of the original system of equations $\sum e_i^2$ (vector norm E).

Due to the successful determination of the model parameters b_i you can equate to zero the first few components of the vector Ξ and obtain the minimum value of the sum of the squares of the residuals

$$\sum e_i^2 \to min.$$

So it is possible to obtain estimates of the parameters of the model by the least squares method (but in a slightly different, non-standard computational way, without first drawing up a system of normal equations).

If the last diagonal element of the triangular matrix R is nonzero $r_{34} \neq 0$, then you can equate to zero the maximum number (m) of the first components ξ_i and find estimates of the model parameters from the triangular system of equations.

This also determines the (minimum) sum of the squares of the residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=m+1}^{n} \xi_i^2 = \sum_{i=m+1}^{n} z_i^2 .$$

If the last diagonal element of the triangular matrix R is exactly equal to zero $r_{34} = 0$, then this means that the last variable x_4 is not independent, but it is a linear combination of other argument variables $x_4 = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3$.

In this case, the estimate $b_4 = 0$ should be equated to zero.

That is, it is not necessary to include in the model a combination of already taken into account variables, and estimates of other parameters of the model should be obtained from the conditions of zero of the smaller number of the first components $\xi_i = 0$.

If the last diagonal element of the triangular matrix R is not exactly zero, but close to it $|r_{34}| \approx 0$, this means that there is a multicollinear relationship between the source variables x_i.

In this case, it makes sense to equate the score $\boldsymbol{b_4} = 0$ to zero and remove the questionable term from the model (otherwise an unstable solution with large errors will be obtained).

So, if we have a QR decomposition of the matrix X, the problem of multi-linearity is solved quite simply.

At the same time, the strong side of the QR decomposition is that it allows you to calculate (find a numerical) solution to the least squares problem.

As is well known, the classical, ordinary least squares method gives us a closed solution in the form of normal equations. But this solution is not always suitable for practical, specific applications.

Therefore, if you need to find the actual numerical solution, the least squares method is not suitable, at the same time, the QR decomposition easily gives such numerical values.

It should be noted that today the most well-known and used methods for obtaining an orthogonal matrix Q are the Gram-Schmidt process, the Householder transformation, and the Givens rotation.

An orthogonal matrix Q (and a triangular matrix R) can be obtained by successive operations with matrices $H_k = I - 2\omega_k\omega_k'$, where I – single matrix, ω_k – normalized vector ($\omega_k' \omega_k$=1), in which the first $(k\text{-}1)$ components are zero.

It's not hard to see that $H_k'H_k = H_kH_k = I$, that is, the matrix H_k is orthogonal. This matrix is also called as the Householder reflection matrix.

Any vector can be represented as $a = a_1\,\omega + a_2\,\varpi$, where ω, ϖ – orthonormal vectors ($\omega'\varpi = 0$; $\omega'\omega = 1$; $\varpi'\varpi = 1$).

Householder reflection $Ha = -a_1\,\omega + a_2\,\varpi$ changes the sign of the first component to the opposite.

We have to take into account some important properties of the Householder matrix: it is Hermitian H=H* and unitary HH*=I, and therefore it is an involution H^2=I. In this case, the transformation $H_u(x)$ displays (reflects) point x to point $x\text{-}2(u, x)u$.

The Householder matrix has one eigenvalue equal to (-1), which corresponds to the eigenvector u, while all other eigenvalues are equal to (+1).

In this case, the determinant of the Householder matrix is (-1), and the Householder transformation in the metric space preserves distances.

3 Converting a rectangular matrix to upper triangular

Consider the process of sequential transformation of a rectangular matrix $A = [a_1, a_2, \ldots, a_m]$ to the upper triangular shape (possibly with permutations of columns).

The first transformation with a matrix H_1 must transform the first vector a_1 (the first column of the matrix A) to vector $\pm a_1\,e_1$, where e_1 a coordinate vector in which only one first element is nonzero (this element is equal to 1, the other elements are zero); as a_1 is marked

length of the vector a_1 (because the orthogonal transformation does not change the length of the vector).

Consider the approach of how to find a vector ω_1, which defines the whole matrix H_1.

To do this, consider that again (re-)reflection restores vector a_1:

$a_1 = H_1(\pm a_1 e_1) = (I - 2\omega_1\omega_1')(\pm a_1 e_1) = \pm a_1 e_1 + 2a_1\omega_{11}\omega_1$, hence the vector ω_1 proportional to the vector a_1, to the first component of which a value $\pm a_1$ is added; sign (+ or –) should be selected by the sign of the element a_{11}.

So we get a (still normalized) vector Ω_1 in the form $\Omega_1 = a_1 + a_1 \cdot \mathrm{sgn}(a_{11}) \cdot e_1$.

The square of the length of this vector is equal to $\Omega_1'\Omega_1 = 2a_1(a_1 + |a_{11}|)$.

Thus the first Hausholder matrix is constructed $H_1 = I - \frac{\Omega_1\Omega_1'}{a_1(a_1+|a_{11}|)}$.

In the transformed matrix A in the first column will have only one non-zero element: $H_1 a_1 = a_1 - \Omega_1 = -a_1 \cdot \mathrm{sgn}(a_{11}) \cdot e_1$.

Other converted columns of matrix A (and the dependent variable column Y) have the form: $H_1 a_j = a_j - \lambda_j \Omega_1$, where λ_j – numerical coefficients $\lambda_j = \frac{(a_1,a_j)+a_1 \cdot sgn(a_{11}) \cdot a_{j1}}{a_1(a_1+|a_{11}|)}$

The first elements of the converted columns will no longer change in subsequent ones Householder reflections.

Without these first elements of the norms of all vectors (columns of the transformed matrix A) are reduced on the value a_{jk}^2 (by now $k = 1$).

Let find the column a_q with the highest residual norm (relative to the original norm): $\frac{a_q^2 - a_{qk}^2}{a_q^2} = \max\left(\frac{a_j^2 - a_{jk}^2}{a_j^2}\right)$.

If the largest relative residual norm is less than some limit value (for instance, 0,01), further transformations are canceled, the process ends prematurely due to the detection of multicollinear connections.

In the second stage we find the matrix H_2, which transform vector a_q (without first component) to vector $\pm(a_q)^* e_2$, where e_2 – the second coordinate vector; by $(a_q)^*$ marked shortened vector a_q length without its first component.

Thus, after the second transformation, the vector a_q will have two non-zero components.

Then again we find the residual norms and determine the next vector a_p, which (without the first two components) will be converted to $\pm(a_p)^* e_3$.

The process will end after m iterations, or early if the residual norms become less than accepted limit level.

Usually in regression analysis the first column of the matrix X (matrix A) there is a column of ones $x_0 \equiv 1$ (to take into account in the regression model the obligatory free term).

Therefore, at the first stage (of Householder reflections), it is the first The order of selection of the following columns for conversion is determined by the values of their relative residual norms.

4 An example of the implementation of the algorithm

The described above algorithm was implemented in an electronic spreadsheet *Excel* in the macros form on the *VBA* language (*Visual Basic for Application*).

Consider numerical example of the analysis of the regression dependence of the resultant feature on 5-factors, the corresponding initial data for this are presented in Table.1.

Table 1. Data for regression analysis

№	y	x_0	x_1	x_2	x_3	x_4	x_5
1	1075,3	1	32,06	17,9	12,08	35,61	8,33
2	1002,7	1	27,57	10,23	14,06	37,48	10,63
3	995,6	1	27,88	10,29	11,26	37,77	12,72
4	872,4	1	31,65	11,72	7,32	34,98	14,25
5	909,3	1	34,81	12,64	1,68	39,04	11,75
6	1009,9	1	29,47	10,87	1,31	46,14	12,15
7	919,7	1	34,42	12,77	1,28	41,04	10,48
8	876,2	1	32,76	12,26	1,06	42,53	11,32
9	908,6	1	31,24	11,65	4,49	41,27	11,28
10	935,8	1	30,4	11,33	6,88	40,07	11,26
11	949,9	1	29,96	11,18	8,84	39,48	10,5
12	927,4	1	30,49	11,41	7,73	39,55	10,78
13	1003,9	1	29,71	11,05	13,08	29,46	16,68
14	1017,6	1	29,02	10,79	14,34	29,06	16,77
15	997,6	1	29,55	10,99	11,75	30,07	17,63
16	958,2	1	30,79	11,44	7,94	31,29	18,54
17	907,5	1	32,55	12,08	1,45	33,03	20,87
18	928	1	33,27	12,35	1,41	32,32	20,63
19	930,1	1	35,34	13,42	0,76	32,24	18,23
20	892,6	1	33,71	12,79	0,78	33,58	19,13
21	917,7	1	32,3	12,03	4,26	32,68	18,7
22	947,2	1	31,32	11,64	6,99	31,65	18,38
23	959,6	1	30,97	11,55	8,94	31,24	17,3
24	943,4	1	31,52	11,7	7,63	31,64	17,5

By using Householder reflections matrix X was transformed into a triangular form (to matrix $R = Q'X$).

The process of transformation ended prematurely.

Variables x_1, x_4 was not connected in the model since their residual norms decreased to values of about 0.1% of the original norms as shown in Table 2 and Table 3.

Let move converted columns x_1 and x_4 (with small relative residual rates less than 0.01) to the left to the column Z.

As result, we can obtain a system of 4 equations with 3 columns of free terms (Z, x_1 and x_4) relative to 4 parameters of the considered model b_0, b_2, b_3, b_5 as shown in Table 4.

Table 2: Data after householder reflections

№	$z = Q'y$	x_0	x_1	x_2	x_3	x_4	x_5
1	-4651,2	-4,8	-153,6	-58,3	-32,1	-174,1	-72,6
2	-168,7		7,6	1,4	-22,2	8,3	3,2
3	-18,0		0,7	-1,4		-19,2	18,2
4	62,9		4,5	6,8		-6,2	
5	-3,9		2,0			-2,5	
6	135,5		-2,4			3,4	
7	5,7		1,6			-2,1	
8	-24,3		0,1			-0,2	
9	-12,4		-0,1			0,1	
10	-1,6		-0,0			-0,0	
11	-4,2		0,2			-0,4	
12	-20,3		0,3			-0,5	
13	21,9		0,5			-0,7	
14	28,9		0,3			-0,4	
15	30,8		-0,1			0,1	
16	20,0		-0,3			0,5	
17	21,8		-1,1			1,7	
18	37,0		-0,6			0,8	
19	20,4		1,0			-1,6	
20	-3,4		-0,3			0,3	
21	3,0		-0,2			0,4	
22	13,8		-0,2			0,3	
23	7,9		0,1			-0,2	
24	1,4		0,2			-0,3	

Table 3: Norms

Output	21691,3	24	23707,1	3461,7	1524,9	30845,7	5617,9
Residual	24827,4	0	16,928	0	0	31,7126	0
Relative	0,00114	0	0,00071	0	0	0,00102	0

Table 4: System of equations with a triangular matrix

y	x_1	x_4	x_0	x_2	x_3	x_5
-4651,2	-153,6	-174,1	-4,8	-58,3	-32,1	-72,6
-168,7	7,6	8,3		1,4	-22,2	3,2
-18,0	0,7	-19,2		-1,4		18,2
62,9	4,5	-6,2		6,8		

The solution to this system is given in Table 5. Explanatory variables are shown in the model x_2, x_3, x_5. Other variables (including y) are expressed through these explanatory variables.

Table 5: Inverse matrix and system solution

Model parameters			**Inverse matrix**			
y	x_1	x_4	x_0	x_2	x_3	x_5
790,78	24,03	67,07	-0,20	0,29	-0,86	-1,97
9,14	0,65	-0,90				0,14
8,14	-0,28	-0,60		-0,04	0,01	0,01
-0,25	0,09	-1,13			0,05	0,01

5 The results of calculations

As a result, the following regression models can be obtained (together with the coefficients of determination):

$y = 790{,}78 + 9{,}14 \cdot x_2 + 8{,}14 \cdot x_3 - 0{,}25 \cdot x_5$; $R^2 = 0{,}56$

$x_1 = 24{,}03 + 0{,}65 \cdot x_2 - 0{,}28 \cdot x_3 + 0{,}09 \cdot x_5$; $R^2 = 0{,}82$

$x_4 = 67{,}07 - 0{,}90 \cdot x_2 - 0{,}60 \cdot x_3 - 1{,}13 \cdot x_5$; $R^2 = 0{,}93$

Coefficients of determination R^2 can be calculated, for example, as follows:

$$R^2 = 1 - \frac{||\boldsymbol{e}||}{||\boldsymbol{y}|| - N \cdot (\bar{y})^2} = 1 - 0{,}4312 = 0{,}5688$$

So we have multicollinear relationships of variables x_1 and x_4 with explanatory variables x_2, x_3, x_5 and with multiple correlation coefficients $R_1 = \sqrt{0{,}8251} = 0{,}9084$ and $R_4 = \sqrt{0{,}9382} = 0{,}9686$, which exceeds the closeness of the relationship of the explanatory variables with the resultant feature $R_y = \sqrt{0{,}5686} = 0{,}7542$.

Note that the matrix of paired correlation coefficients r_{xy} does not show (not demonstrate) any effect of multicollinearity at once.

But it turns out that the determinant of this correlation matrix is almost zero $|r_{xy}| = 0{,}0000656$, that is, all explanatory variables are interconnected by a precise multicollinear relationship as shown in Table 6.

So, with the help of QR decomposition of X rectangular matrix it is possible (without first compiling a system of normal equations) to obtain least square methods' estimates of the parameters of the regression model together with all multicollinear connections.

Table 6: Correlation matrix

r_{xy}	x_1	x_2	x_3	x_4	X_5	y
x_1	1	0,5817	-0,7796	-0,0065	0,2102	**-0,5863**
x_2	0,5817	1	-0,2002	-0,0160	-0,1664	**0,1259**
x_3	-0,7796	-0,2002	1	-0,3694	-0,1748	**0,7031**
x_4	-0,0065	-0,0160	-0,3694	1	-0,7743	**-0,1736**
x_5	0,2102	-0,1664	-0,1748	-0,7743	1	**-0,1969**

6 Analysis of the stability of the obtained results

We turn to consider a very important problem of the stability of the results obtained from some particular observations, which have an excessive effect on the values of the model parameters. Using a matrix $\boldsymbol{Q}$ you can detect all such questionable observations, such as observation №1 in the above tables.

If you delete the observation №1, the values of the model parameters will change significantly:

With observation №1:

$y = 790{,}78422 + 9{,}14172 \cdot x_2 + 8{,}14859 \cdot x_3 - 0{,}25244 \cdot x_5$;

Without observation №1:

$y = 1300{,}5080 - 34{,}4492 \cdot x_2 + 1{,}75898 \cdot x_3 + 2{,}25143 \cdot x_5$.

This unpleasant effect is shown in the graphs of component effects in Figure 1 (with), Figure 2 (without).

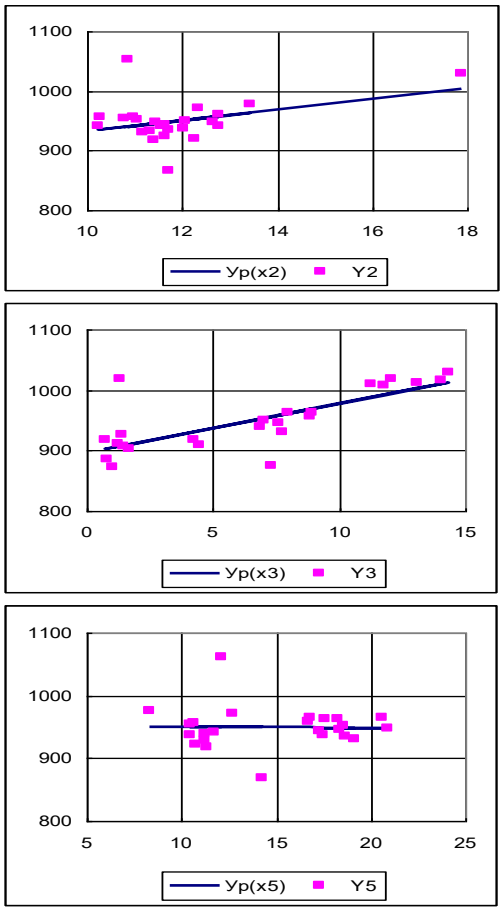
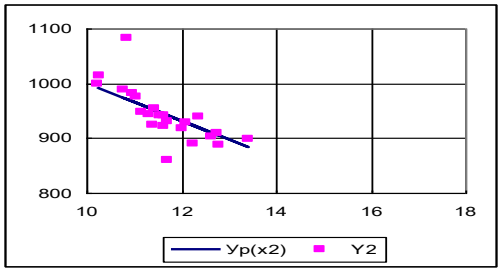


Figure 1: Along with observation № 1.

On the graphs empirical points are superimposed on theoretical regression lines (calculated values). It helps to find out which dependencies are significant and which are not. Indeed, it is always possible to choose such scales of coordinate axes that all theoretical lines will have the same slope in 45°.

The presence of empirical points (or 95% confidence bands) will not allow incorrect conclusions in this case.

However, the question arises, how to determine such empirical points so that only one variable varies on each graph, and the rest is fixed at the average levels?



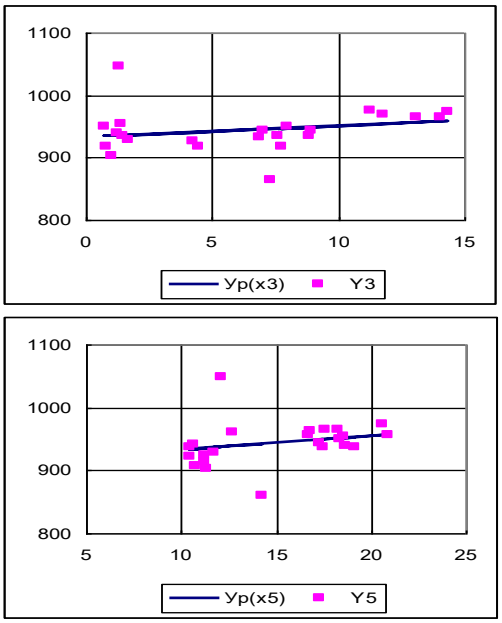


Figure 2: Without observation № 1.

It turns out that this is a problem whose solution is insufficiently covered in the literature. Therefore, we propose such a solution.

After determining the parameters of the model, you can find the deviation of each observation from the theoretical values (residuals): $e = y - y_p$.

Now, for each observation, you can write down the identity:

$$y_i = \bar{y} + b_2(x_{2i} - \bar{x}_2) + b_3(x_{3i} - \bar{x}_3) + b_5(x_{5i} - \bar{x}_5) + e_i$$

Members $b_k(x_k - \bar{x}_k)$ are called "component effects".

Sum $\bar{y}$ and the corresponding component effect is the equation of the theoretical dependence of the resultant feature y on one variable x_k at the average values of the remaining explanatory variables.

We will fix in identities all variables (except one) on average values and we will receive the corrected data as the sum of the total mean $\bar{y}$, the corresponding component effect $b_k(x_k - \bar{x}_k)$ and the residual:

$$Y_2 = Y_p(x_2) + e = \bar{y} + b_2(x_2 - \bar{x}_2) + e;$$
$$Y_3 = Y_p(x_3) + e = \bar{y} + b_3(x_3 - \bar{x}_3) + e;$$
$$Y_5 = Y_p(x_5) + e = \bar{y} + b_5(x_5 - \bar{x}_5) + e.$$

7 Conclusions and further research

Since the regression analysis does not end only with the assessment of the parameters of the model, it is necessary to identify all the most influential observations and assess their negative contribution.

Modern mathematical theory offers methods for identifying such components and assessing their acceptability in the data. These methods are based on the previous *QR* decomposition of the data matrix.

The extra-large amount of computing work is no longer an obstacle in the presence of modern computers.

Comparing the results of calculations of multifactor regression analysis by the method of *QR* decomposition of rectangular matrices by House-Holder mappings with the calculations performed in the *StatGraphics* package, we obtained a 100% match.

Checking the significance of the parameters of the equations by the *student's* criterion shows that not all of them are significant:

$y = 899{,}649 + 7{,}59366 * x_3$;
$x_1 = 26{,}1328 + 0{,}60714 * x_2 - 0{,}305855 * x_3$;
$x_4 = 67{,}0703 - 0{,}906909 * x_2 - 0{,}600053 * x_3 - 1{,}13156 * x_5$.

In the following, we can consider how the procedures for estimating the parameters of the equation can be performed in the *StatGraphics* package.

Summarizing all the above, as a conclusion, we can say that the use of *QR* decomposition matrices has significant advantages over the standard procedure of the least squares' method in the presence of multicollinearity of data and is a reliable computational procedure.

References

[1] Dorokhov, O., Dorokhova, L. (2011). Fuzzy model in fuzzy-tech environment for the evaluation of transportation's quality for cargo enterprises in Ukraine. *Transport and Telecommunication*. 12(1):25–33

[2] Dorokhova, L., Dorokhov, O. (2017). Computer fuzzy model regarding pharmacies integral perceptions by visitors. *Bulletin of the Transilvania University of Brasov, Series III: Mathematics, Informatics, Physics*. 10(2):155–170
https://doi.org/10.31926/but.mif.2019.12.61.2.23

[3] Omelchenko, O., Dorokhov, O., Kolodiziev, O., Dorokhova, L. (2018). Fuzzy modeling of the creditworthiness assessments of bank's potential borrowers in Ukraine. *Ikonomicheski Izsledvania*. 27(4):100–125

[4] Susanto, S., Utama, D. (2022). Fuzzy based decision support model for health insurance claim. *Informatica*. 46:119–130
https://doi.org/10.31449/inf.v46i7.4325

[5] Narang, M., Joshi, M., Pal, A. (2022). A hesitant fuzzy multiplicative base-criterion multi-criteria group decision making method. *Informatica*. 46:235–242
https://doi.org/10.31449/inf.v46i2.3452

[6] Malyarets, L., Kovaleva, K., Lebedeva, I., Misiura, I., Dorokhov, O. (2018). The Heteroskedasticity Test Implementation for Linear Regression Model Using MATLAB. *Informatica*. 42(4):545–553
https://doi.org/10.31449/inf.v42i4.1862

[7] Vigneau, E. (2021). Clustering of variables for enhanced interpretability of predictive models. *Informatica.* 45:507–516
https://doi.org/10.31449/inf.v45i4.3283

[8] Malyaretz, L., Dorokhov, O., Dorokhova, L. (2018). Method of constructing the fuzzy regression model of bank competitiveness. *Journal of Central Banking Theory and Practice*. 7(2):139–164
https://doi.org/10.2478/jcbtp-2018-0016

[9] Ishikawa, A., Fujimoto, S., Mizuno, T. (2021). Why does production function take the Cobb–Douglas form? *Evolutionary and Institutional Economics Review*. 18(1):79-102.
https://doi.org/10.1007/s40844-020-00180-3

[10] Yang, Y., Schmidt, P. (2021). An econometric approach to the estimation of multi-level models. *Journal of Econometrics.* 220(2):532-543.
https://doi.org/10.1016/j.jeconom.2020.04.012

[11] Liu, L., Fan, T. (2021). The Impact of Identity on Satisfaction toward Entrepreneurial Environment: A Case Study. *Discrete Dynamics in Nature and Society.* 2021:1-8.
https://doi.org/10.1155/2021/7234635

[12] Borisov, I. (2021). Agricultural personnel and rural labour: How is their reproduction related? *Journal of New Economy*. 22(3):161-183.
https://doi.org/10.29141/2658-5081-2021-22-3-9

[13] Varbanova, V., Beutels, P. (2020). Recent quantitative research on determinants of health in high income countries: A scoping review. *PLOS ONE*. 15(9):1-21.
https://doi.org/10.1371/journal.pone.0239031

[14] Wang, W., Zhou, Y. (2021). Eigenvector-based sparse canonical correlation analysis: Fast computation for estimation of multiple canonical vectors. *Journal of Multivariate Analysis*. 185(C): 30-41.
https://doi.org/10.1016/j.jmva.2021.104781

[15] Cheng, H. (2021). Importance sampling imputation algorithms in quantile regression with their application in CGSS data. *Mathematics and Computers in Simulation (MATCOM)*. 188(C):498-508.
https://doi.org/10.1016/j.matcom.2021.04.014

[16] Li, T., Song, X., Zhang, Y., Zhu, H., Zhu, Z. (2021). Clusterwise functional linear regression models. *Computational Statistics & Data Analysis*. 158(C): 101-112.
https://doi.org/10.1016/j.csda.2021.107192

[17] Jian, D., Yu, C., Aihua, G., Jingwei, C., Lili, L., Qinling, C., Shujun, L., Qifeng, X. (2020). Potential Trend for Online Shopping Data Based on the Linear Regression and Sentiment Analysis. *Mathematical Problems in Engineering*. 2020:1-11.
https://doi.org/10.1155/2020/4591260

[18] Bagnato, L., Punzo, A. (2021). Unconstrained representation of orthogonal matrices with application to common principal components. *Computational Statistics*. 36(2):1177-1195.
https://doi.org/10.1007/s00180-020-01041-8

[19] Wei, C., Bo, Z. (2021). Controllability of Flow-Conservation Transportation Networks with Fractional-Order Dynamics. *Complexity*. 2021:1-13.
https://doi.org/10.1155/2021/8524984

[20] Wireko, F., Barnes, B., Sebil, C., Ackora-Prah, J. (2021). The Eigenspace Spectral Regularization Method for Solving Discrete Ill-Posed Systems. *Journal of Applied Mathematics*. 2021:1-11.
https://doi.org/10.22541/au.161212952.20018992/v1

[21] Jiayan, W., Lanlan, G., Zongmin, L., Xueqin, W., Zhengqing, F. (2021). An Efficient Algorithm for Ill-Conditioned Separable Nonlinear Least Squares. *Journal of Mathematics.* 2021:1-8.

Smart Curriculum Mapping and Its Role in Outcome-based Education

Tanweer Alam[1], Mohamed Benaida[1]
[1]Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia
E-mail: tanweer03@iu.edu.sa, md.benaida@gmail.com

Keywords: Curriculum mapping, Student learning outcomes, Assessment Methods, Machine Learning Techniques, K-Map Algorithm.

Received: September 1, 2021

Educational development processes are essential for successful academic performance in educational and technical environments. Teachers and students also need a model and guidelines required for effective learning. Without effective curriculum mapping, the institutions cannot accurately estimate outcomes and maximize potential performance on resources. A matrix depicts the relationship between student learning outcomes (SOs) and topics on the curricular map. The need to earn satisfying produce of education and achieve considerable progress in the visibility of education equity in completing professional duties is a primary motivation for learning the curriculum. One of the most effective strategies to increase overall teaching effectiveness, involvement, or curricular interaction is curriculum development. The mapping connects all disciplines to academic outcomes and displays well-planned teaches. An excellent example of a curriculum should be well-prepared and purposefully encourage expertise acquisition. This paper describes a set of range standards and recommendations for this technique and challenges that affect curricular map construction. As a result, this strategy will increase the overall performance of education and the quality of the curriculum.

Povzetek: Prispevek opisuje novo arhitekturo veriženja IBchain, ki uporablja internet stvari in verigo blokov za varno komunikacijo

1 Introduction

Curriculum maps offer data that display how essential curriculum parameters are related and aligned. The mapping with the parameters at the upper left and down the proper columns regularly reflects it. Then, at some stage in the essential cells, association signs signify relation to some of the vital issue additives, using a curriculum mapping model to demonstrate where results are interpreted via courses. Curriculum mapping is a program that has proven to be effective. Within the current phase of innovative system recommendations, it is widely utilized as a monitoring mechanism in better education. It gives program directors and teacher guides the ability to regularly direct their curriculum files and obtain outcomes data [1]. Educational institutions have established curriculum mapping solutions to increase program quality, participation, and collaboration. Such methodologies can effectively assist college students in understanding curriculum results and identifying problems in the functions [2]. Curriculum mapping is a term used to describe the process of controlling syllabus learning outcomes using guidelines to identify or highlight academic discrepancies, inefficiencies, and inconsistencies in conceptual curriculum results [3]. The outcomes include improving the syllabus map's standard alignment and more significant student learning outcomes [4]. Furthermore, the curriculum map style indicates how well the issue educator will illustrate, and the content structure will cover the teaching principles established by professional learning outcomes [5]. As a result, experts have developed an intelligent solution, including a current curriculum map as an addition that allows us to explore road maps, increase efficiency, and offer recommendations for refining the lesson map (Figure 1). Our goal is to create a platform that intelligently carries out the program's instructions and skilled outcomes. Accreditation organizations approve curriculum courses following the plan's educational aims and the position's requirement for a curriculum map to determine the effectiveness of each resource / expert outcome.

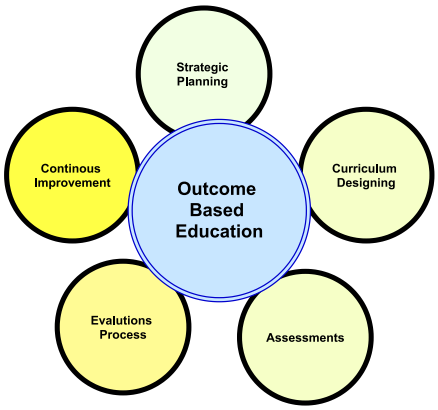


Figure 1: Outcome-based education tools

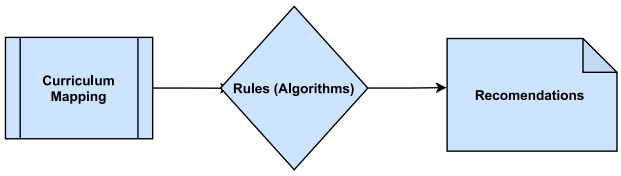


Figure 2: Smart Curriculum Mapping Process

The following research questions are suggested as part of this study:

1. What are the most successful ways for locating a functional map in a curriculum?

2. How will we include such procedures into the intelligent map reaction?

The standard exceeds the following measures, special curriculum map assessment presented.

1. Describe the syllabus's topics and learning outcomes for each problem.

2. To satisfy the program's needs, write measurable learning outcomes for learners. Several platforms, such as ABET, receive findings from accredited authentication organizations.

3. professional program outcomes with a healthy education. This alignment should be as precise as feasible to aid in the full achievement of student learning outcomes.

4. Check the curricular map using a set of expertise rules. The initial level of a curriculum map is evaluated using this set of guidelines.

5. Instructions for creating and revising curricular maps to improve fundamental learning skills. A few tips, for example, are unrelated to a broad publishing style relating to software outcomes or additional learning after appearance.

The curriculum mapping evaluation is completed in the following stages.

Stage 1. Self-study strategic planning for each course in a program. Figure 2 shows the smart curriculum mapping process.

Stage 2. Writing good student learning outcomes for fulfilling all requirements of a program.

Stage 3. Align the courses with standard students learning outcomes.

Stage 4. Apply the intelligent algorithm to evaluate the curriculum mapping.

Stage 5. Find all recommendations from the algorithm and rebuild the curriculum mapping to improve the quality of learning.

This material is divided into the sections below. The second section depicts activities connected to curricular mapping; the third section details the methods and rules for refining the map; the fourth section describes the proposed algorithm. The fifth section displays and discusses the findings of the issues. Finally, section 6 brings the work to a close with recommendations for further research.

2 Related works

Plaza et al. (2007) published a paper on curriculum mapping and application evaluation. The authors have combined work with students and teachers with a picture relevant to the curriculum mapping in this research [6], which shows the correlation between motivated/enhanced and acquired ideas.

Uchiyama and Radin (2009) published articles on curriculum mapping for better learning. They represent a curriculum map that can develop recognition based on real-time data to improve school education quality [7].

Perlin (2011) stared at the curricular map for this evaluation program, solely centred on using the George Mason University Fitness Management Method. It aided in developing a framework for integrating high-level analysis with a mapping exercise. This framework begins with acquiring analytical and technical methods to study the university's mapping approach and applications.

As a result, interruptions have emerged. It has been discovered that they can be adaptable in the end, and users can grasp how to overcome those challenging conditions, so learners of this mapping gadget should develop the first college software [8].

Spencer et al. (2012) provided a curricular map related to the possibility of adjusting. The collecting, assessment, and innovative exercise and evaluation skills are discussed. This method of questioning encourages the growth of physical exercise collaboratively. Incremental graphs provide conceptual presentations of essential techniques and requirements [9].

Lam and Tsui have presented a research article on course mapping experiments. (2013). It suggests that creating a curriculum map could be a good technique for determining how much agreement university students have in producing results that fit the expectations. It means that developing a curriculum map could be a valuable tool for assessing the level of agreement among university students in creating outcomes that correspond to the university organization's learning [10]. Although the use of mathematics in a better university is promising, Avela et al. (2016) believe that the best errors are related to the sharper development of examination statistics in the concern that consumers would be unable to use this system successfully or with its measureless interval. As a result, academics have looked into scientific intelligence-sharing methodologies for visual data analysis, relationship mining, and other issues. As a result, the advantages and problems of acquiring scientific knowledge have been identified and recognized for better education, allowing educators to use this powerful and effective tool to enhance the happiness of learning in a better education [11].

Pat Hutchings (2016) presented a paper on the focus of education and the student's outcomes. It focuses on identifying and spreading approaches for applications and organizations to effectively use assessment data to notify and improve courses and convey satisfactorily with candidates [12].

A research article on advanced program progress processes was published by M Jacobsen et al. (2018).

	Basic Courses	Compulsory Courses	Compulsory Courses	Compulsory Courses	Compulsory Courses	----	Capstone Course
SO-1	I	R	R		R	--	E
SO-2	I			R	R	--	E
SO-3	I		R	R		--	E
SO-4	I			R	R	--	E
SO-5	I	R			R	--	E
SO-6	I			R	R	--	E
-------	I	R	R	R		--	E

Table 1: Curriculum Matrix (Model)

They have presented the continuous development methods based on flexibility and assistance, occurring in a program led by an annual organization headed by organized platforms [13].

The usage of the curriculum map in the modern science school has been examined, according to Gaith et al. 2018, to provide greater awareness of the program in various health institutions and surrounding academic contexts. It has evolved from a participatory development initiative and specific information about the path's flexibility [14] to a participatory development initiative and a simple idea for flexibility. M. Jacobsen et al. (2018) discovered protective features that contribute to particular system development and characteristics that contribute to effective and efficient academic communication, offering excessive professional support within graduation application. This observation has aided in determining the program's strength for graduate students and how it can appropriately make mistakes [15]. Figure 3 shows the assessment cycle.

Buker and Niklason improved the map version in 2019. They seek to establish fundamental guidelines for curriculum development. Assessing the program's cause and analyzing the findings of compliance-based research are two recommended ways [16]. Treadwell et al. (2019) this study looks at how curriculum mapping is presented in interactive web-based learning opportunities, objectives, and outcomes in consultation with 40 instructors to see how they like it. Although participants did not immediately understand how to use the system and experienced some developmental issues, the results consistently showed that they believed the learning outcomes with university students with many inclusive motivations such as communication, usability, and transparency [17]. The Lacerda and Sepel (2019) study examine educators' views to gather their experiences and generate curricular transformation proposals. Their findings revealed that publishing-interest had become more widely accepted than critical theories and that the outcomes with traditional notions were not sufficiently good. Excellent exercise planning was directly related to good exercise planning [18]. Lindén et al. (2017) discuss the impact of the teaching concept on the development of more excellent knowledge while focusing on its historical and psychological foundations. The findings demonstrate that it adjusts to higher education emergencies. It is necessary to grasp the distinctions between the entire curriculum and the critical points simultaneously and consider the curriculum's influence [19]. There have been significant changes in curriculum thinking over the last decade. The number of troubling issues covered has substantially grown in developing solutions and concepts.

Their findings show that academic research is becoming more comprehensive and prone to educational improvements [20]. They matched their curriculum to those of the United Kingdom and Latvia, simultaneously focusing on historical and theoretical subjects. A physics curriculum was gained by Yates and Millar (2016), notably at universities with Australian inner skills.

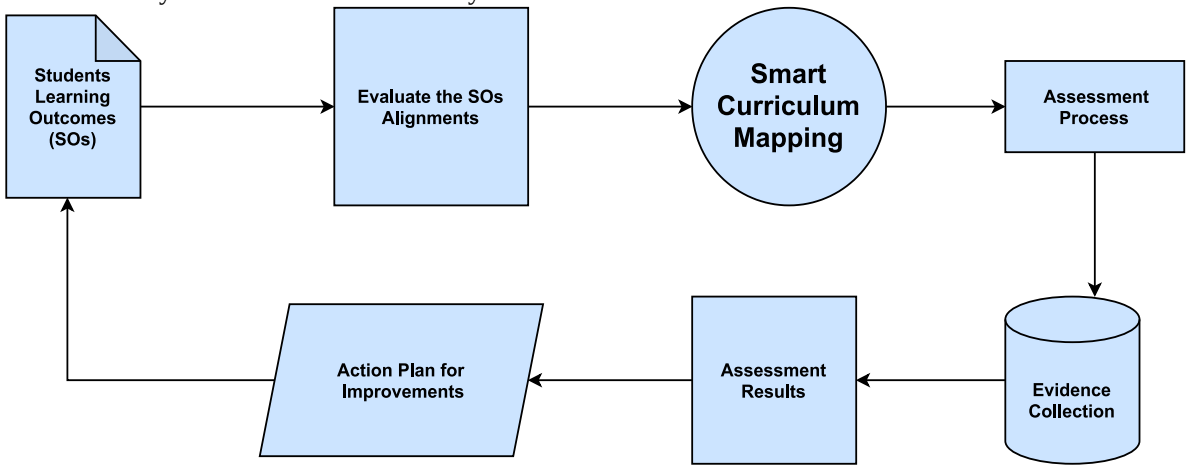


Figure 3: Assessment cycle

Physics has evolved into a unique matter, owing to the vast technical knowledge that must be updated in trading and reviewed over time. Its popularity is based on whether the curriculum can be taken seriously in the workplace [21]. Sweden's new curriculum, which was implemented in 2011, was examined in such a way (Alvunger 2018). According to their findings, the instructor's curriculum establishes three roles inside the class: the interaction area, the arrangement, and the person's view. The answers of academics have provided persuasive evidence that curriculum modifications necessitate a high degree of material content and that lecturers can assess skills without difficulty because they will desire to connect with university college students at the same time [22]. Ghaderia (2011) represents curriculum principles focused on total peace and considers peace necessary for conservation maintenance as technology wars evolve. They believe that combining modern ideas with freedom ideas (that is, by combining similarities and differences) can create a final curriculum that promotes peace [23, 24, 25].

3 Methodology

The suggested algorithm analyses the curriculum map based on the students' learning outcomes. In addition to the existing scientific procedures used in curriculum map planning and evaluation, carefully examines every subject-related lesson and curriculum map development to establish the primary method utilized to attain a satisfactory level.

To fulfil the visions, we took a four-step approach:

1. Create curriculum maps and their setups.
2. Guidance on measuring a high-quality course map using an algorithm.
3. Plan a set of suggested guidelines and review them.
4. Research-based expert assessment of outcomes

The following terms I, R and E are employed within the program to connect the curriculum and know the program's outcomes, depending on the model review and the needs for numerous educational accrediting structures (e.g. ABET, NCAAA, etc.).

Introduced (I): students are not expected to become curriculum experts. Students learn to distinguish subjects using basic information, understanding, skills, and talents (first and second year).

Reinforcement(R): students are expected to have a basic understanding and knowledge of the topic or skills. Learning activities help students develop and combine their awareness, skills, and growth challenges (guidelines).

Expert (E): University students are expected to have a strong foundation and understand basic, technical, or fundamental skills. Utilizing information or abilities in various circumstances and varying difficulty levels emphasizes education and knowledge (capstone).

Figure 4 depicts a sample curriculum map. Many of the program findings are no longer matched with themes that are consistent with the superb analysis map's general policies.

The following method must be utilized to generate a wonderful map that provides great satisfaction in knowing the outcomes.

1. Understanding the outcomes: The outcomes should be explained in conjunction with the program and problem headings. The results must be compatible with the program's aims, objectives, and goals.

2. Mapping: Understanding system results should go hand in hand with knowing the consequences. The required subjects should be considered to ensure that students' knowledge is collected and sequenced.

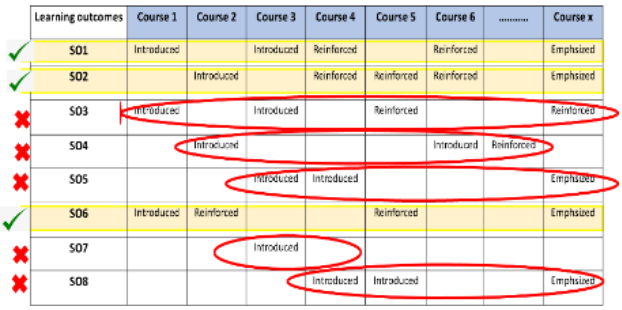


Figure 4: Sample of Curriculum Mapping

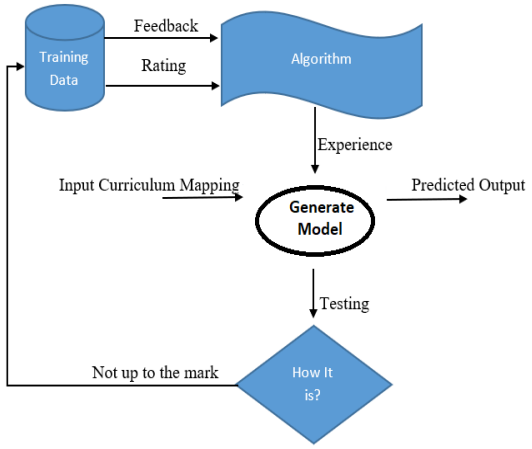


Figure 5: Evaluation of curriculum mapping

Alignment strategies: A set of actions must be used to develop a curriculum map. For instance, a high level of alignment (I, R, E), a list of courses specific to each program, and so on.

Participants should be included in the development of map information acquisition, including professors, and students.

Table 1 displays an example of a curriculum matrix that depicts the relationship between the system's numerous learning outcomes and the curriculum courses.

The following strategies should be employed to earn the first report of any provided curriculum map, based on quality assurance information and permitted enterprises [26, 27, 28].

1. At least three studies must agree with all test outcomes (I, R and E).

2. All lessons should no longer be challenging in all student's outcomes. However, each view must keep at least 3-5 outcomes.

3. The study's findings must be interpreted logically for learners to proceed from the early to the later stages of the period.

4. "I" must be the initial stage, and all of the first lessons must be covered. The "R" should begin at a third of a turn to the stop and cover all crucial and necessary issues. Also, "E" aims to incorporate expertise (capstone).

The ideas listed above can be used to evaluate a curricular map that follows a set of rules. Within the suggested implementation tool, the following sections are used [29, 30, 31, 32].

Phase 1: Discover study and assessment data that match the model and values for analysis procedures. All teachers must be assessed the related information.

Phase 2: Practice learning and data collection to enhance statistical abilities.

Phase 3: Request feedback and a report from the university.

Train the machine with this data, and it will create comments and recommendations with standard rules.

Phase 4: Locate the essential first steps using the curriculum map. Phase four saw an improvement in implementation.

Phase 5: Create a curriculum map model using all factors and clarify it according to the educational outcomes.

After some experience, a set of rules will yield a blueprint for creating a workable curricular map.

Phase 6: Create a curriculum map and determine the expected outcome.

A collection of rules will take your input in the form of a curriculum map and use it to generate output.

Input, processing, and output are suggested in figure 5. The flow of a set of curricular map rules is improved by figure 6.

The above levels will be utilized to assess curriculum map development using digital learning methodologies as described in the recommendations.

The following is a summary of the proposed set of rules:

(1) put in (curriculum map)

(2) Algorithm for mapping curriculum

(3) Output (recommendations)

4 Procedure for generating smart curriculum mapping

Curriculum Mapping test method

i) Input- Curriculum map design: says CM

ii) Return information for each column to the CM. According to the data, we define it

$D_{ij} = (C_j L_k \sum_i SO_i)$ where i=1,2,3 ….., j=1,2,3…,

where SO is the outcome of students, L is the grade level, and C is the curriculum's courses.

iii) In step 1, we look for special character sets (D_{ij}). For instance: L_1C_1{I R E R},
L_2C_2{I I I I E I},
L_3C_3{I R E R}, and so on.

Each set should be divided into three characters: L, C and {}. Divide each set into three arrays of characters say-Level[i], Courses[j], str[j].

iv) in this step,

a) if i=1 or 2, then go to step 5

b) else if i=3 or 4 or 5, then go to step 6

c) else if i=6 or 7 or 8, then go to step 7

d) else go to Step 8.

v) locate an array of courses[j] and corresponding string str[j]

a) The event that string str[j] has all blank, then Print Error "Course [j] has not aligned with any SO".

b) Differently, if string str[j] has no blank, then Print Error "Course [j] has aligned with all SO".

c) Differently, if str[j] has L and blank mixed characters, then Print "Course [j] has aligned with SOs successfully".

d) Else go to step 4

vi) locate an array of courses[j] and corresponding string str[j]

a) If the str [j] series is empty, then the "course [j] does not match any SO" error message will be displayed.

b) Otherwise, print error "course [j] aligned with all SO" if string str [j] is empty.

c) Print "course [j] aligned with SOs correctly" if str [j] contains M and pure mixed letters.

d) Alternately, proceed to step 4.

vii) Locate a series of course [j] and a SO [j] that corresponds to them.

Print errors "Course [j] is not compatible with any SO" if the string str [j] is empty.

Otherwise, if string str [j] is empty, the error "Course [j] aligned with all SO" will be printed. If str [j] does not contain H or mixed blank characters, print "line [j] aligned with SO nicely."

Alternately, proceed to step 4.

viii) identify all sets of characters from $\sum_i SO_j$

where i = 1, 2, the number of guides and j = 1, 2, number of SOs Keep in mind the vast range of spaces in each set and follow the steps below:

If the total number of empty spaces in any SO set equals the total number of courses, print error "all courses must follow with the rules."

Else Print errors occur when the maximum spacing width is less than 3 or more than 5 lessons for the total number of recommendations. "At least three outcomes must be aligned in all courses (including I, R, and E)."

Print- "Mapping is the best".

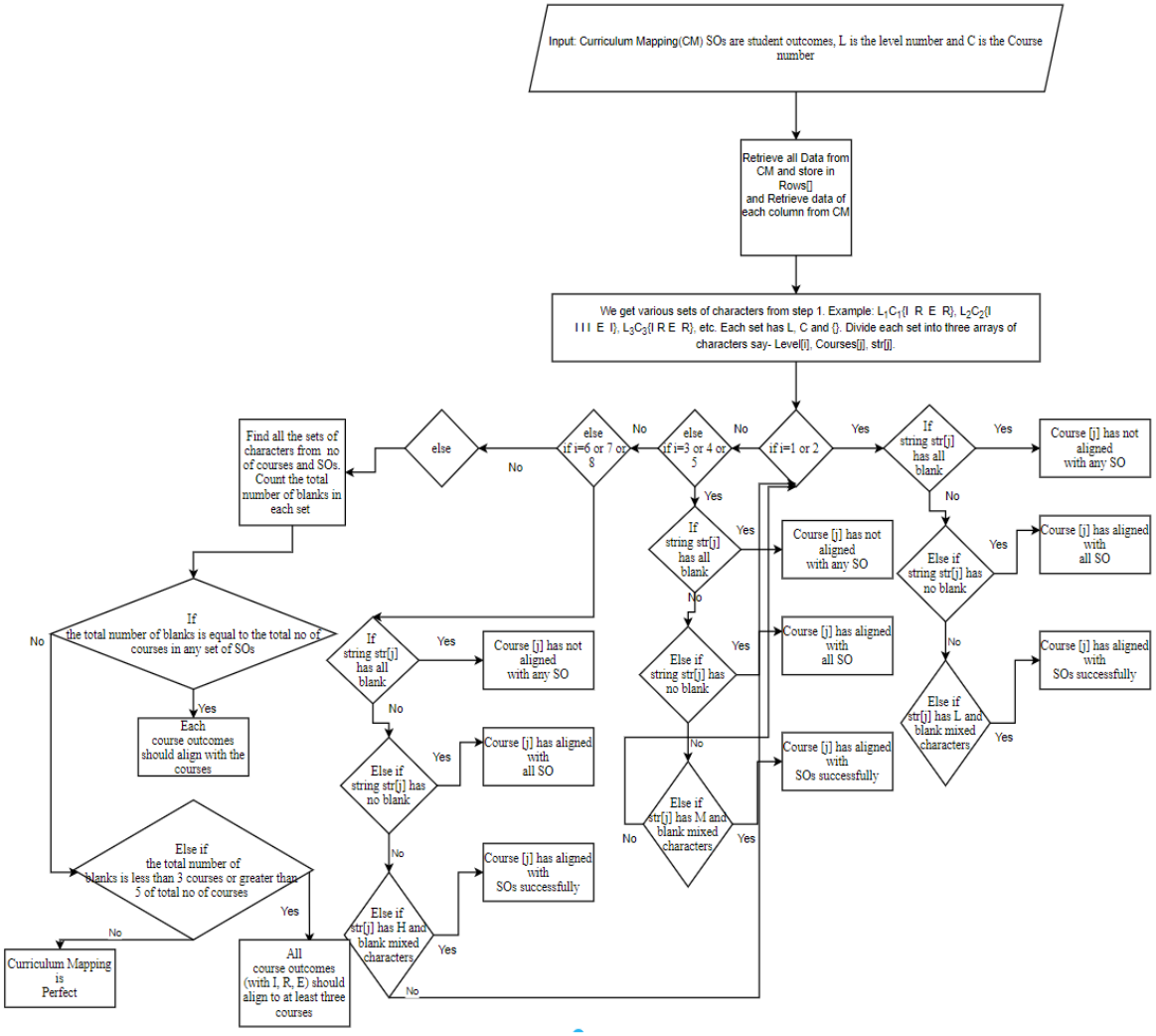


Figure 6: Flow Chart of Curriculum Mapping evaluation

Recommendations

1. Course # has not aligned with any SO#.
2. Course # has aligned with all SOs.
3. Course # has not aligned with any SO.
4. Every course outcomes should align with the courses.
5. Every course outcomes should align to at least three courses.
6. Curriculum Mapping is Perfect.
7. Percentage of the perfectness of the curriculum mapping.

Figure 7: Recommendations (Output)

This algorithm generates curriculum map recommendations. Figure 7 shows a list of recommendations. Suggestions can be "Study number does not correspond to any SO", "Course number corresponds to SO successfully", "Study number corresponds to all SO", "The result of each study must correspond to the course.", "At least three subjects must accompany all course outcomes (I, R, E)".

5 Results and discussion

Curriculum mapping is usable thru several sources, along with newcomers' students, instructors, universities, and academic groups. Commonly every person in the class is meant to be privy to how the program influences them presently and in the coming years. This research explains the evaluation of curriculum mapping and structures, structures, and guidelines [33, 34, 35]. Developing a conceptual image of an advanced mapping processing method is a tremendous advancement. The mapping is designed for each consultation and needs to be done by every teacher when the bankruptcy content material is changed into created. Further, the researchers additionally protected the views of curriculum mapping specialists, educational professionals, and higher training specialists. There is a diffusion of techniques for collecting statistics, but authors have chosen to apply surveys because those are enormously popular. This collects vital information concerning respondents' observations in each program and assesses their remarks [36, 37, 38, 39]. Forty experts participated in this evaluation, every of whom completed a ten questions survey designed to decide the experts' professions. The following nine questions are regarding curriculum mapping, and they're graded as proven in very satisfied, satisfied, natural, dissatisfied, and eventually very dissatisfied. Curriculum specialists contribute 25% of

people who spoke back to the survey, and educational experts (12.5%) and senior faculty individuals account for the relaxation (37.5%). Absolutely everyone's valued understanding and know-how have enabled them to study how curriculum mapping has been investigated, with more than half mentioning that they're very thrilled or satisfied with what they were provided (12.5% and 12.5%, respectively). The satisfaction level of the question "Do you find the smart curriculum mapping given to you to be useful?" is 70%. The satisfaction level of the question "Would you believe curriculum mapping can aid in the achievement of learning objectives?" is 65%. The satisfaction level of the question "Would you believe smart curriculum mapping can determine curriculum mappings' weak points?" is 88%". The satisfaction level of the question "Would you feel curriculum mapping should be able to make the curriculum more accessible and show the links between the subjects and the learning outcomes of the students?" is 70%. The satisfaction level of the question "Would you accept Smart Curriculum Mapping can assist in suggestions for curriculum design developments?" is 88%. The satisfaction level of the question "Would you believe the Smart Curriculum Mapping would be transparent?" is 85%. The satisfaction level of the question "Would you believe Smart Curriculum Mapping would equally align all student learning outcomes?" is 88%. The satisfaction level of the question "Would you believe the Smart curriculum mapping offered the appropriate balance between theory and practice?" is 85%.

Finally, "the rate of how satisfied you are with the smart curriculum mapping." is 88 %.

The given desk deals with a list of questions in a tabular shape; although you would no longer use a score and instead advocate a consultant place of knowledge, query 1 is isolated from the other questions (Table 2). Table 3 lists each professional's solutions to each question; such solutions are accumulated, and every question's suggestion is determined. Curriculum map visualization has many benefits. A listing of advantages is given below.

1) continuity of concern counts and gaining knowledge of dreams.

2) ongoing, regular, minimum, and repeated development initiatives have all been powerful.

3) aid in furthering one's research (consensus or inclusion).

4) observe and examine the learning effects.

5) enhance the academic body of workers' excellence (capable of percentage the getting to know the process).

6) talk about clarity problems (precise information, plan opinions, academic support, and higher application results).

7) method improvement requirements. Each path might need to broaden a method to evaluate scholar success, thinking of the program's curriculum.

Questions	10 Expert	5 Expert	15 Expert	5 Expert	5 Expert
Kindly select the option that perfectly represents yourself from the list below.	Educational Specialist	Curriculum Advisor	Professor	Curriculum Professional	Academic Expert

Table 2. Question 1 presenting the specialists' area of professionalism

Questions	Expert	Expert	Expert	Expert	Expert	Expert	Expert	Expert	Expert	Expert	otal (max 200)	Percentage(%)
Do you find the smart curriculum mapping given to you to be useful?	16	20	20	20	12	8	8	8	8	12	132	66
Would you believe curriculum mapping can aid in the achievement of learning objectives?	16	20	20	20	12	8	8	8	8	16	136	68

Would you believe smart curriculum mapping can determine curriculum mappings' weak points?	16	20	20	20	12	16	16	16	16	16	168	84
Would you feel curriculum mapping should be able to make the curriculum more accessible and show the links between the subjects and the learning outcomes of the students?	20	20	20	20	12	8	8	8	8	12	136	68
Would you accept Smart Curriculum Mapping can assist in suggestions for curriculum design developments?	16	20	20	20	12	16	16	16	16	16	168	84
Would you believe the Smart Curriculum Mapping would be transparent?	16	20	20	20	12	16	16	16	16	12	164	82
Would you believe Smart Curriculum Mapping would equally align all student learning outcomes?	16	20	20	20	12	16	16	16	16	16	168	84
Would you believe the Smart curriculum mapping offered the appropriate balance between theory and practice?	16	16	20	20	12	16	16	16	16	12	160	80
Also rate how satisfied you are with the smart curriculum mapping.	16	20	20	20	12	16	16	16	16	16	168	84

Table 3. Questionnaire results (Q2-10) with averages

	Introduced	Reinforced	Emphasized
Corrective Expertise	4	5	3
Essential Imagining	7	2	1
Interaction	2	3	9
Study Abilities	3	2	2
Moral Analysis	3	1	1

Table 4. Statistics of learning outcomes vs strength of curricula

The comprehensive evaluation software includes getting to know results, the manner used to evaluate each outcome, every strategic mark, the individual responsible for collecting records, the accrued information, the responsibility for decoding outcomes and acquiring corrections, and evaluating enhancements (figure 6, table 4).

We expect that the getting to know outcomes of the 14 students are divided into discipline understanding, vital questioning, verbal exchange, research abilities, and a behavioural session on the desired topics. In keeping with the effects of this take a look at, most professionals are satisfied with the proposed approach. Desk 3 shows the outcomes of the expert evaluations. Curriculum evaluations can be trendy, but the wide variety of variables, including entering records and available assets, relies upon whether the entire curriculum is classed. Establishments should be privy to the need for high stage checking out and strategies for undertaking all assessments. On every occasion, the consequences of the instructions are explained, they allow the mapping of their classes.

Figure 7: Student's learning outcomes for mandatory courses

6 Conclusion

In the meantime, several universities employ curriculum map modification tools to analyze and improve curriculum consistency and give great development accuracy in higher education. They developed and integrated a good (developmental, organized, and meaningful) learning environment. Once SO has agreed to the design, ensure that activities are grouped so that the most common learning technique can learn. Building a curriculum map helps demonstrate the relationship between instruction and student outcomes understanding and provides a more thorough curriculum overview. This study has produced, examined, and algorithmically suggested an outstanding curricular map. The syllabus set of curriculum maps analyses the curriculum and provides recommendations for the map's correctness based on all of the principles used to assess the syllabus's suitable quality. In conclusion, this algorithm could be developed in higher education institutions to boost the acquisition of specialized knowledge even more.

References

[1] Bone, E. K., & Ross, P. M. (2021). Rational curriculum processes: revising learning outcomes is essential yet insufficient for a twenty-first century science curriculum. Studies in Higher Education, 46(2), 394-405. https://doi.org/10.1080/03075079.2019.1637845

[2] Lafave, L. M., Yeo, M., & Lafave, M. R. (2021). Concept mapping toward competency: Teaching and assessing undergraduate evidence-informed practice. The Journal of Competency-Based Education, e1242. https://doi.org/10.1002/cbe2.1242

[3] Webb, K. K. (2020). Curriculum mapping in academic libraries revisited: Taking an evidence-based approach. College & Research Libraries News, 81(1), 30. https://doi.org/10.5860/crln.81.1.30

[4] Joyner, H. S. (2016). Curriculum mapping: A method to assess and refine undergraduate degree programs. Journal of Food Science Education, 15(3), 83-100. https://doi.org/10.1111/1541-4329.12086

[5] Herrmann, T., & Leggett, T. (2019). Curriculum Mapping: Aligning Content and Design. Radiologic technology, 90(5), 530-533.

[6] Plaza, C. M., Draugalis, J. R., Slack, M. K., Skrepnek, G. H., & Sauer, K. A. (2007). Curriculum mapping in program assessment and evaluation. American Journal of Pharmaceutical Education, 71(2). https://doi.org/10.5688/aj710220

[7] Uchiyama, K. P., & Radin, J. L. (2009). Curriculum mapping in higher education: A vehicle for collaboration. Innovative Higher Education, 33(4), 271-280. https://doi.org/10.1007/s10755-008-9078-8

[8] Perlin, M. S. (2011). Curriculum mapping for program evaluation and CAHME accreditation. Journal of Health Administration Education, 28(1), 33-53.

[9] Spencer, D., Riddle, M., & Knewstubb, B. (2012). Curriculum mapping to embed graduate capabilities. Higher Education Research & Development, 31(2), 217-231. https://doi.org/10.1080/07294360.2011.554387

[10] Lam, B. H., & Tsui, K. T. (2013). Examining the alignment of subject learning outcomes and course curricula through curriculum mapping. Australian Journal of Teacher Education, 38(12), 6. https://doi.org/10.14221/ajte.2013v38n12.8

[11] Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. Online Learning, 20(2), 13-29. https://doi.org/10.24059/olj.v20i2.790

[12] Hutchings, Pat. (2016). "Aligning educational outcomes and practices." https://files.eric.ed.gov/fulltext/ED567005.pdf

[13] Jacobsen, M., Eaton, S. E., Brown, B., Simmons, M., & McDermott, M. (2018). Action research for graduate program improvements: A response to curriculum mapping and review. Canadian Journal of Higher Education/Revue canadienne d'enseignement supérieur, 48(1), 82-98. https://doi.org/10.47678/cjhe.v48i1.188048

[14] Al-Eyd, G., Achike, F., Agarwal, M., Atamna, H., Atapattu, D. N., Castro, L., ... & Nausheen, F. (2018). Curriculum mapping as a tool to facilitate curriculum development: a new School of Medicine experience. BMC medical education, 18(1), 1-8. https://doi.org/10.1186/s12909-018-1289-9

[15] Jacobsen, M., McDermott, M., Brown, B., Eaton, S. E., & Simmons, M. (2018). Graduate students' research-based learning experiences in an online Master of Education program. https://doi.org/10.53761/1.15.4.4

[16] Buker, M., & Niklason, G. (2019). Curriculum Evaluation & Improvement Model. The Journal of Health Administration Education, 36(1), 37.

[17] Treadwell, I., Ahlers, O., & Botha, G. C. (2019). Initiating curriculum mapping on the web-based, interactive learning opportunities, objectives and outcome platform (LOOOP). African Journal of Health Professions Education, 11(1), 27-31. https://doi.org/10.7196/AJHPE.2019.v11i1.1073

[18] Lacerda, C. C., & Sepel, L. M. N. (2019). Basic school teachers' perceptions about curriculum theories. Educação e Pesquisa, 45. https://doi.org/10.1590/s1678-4634201945197016

[19] Lindén, J., Annala, J., & Coate, K. (2017). The role of curriculum theory in contemporary higher education research and practice. In Theory and method in higher education research. Emerald Publishing Limited. https://doi.org/10.1108/S2056-375220170000003008

[20] Rouk, V. (2013). From Times of Transition to Adaptation: Background and Theoretical Approach to the Curriculum Reform in Estonia 1987-1996. Bulgarian Comparative Education Society.

[21] Yates, L., & Millar, V. (2016). 'Powerful knowledge'curriculum theories and the case of physics. The Curriculum Journal, 27(3), 298-312. https://doi.org/10.1080/09585176.2016.1174141

[22] Alvunger, D. (2018). Teachers' curriculum agency in teaching a standards-based curriculum. The Curriculum Journal, 29(4), 479-498. https://doi.org/10.1080/09585176.2018.1486721

[23] Ghaderia, M. (2011). Peace-based curriculum based on the theories of "difference" and "similarity". Procedia-Social and Behavioral Sciences, 15, 3430-3440. https://doi.org/10.1016/j.sbspro.2011.04.314

[24] https://cte.tamu.edu/getattachment/Faculty-Teaching-Resource/Program-ReDesign/Curriculum-Mapping/Curriculum-Process-Overview-and-Instructions-SS-DF-(1).pdf.aspx?lang=en-US accessed on: 21-12-2021

[25] Assessment and Curriculum Support Center, University of Hawai‘i at Mānoa (2020). http://manoa.hawaii.edu/assessment/howto/mapping.htm accessed on: 21-12-2021

[26] Curriculum Design, University of South Florida. (2020). https://www.usf.edu/atle/teaching/curriculum-design.aspx accessed on: 21-12-2021

[27] Codó, E., Dans, L., & Wei, M. M. (2008). Interviews and questionnaires. The Blackwell guide to research methods in bilingualism and multilingualism, 158-176. https://doi.org/10.1002/9781444301120.ch9

[28] Patten, M. L. (2016). Questionnaire research: A practical guide. Routledge. https://doi.org/10.4324/9781315265858

[29] Champlain College, Curriculum mapping. (2020). Assessing Learning Outcomes. https://champlain.instructure.com/courses/200147/pages/curriculum-mapping accessed on: 21-12-2021

[30] Abdullah Alshanqiti, Tanweer Alam, Mohamed Benaida, Abdallah Namoun and Ahmad Taleb, (2020) "A Rule-based Approach toward Automating the Assessments of Academic Curriculum Mapping" International Journal of Advanced Computer Science and Applications (IJACSA), 11(12). http://dx.doi.org/10.14569/IJACSA.2020.0111285

[31] Melody, K., Quinn, D.H., Waite, L.H., Mandos, L.A. and Tietze, K.J., 2021. Curriculum mapping: A process to revise the path to achieving student competency. *Education in the Health Professions*, *4*(1), p.1. https://doi.org/10.4103/ehp.ehp_41_20

[32] Cooper, B., Cowie, B., & Furness, J. (2021). Curriculum mapping as a boundary encounter: meeting the demands of multiple agendas. Educational Research for Policy and Practice, 1-22. https://doi.org/10.1007/s10671-021-09299-5

[33] Willans, F. (2021). Mapping the curriculum or doing curriculum mapping? A view from Linguistics and Language programmes. Directions: Journal of Educational Studies, 35(1), 29-37. http://repository.usp.ac.fj/id/eprint/12931

[34] Harden, R. M. (2001). Curriculum mapping: a tool for transparent and authentic teaching and learning. AMEE. https://doi.org/10.1080/01421590120036547

[35] Aljohani, M., & Alam, T. (2015, December). Design an M-learning framework for smart learning in ad hoc network of Android devices. In 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-5). IEEE. https://doi.org/10.1109/ICCIC.2015.7435817

[36] Alam, T., & Aljohani, M. (2020). M-Learning: Positioning the Academics to the Smart devices in the Connected Future. JOIV: International Journal on Informatics Visualization, 4(2), 76-79. http://dx.doi.org/10.30630/joiv.4.2.347

[37] Alepis, E., & Troussas, C. (2017). M-learning programming platform: Evaluation in elementary schools. Informatica, 41(4). https://www.informatica.si/index.php/informatica/article/view/1496

[38] Vičič, J., & Šukljan, T. (2016). Motivating cultural heritage artifacts presentation using persuasive technology. Informatica, 40(4). https://www.informatica.si/index.php/informatica/article/view/1466

[39] Zhu, F. (2018). Research on Intelligent English Oral Training System in Mobile Network. Informatica, 42(2). https://www.informatica.si/index.php/informatica/article/view/2216

A Novel Method for Multiple Object Detection on Road Using Improved YOLOv2 Model

P.Gunasekaran[1], A.Azhagu Jaisudhan Pazhani[2] and T.Ajith Bosco Raj[3]
[1, 2]Department of ECE, Ramco Institute of Technology, Rajapalayam, Tamilnadu, India.
[3]Department of ECE, PSN College of Enginnering and Technology, Tirunelveli, India
E-mail: mailtogunasekar@gmail.com, alagujaisudhan@gmail.com, ajithboscoraj@gmail.com

Keywords: AI, Object, Vehicle, Convolutional, VOC, COCO, KITTI, YOLO

Received: December 30, 2021

Object detection is a branch of machine vision and image processing that deals with instances of a certain class of semantic items. One of the most significant habits of object detection in intelligent transportation schemes is vehicle detection. Its aim is to extract clear-cut vehicle-type information from photographs or videos of automobiles. A fully convolutional network (FCN) is employed in sophisticated driver assistance systems for high performance and quick object identification (ADAS). A novel vehicle detection model employing YOLOv2 is presented to tackle the difficulties of prevailing vehicle detection, such as the absence of vehicle-type recognition, stumpy detection accuracy and sluggish speed. The detection model is trained using the VOC and COCO datasets, and the detection enactment is evaluated quantitatively using KITTI training pictures. In addition, the performance of the YOLOv2 model was compared to that of prior models.

Povzetek: Razvita je nova metoda zaznavanja več objektov na cesti s pomočjo YOLOv2 modela.

1 Introduction

Moving object detection is a computer technique that compacts with recognizing occurrences of semantic matters of a precise class (such as humans, automobiles, etc.) in a digital picture or video. It is connected to computer vision, image processing, and neural networks. Vehicle detection and pedestrian detection are two well-studied fields. In the field of machine vision, moving object detection has a variety of applications, including picture retrieval and video monitoring.

While new research datasets have increased the number of training sets and testing instances to get closer to real-world situations, detectors' capacity to process big data sets in an acceptable period of time has become a significant concern in addition to accuracy. It is not just the number of classes that matters, but also the training examples.

Detecting moving items in a video clip entail finding them in the frame. Item detection is required by every tracking technique, whichever in all frame or when the object first shows in the video. Various backdrop removal approaches from the literature were simulated for moving object detection. Background subtraction uses the relative difference between the current image and the reference updated backdrop over time. Background subtraction that works well should be able to deal with fluctuating lighting conditions, background clutter, shadows, camouflage, bootstrapping, and foreground segmentation in real time.

The tracking of moving objects in video images has flickered a lot of interest in machine vision. Surveillance systems, navigation systems, and object identification all flinch with object tracking. Object tracking is extremely important in a real-time environment because it allows for an improved sense of refuge through visual information, security and surveillance to recognize people, analysis of customer shopping behavior in retail spaces, video abstraction to attain involuntary annotation of videos, generation of object-based synopses, traffic management to examine flow, and design futuristic video effects.

Huieun Kim et al. offered "On-road object identification using Deep Neural Network" [4], which advocated SSD as a quicker object detection method than R-CNN by 41 frames per second. The model is built on SSD and tweaked with the KITTI dataset, which is made up of on-road environment object classes (SSD is a pre-trained model by Pascal VOC pictures). This work proposes an on-road object identification method based on SSD that overcomes the difficulties of detecting on-road objects using a camera in instantaneous and allows for robust object detection. It creates appearance characteristics from input pictures using convolutional layers and trains object position in 2D image coordinates by calculating loss of object box position (IoU) during the training step. SSD, on the other hand, has the disadvantage of overlooking tiny things due to its grid methodology.

The furthermost representative FCN-based object identification approaches are region-based fully convolutional networks (R-FCN), single shot multi-box detector (SSD), and you only look once (YOLO). To obtain good detection performance, these approaches

need a large amount of labeled training data. Most deep learning-based detection algorithms train the classification model using millions of ImageNet classification datasets and fine-tune it by detection training data such as tens of thousands of PASCAL VOC and COCO datasets [13]. The detection approaches based on deep learning, on the other hand, need a high level of computing complication to train the detection models.

A FCN-based object identification approach that enhances performance in a road environment was suggested in the publication "High Performance and Fast Object Detection in Road Environments" [9]. Although the SSD input network is slighter than that of YOLO, the processing time is significantly longer. The classification-specific layer design and the amount of default boxes account for the performance disparity. The VGG-16 model castoff in SSD requires around four times more processing resources than the Darknet-19 model used in YOLO in the classification-specific layer.

Hui-Lee Ooi et al. used an object detector to evaluate the MOT in urban traffic sceneries with road users of various vehicle sizes, whereas earlier work in this area has used background removal or optical flow to excerpt the items of interest regardless of size. The work involves a review of a common model object detector for tracking in urban traffic divisions, as well as the addition of label information to describe the items in the scene. The label information should be a valuable signal to differentiate and associate the objects of interest through frames, resulting in a more precise trajectory, due to the diversity of objects prevalent in urban landscapes. This is documented in the paper “Multiple Object Tracking in Urban Traffic Scenes with a Multiclass Object Detector” [5].

Due to its efficiency and accuracy, a deep-learning object identification model from the Region-based Fully Convolutional Network (R-FCN) framework is used to recognize the road users in each frame. The top performing technique on the MIO-TCD localization contest led to the selection of this detector. The pre-trained model is refined further by using the MIO-TCD dataset to deliver labels for the various road users seen in traffic scenes, each of which falls into one of eleven classes or labels.

The work "Survey of Pedestrian Detection for Advanced Driver Assistance Systems" [3] focuses on one form of ADAS in particular, pedestrian protection systems (PPSs). This study focuses on pedestrians since, according to accident data, 70% of persons engaged in car-to-pedestrian collisions were in front of the vehicle, with 90% of them moving. As a result, PPSs frequently employ forward-facing sensors. The foreground segmentation algorithm detects moving people. The INRIA Person Data set, which is now fairly popular for general human categorization assessment but comprises a significant number of samples derived from high resolution pictures, was employed in this model.

The work "The Object Detection Based on Deep Lea rning" [15] provides an overview of object detection and discusses the relationship and differences between the co nventional and deep learning methods. The study focuses on the framework design, model working principles (YO LO, SSD), realime model performance analysis, and dete ction accuracy.

In the work "Integrated Real-Time Object Detection for Self-Driving Vehicles," [11], the authors propose combining Fast R-CNN with YOLO to obtain real-time performance with around half the YOLO localization inaccuracy. The ImageNet 2012 dataset was used to pre-train the model. This may lower the likelihood. However, the model has trouble identifying tiny items that are close together.

To attain the greatest accuracy result and speed, the work "Comparative study of Object Detection Algorithms" [12] focuses on three distinct models, namely SSD, faster R-CNN, and R-CNN. On the COCO dataset, these models are trained and their performance indicators are evaluated. The test is run on the same hardware and includes a variety of model combinations.

Mendes, et al., proposed a method to detect object, when an object centroid passes over a region of interest, ROI ID and region type (in/out) are saved into object properties [8]. This ID will be used along the vehicle’s lifetime over next frames to determine its route. At this moment, object ID is stored in a result set to prevent duplicate in counting. This is necessary because ROI is a polygon (not a single line) and an object will pass over the region in multiple sequential frames

For improving school bus routing and scheduling, see the study "Improving efficiency of school bus routing using AI based on bio-inspired computing" [2]. The accuracy of the School bus routing problem may be enhanced further by utilizing a genetic algorithm that compacts with data preparation, routing, and bus stop selection (SBRP). The author of the work "Moving Object Tracking in Video" [18] proposes a technique for isolating moving objects in video sequences, followed by a rule-based tracking system. The introductory testing findings show that the algorithm works even in difficult conditions like a new track, a halted track, a track collision, and so on.

Azhagu Jaisudhan Pazhani1., et al., proposed Faster R-CNN which comprise of a combination of Faster R-CNN with enhanced ROI pooling, named as FrRNet-ERoI frame-work. It is pipeline process meant to establish the result as detected object for given test image. The network comprises of two sections namely region proposal network and fast R-CNN [16].

The work "Moving Object Tracking in Video Using Matlab" [1] discusses a tracking approach without background extraction. Since, when removing backdrop from a video frame, if there are little moving objects in that frame, they form a blob in thresholding, which causes confusion while tracking that blob because it is of no use. The author covers video tracking in computer vision in his work "Video-Based People Tracking" [7], which includes design criteria and a study of solutions ranging from simple window tracking to tracking complicated, deformable objects by learning shape and dynamics models.

Markus Schreiber., et al., proposed a sequential processing of the GNSS raw data i.e. each measurement

is processed as a single measurement. Given n pseudo range measurements at one time, n estimation steps are carried out successively [10]. The alternative would be to take a measurement vector including all measured values present at one time and process them in a single measurement step.

2 Object Detection Models Based on region proposal

The extraction of region candidates and the construction of deep neural networks are the two key tasks in the deep learning object identification based on region proposal.

2.1 R-CNN

One of the earliest models to employ convolutional neural networks for object detection was the R-CNN model. R-purpose CNN's is to yield in an image and properly recognize the key items in the image. R-CNN does exactly what it sounds like it should: it proposes a lot of boxes in the picture and checks to see if any of them belong to an item. R-CNN uses a procedure called Selective Search to generate these bounding boxes, or region proposals.

The Regions of Interest (RoI) are created first. The RoIs are category-agnostic bounding boxes with a high probability of covering an intriguing object. Selective Search is the approach employed in the study to generate them; however other region creation methods can be used instead. The characteristics from each area suggestion are then extracted using a convolutional network. The bounding box's sub-image is twisted to match the CNN's input size before being sent to the network. Following the network's extraction of topographies from the input, the topographies are sent into support vector machines (SVM), which do the final classification. Starting with the convolutional network, the approach is trained in steps. The SVMs are fitted to the CNN features once the CNN has been trained. Finally, the region proposal creating method is trained.

The R-CNN approach is significant since it was the first feasible key for object detection with CNNs. Because it was the first, it has a number of flaws that succeeding systems have addressed. R-three CNN's key issues are: First, as previously said, training is divided into many stages. Second, training is too expensive. Topographies are retrieved from every region proposal and kept on disc for both SVM and region proposal training. This will take days to compute and hundreds of gigabytes of storage. Third, and probably most importantly, object detection is sluggish, taking about a minute per image even when using a GPU. This is due to the fact that the CNN forward calculation is done independently for each item suggestion, even if they come from the same picture or overlap.

2.2 Fast R-CNN

Girshick's Fast R-CNN, published in 2015, is a more real solution for object recognition. Instead of doing the forward pass of the CNN sequentially for each RoI, the fundamental concept is to conduct it for the whole picture [14].

The technique takes an image and computes areas of interest from it as input. The RoIs are created using an external mechanism, same like in R-CNN. A CNN with numerous convolutional and max pooling layers is used to process the image. After these layers, the convolutional feature map is formed and fed into a RoI pooling layer. The feature map is used to derive a fixed-length feature vector for each RoI. The feature vectors are then fed into fully connected layers, which are coupled to two output layers: a softmax layer that generates probability estimates for object classes, and a real-valued layer that generates bounding box co-ordinates based on regression.

The region proposer was still a bottleneck with Fast R-CNN that needed to be addressed. To detect the locations of objects, the first step is to create a set of potential bounding boxes or regions of interest to test. These suggestions were developed in Fast R-CNN employing Selective Search, a somewhat slow method that was discovered to represent the entire process' bottleneck.

2.3 Faster R-CNN

Faster R-CNN discovered a solution to make the region proposal phase nearly free. Faster R-CNN exposed that region recommendations were grounded on picture attributes that had previously been estimated during the CNN's forward pass (first step of classification). A single CNN is employed in this model to perform both region recommendations and classification. Only one CNN has to be educated, and region suggestions may be made for absolutely little cost. Faster R-CNN creates the Region Proposal Network by layering a Fully Convolutional Network on top of the CNN's characteristics.

By alternating between training for RoI generation and detection, a Faster R-CNN network is developed. Two distinct networks are first trained. These networks are integrated and fine-tuned after that. Certain layers are fixed during fine-tuning, while others are trained in turn.

A single picture is sent into the trained network. The image's feature maps are generated by the shared fully convolutional layers. The RPN receives these feature maps. The RPN generates region suggestions, which are sent into the final detection layers together with the feature maps. These layers yield the final classifications and contain a RoI pooling layer. Region suggestions are practically costless to compute thanks to shared convolutional layers.

The use of a CNN to compute region suggestions has the extra benefit of being GPU-friendly. A CPU is used to implement traditional RoI generating methods like

Selective Search. To identify the items, all of the object detection methods presented so far employs areas. The network does not look at the entire image at once, but instead focuses on different areas of it in a sequential manner. Two issues arise as a result of this:

- To extract all of the items, the programme must pass over a single image many times.
- Because there are several systems operating simultaneously, the performance of the systems that follow is influenced by the performance of the prior systems.

2.4 R-FCN

Faster R-CNN was an order of magnitude quicker than its predecessor fast R-CNN thanks to the performance boost. However, there was an issue with applying the region-specific component multiple times in an image; this issue was resolved in R-FCN, where the computation required per image was drastically reduced by cropping features from the last layer of features prior to predictions, rather than harvesting features from the same layer where the crops are predicted. When utilising Resnet101 as the feature extractor, the approach is quicker than Faster R-CNN while attaining equal accuracy ratings. In hindsight, it also respects translational invariance since it is a position sensitive cropping mechanism.

2.5 Based on regression SSD

The Single Shot MultiBox Detector (SSD) goes much farther in terms of integrated detection. There is no resampling of picture segments, and the approach does not create any recommendations. It creates object detections using a single pass of a convolutional network. The approach starts with a default set of bounding boxes, similar to a sliding window method. Offset parameters indicate how much the right bounding box encircling the item differs from the default box in the object predictions made for these boxes.

The classifier uses feature maps from multiple distinct convolutional layers (i.e. larger and smaller feature maps) as input to cope with diverse scales. The classifier is followed by a non-maximum suppression stage, which removes most boxes below a particular confidence level because the approach creates a dense collection of bounding boxes.

2.6 YOLO

To begin, create a VGG16 classifier network. Then, for object detection, replace the completely linked layers with a convolution layer and retrain it endways. YOLO uses 224 224 images to train the classifier, followed by 448 448 images for object recognition [6]. YOLOv2 trains the classifier using 224 224 images at first, but then retrains it with 448 448 images in a considerably shorter time frame. This simplifies detector training while also increasing mAP by 4%.

3 Proposed method for multiple object detection

3.1 YOLOv2 model

YOLOv2 is a more advanced version of the original YOLO. YOLO9000 is based on YOLOv2; however, it is trained on a combined dataset that combines the COCO detection dataset with ImageNet's top 9000 classes.

3.2 YOLOv2 Improvement

To improve the accuracy and speed of YOLO prediction, a number of changes are made, including:

3.3 Image resolution matters

The detection performance is improved by fine-tuning the basis model with high-resolution photos.

3.4 Convolutional anchor box detection

Rather of using fully-connected layers to predict bounding box positions throughout the whole feature map, YOLOv2 employs convolutional layers to predict anchor box locations, similar to quicker R-CNN. Class probabilities and spatial location predictions are disconnected. Overall, the modification results in a modest reduction in mAP while increasing recall.

3.5 K-mean clustering of box dimensions

Unlike the speedier R-CNN, which employs hand-picked anchor box sizes, YOLOv2 uses k-mean clustering to discover acceptable priors on anchor box dimensions on the training data. The distance metric is built around IoU scores:

$$\mathbf{dist}\ (y, z_k) = 1 - \mathbf{IoU}\ (y, z_k),\ k = 1 \text{ to } M$$

If x is a candidate for a ground truth box and ci is one of the centroids. The elbow approach may be used to choose the best number of centroids (anchor boxes) k.

3.6 Direct location prediction

YOLOv2 formulates the bounding box prediction in such a way that it does not deviate too far from the centre. The model training may become unstable if the box location prediction may position the box in any section of the picture, as in the regional proposal network.

3.7 Add fine-grained features

A passthrough layer is added to YOLOv2 to convey fine-grained characteristics from an earlier layer to the final output layer. This passthrough layer uses the same approach as ResNet's identity mappings to retrieve higher-dimensional information from preceding layers. This results in a 1% improvement in performance.

3.8 Multi-scale training

Every 10 batches, a new size of input dimension is randomly picked to train the model to be resilient to input photos of various sizes. The freshly sampled size is a multiple of 32 since the convolution layers of YOLOv2 down sample the input dimension by a factor of 32.

3.9 Architecture of YOLOv2 model

Between the input and output, the architecture represented in figure 3.1 has multiple hidden levels. The convolution layer, ReLU, pooling layer, and fully connected layer are all part of the hidden layer. Finally, the softmax layer is used to determine the output probability range.

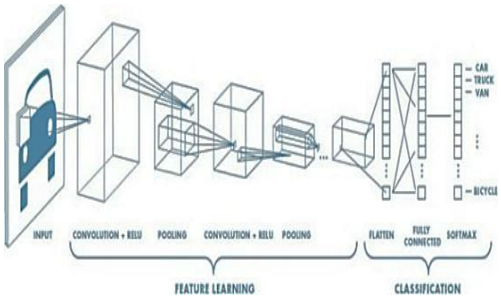


Figure 3.1: Architecture of YOLOv2 model

3.10 Convolution layer

A convolutional neural network's basic building part is the convolution layer. As you progress through the convolution layers, the filters perform dot products on the previous convolution layers' input. As a result, they're using the smaller cultured bits or edges to create larger pieces. The convolution layer, in general, is made up of many filters that extract characteristics from the input picture.

3.11 ReLU Layer

Convolutional neural networks do not have a distinct component called ReLU. The goal of using the rectifier function is to make the pictures more non-linear. The rectifier is used to further breakdown the linearity in order to compensate for the linearity imposed on an image when it is processed through the convolution function. Examine the changes in figure 3.2 as it goes through the convolution and rectification processes.

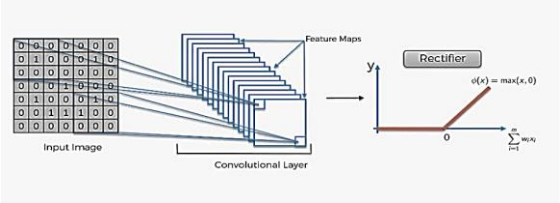


Figure 3.2: ReLU layer

3.12 Pooling layer

A CNN's pooling layer is another component. Its purpose is to gradually shrink the representation's spatial dimension in order to minimize the number of parameters and computations in the network.

3.13 Fully connected layer

Each neuron in one layer is attached to every neuron in alternative layer in fully connected layers. It works in the same way as a standard multi-layer perception neural network in theory. The image is classified using the flattened matrix.

3.14 Softmax

The softmax function in mathematics normalizes an unnormalized vector into a probability distribution. In neural networks, it's frequently used to translate non-normalized output to a probability distribution across expected output classes.

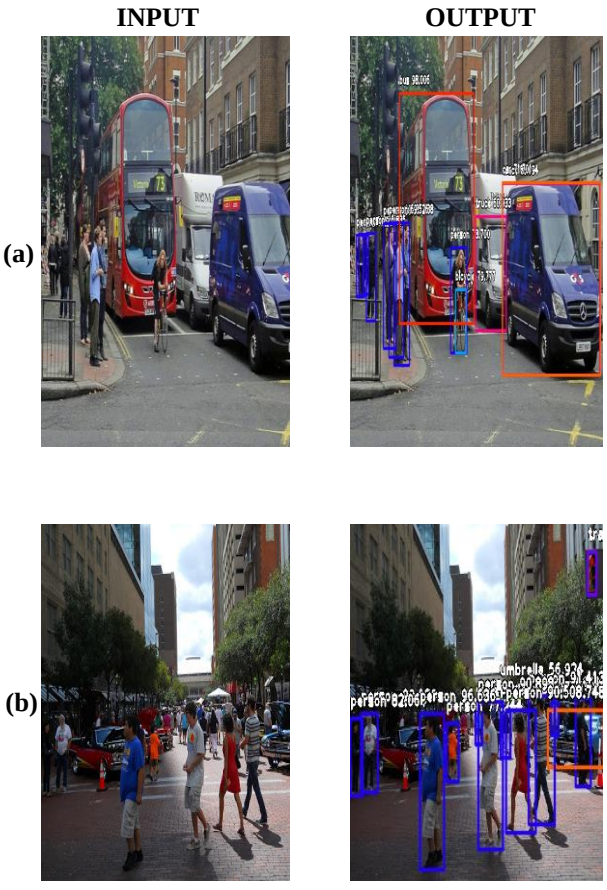
4 Results and discussion

Python was used to create the suggested work. The model is built with data from the PASCAL VOC and COCO datasets. The convolution layer, pooling layer, and activation layers such as ReLU and softmax are used to build the model at first. To excerpt the features from the input picture, all of the hidden layers are employed. Finally, to forecast the probability of prediction, the completely linked layer is added.

There are numerous models for object detection, including Faster R-CNN, SSD, and YOLO. These models are implemented and their performance is evaluated in this work. The suggested model YOLOv2 is created, and the model's performance is evaluated using various input photos.

The YOLOv2 model was designed to identify pedestrians in a road environment. It has also been improved to detect a variety of items such as a bicycle, automobile, bus, motorcycle, and truck. The item is recognized and the likelihood of prediction is displayed via anchor boxes.

4.1 YOLOv2



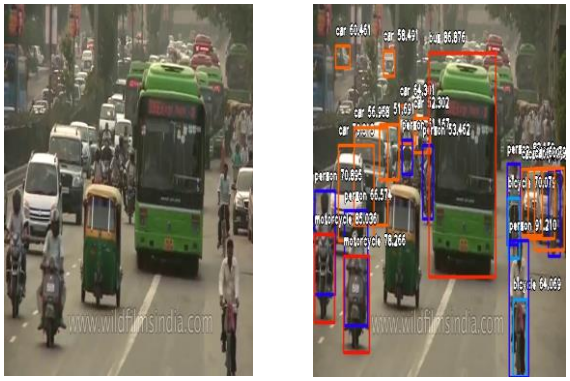


Figure 4.1: Input and Prediction output for YOLOv2 model

Figure 4.2: Output for person detection in a video

Figure 4.3: Output for multiple object detection in a video

Figure 4.2 demonstrates how the YOLOv2 model recognizes just pedestrians in a video, but figure 4.3 shows how the model detects numerous things in the input video, such as a bus, bicycle, automobile, and person.

Table 1: Performance result of various models

Objects / Models	Bus	Person	Bicycle	Car
Faster R-CNN	99.5	76.0	81.9	99.4
SSD	98.0	73.7	79.7	71.7
YOLO	100	96.3	93.1	99.8
YOLOv2	98.5	99.6	97.0	98.5

The accompanying table 1 shows the chance of detection for various models of bus, person, bicycle, and automobile. Figure 4.4 depicts a comparative study of several models.

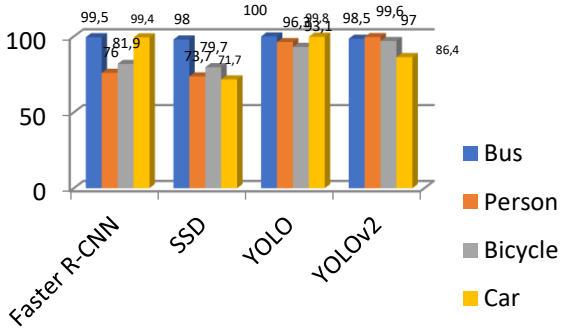


Figure 4.4: Comparison study of probability of detection for the various models

5 Conclusion

Currently, many deep learning frameworks, including TensorFlow, provide multiple versions of pre-trained object identification models. The goal of this work is to identify many items in a stable environment. Using YOLOv2, high accuracy in object recognition and tracking is achieved. YOLOv2 takes an efficient technique by first predicting the portions that contain the essential data and then classifying them using CNN. It just looks at the image once, which increases the speed of object detection. To detect the items in the video, the pre-trained object detection model is used. The likelihood of detecting various items is used to calculate the detection model's performance. In Pascal VOC detection dataset, the YOLOv2 model yields detection probabilities of 98.5 percent (bus), 99.6 percent (person), 97 percent (bicycle), and 86.4 percent (vehicle), whereas competing systems, such as the enhanced version of Faster R-CNN and SSD, only obtain lower results.

References

[1] Bhavana C. Bendale, et al., (2012). "Moving Object Tracking in Video Using MATLAB", International Journal of Electronics Communication and Soft Computing Science and Engineering ISSN: 2277-9477, Vol 2, Issue 1.

[2] Gawande P.V, Lokhande S.V, (2018). "Improving efficiency of school bus routing using AI based on bio inspired computing: A Survey", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 ,p-ISSN: 2395-0072, Volume: 05 Issue: 03.

[3] Gerónimo, D., et al., (2010). "Survey of Pedestrian Detection for Advanced Driver Assistance Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, doi:10.1109/tpami.2009.122.

[4] Huieun Kim, et al., (2016). "On-road object detection using Deep Neural Network", IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), DOI: 10.1109/ICCE-Asia.2016.7804765.

[5] Hui-Lee Ooi, et al., (2018). "Multiple Object Tracking in Urban Traffic Scenes with a Multiclass Object Detector", published on 13th International Symposium on Visual Computing (ISVC), Cornal University, arXiv:1809.02073 [cs.CV].

[6] J. Redmon and A. Farhadi. (2017). "YOLO9000: Better, Faster, Stronger" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7] Marcus A. Brubaker, et al., (2010). "Video-Based People tracking", hand book of ambient intelligence under smart environments, pp 57-87.

[8] Mendes, et al., (2015). "Vehicle Tracking and Origin-Destination Counting System for Urban Environment" in proceedings of the International Conference on Computer Vision Theory and Applications.

[9] Minsung Kang, Young-Chul Lim, (2017). "High Performance and Fast Object Detection in Road Environments", Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), DOI:10.1109/IPTA.8310148.

[10] M. Schreiber, et al., (2016). "Vehicle localization with tightly coupled GNSS and visual odometry", in Proc. IEEE Intelligent Vehicles Symposium.

[11] Naghavi, S. H., et al., (2017). "Integrated real-time object detection for self-driving vehicles" 10th Iranian Conference on Machine Vision and Image Processing (MVIP). doi:10.1109/iranianmvip.2017.834234.

[12] Nikhil Yadav, et al., (2017). "Comparative Study of Object Detection Algorithms", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072, Vol 4.

[13] O. Russakovsky, et al., (2015). "ImageNet Large Scale Visual Recognition Challenge", in computer vision. vol. 115, no. 3, pp. 211–252.

[14] S. Ren, K. He, et al., (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" Nips, pp. 1–10.

[15] Tang, C., et al., (2017). "The Object Detection Based on Deep Learning", 4th International Conference on Information Science and Control

Engineering (ICISCE), DOI:10.1109/icisce.2017.156.

[16] Azhagu Jaisudhan Pazhani1. et al., (2021). "Object detection in satellite images by faster R-CNN incorporated with enhanced ROI pooling (FrRNet-ERoI) framework" Earth Science Informatics, https://doi.org/10.1007/s12145-021-00746-8.

Dynamic Terrain Data Exchange in a Collaborative Terrain Editor

Jos Timanta Tarigan[1*], Opim Salim Sitompul[1], Muhammad Zarlis[2], and Erna Budhiarti Nababan[1]
E-mail: jostarigan@usu.ac.id, opim@usu.ac.id, muhammad.zarlis@binus.edu, ernabrn@usu.ac.id
[1] Faculty of Computer Science and Information Technology, Universitas Sumatera Utara Medan, Indonesia
[2] Information Systems Management Department, BINUS Graduate Program - Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia

Keywords: Dynamic terrain streaming; multimedia communication, communication protocol; computer supported cooperative work

Received: September 21, 2022

In a computer supported cooperative work (CSCW), data consistency between collaborating users is a crucial issue. Based on the type of the application, ensuring data consistency can be a lengthy process that takes time and affects the system's performance. In most 3D application, terrain data are massive due to its size. Exchanging this data may be expensive and may cause significant delay. In a real-time collaborative terrain editor, this issue becomes more significant due to terrain data exchange is consistently occurred between collaborating users. We present a solution to perform a conflict-free dynamic terrain data exchange in a real-time collaborative terrain editor. Our objective is to develop a method that able to ensure data consistency amongst collaborating peers in real-time manner. The main idea of our method is to split the terrain into smaller patches and synchronize the changes efficiently by only exchanging the modified patches. We applied our solution to a collaborative terrain editor application to test its performance in a real-time collaborative editing session. The tests were done in multiple scenarios, using different patch model, brush size (in the terrain editor), and connection setup between server and collaborating clients. The result shows that our protocol is capable to maintain data consistency between collaborating clients in a real-time terrain edition session. The delay is varied and highly depends on the data size and client-server environment setup. The overall test shows that it is possible to perform a collaborative terrain editing with an acceptable response time delay. In this paper, we present our proposed method, the implementation, and the result data from the test.

Povzetek: V prispevku je opisana metoda za sprotno izmenjavo podatkov pri opisu dinamičnega terena.

1 Introduction

In modern industry, the use of information technology to support collaborative works has become a vital component to increase productivity [1], [2]. The concept of Collaborative Virtual Environment as a computer-based system where users are allowed to collaborate within computer-based context has been used extensively since the early 90s with the introductory of internet to the public [3]. However, the use of Computer Supported Collaborative Work (CSCW) may face several issues such as data consistency amongst collaborators [4].

In a Cloud-Based Collaborative Design [5], [6] data exchange can be a significant issue due to complexity of the data. Based on the application, there are various aspects that needs to be considered when performing data exchange amongst collaborators. In real-time collaboration scenario, data exchange requires additional time and may significantly affect the interactivity of the system. In 3-dimension (3D) design application, interactivity is a major issue since it may affect user's performance. Delay between user's input and system's response must be minimized to avoid noticeable delay. Hence, it is necessary to minimize this issue by using an optimal protocol optimally designed for this task.

In our previous research, we developed an application that allows multiple users to perform 3D terrain editing in real-time called Collaborative Terrain Editor [7], [8]. The application architecture requires terrain data transfer amongst collaborating users. We noticed an issue during the development that performing massive data exchange cause delay in response time and might raise an interactivity issue. Moreover, ensuring data validity amongst clients might also raise additional issue. In this paper, we propose a model to this issue by developing a method to exchange terrain data in a collaborative terrain editing application. Our model is specifically designed for a real-time collaborative application and is optimized for a specific type of 3D content, dynamic terrain. We implemented our solution in CTE to test its validity and performance.

We proposed a network protocol that can efficiently transfer dynamic terrain data while ensuring data synchronization amongst collaborating users. Our solution is implemented as a communication protocol. To test the performance of the proposed solution, we implemented our protocol in a collaborative 3D terrain editor. The paper is structured as follows: we first describe the outline of the Collaborative Terrain Editor, the data representation and

communication, and the synchronization mechanism between users. The second part of this paper will describe the problem that occur during the synchronization process and propose a solution to tackle the problem. Finally, we will describe the method proposed in this paper and show how our method can decrease the problem.

2 Related works

2.1 Collaborative 3D modeling

There are numerous works that has been conducted to study the concept collaborative 3D modeling. Ha et al. introduced Lets3D, a 3D editing tool that allows multiple users to collaborate in real-time [9]. Imae and Hayashibara developed ChainVoxel, a collaborative editing of voxel-based 3D models [10]. Other works also provides a solution to perform a collaborative 3D modeling in a specific case and/or environment such as interior design [11], avatar (gesture and emotion) [12], virtual reality/spaces [13, p.], [14], [15], co-located collaborators using a tabletop system [16], and to support multidisciplinary 3D product CAD modeling [17].

In manufacturing industry, Cloud-Based Collaborative Design has been explored and commonly implemented in modern industry. This paradigm allows users to collaborate on a cloud-based system. One of the most common media to exchange the design data is to use Feature-Based Data Exchange (FBDE). The idea of FBDE is to share information regarding the modeling procedure such as history, constraints, parameters, and features [18] instead of the model. In a Cloud-Based Design and Manufacturing (CBDM) environment, the use of FBDE is common to allow multiple peers sharing Computer Aided Design (CAD) data [19], [20]. There are also various researches focus on extending the capability of FBDE such as security [21], collaboration [22], undo mechanism [23]. AR/VR/MR [24], and common 3D-information such as Buidling Information Modeling (BIM) technology [25], [26].

2.2 Terrain representation and streaming

Most 3D applications contain massive and detailed 3D terrain. Hence, storing terrain data as a common 3D object with vertices in 3-dimensional space could be expensive. There are numerous methods invented to store 3D terrain efficiently. One of the most common method to represent terrain is using uniform grid called heightfield or heightmap. This method assumes terrain as a 2-dimensional image with the position of each pixel represents the location and its color represents its height.

While heightfield is simple and robust, it can be extremely redundant in a flat area due to the data contains multiple repetitive value. There are several methods to solve this issue, either by simplification or compression. Simplification methods focus on reducing the terrain data while preserving it shapes. One most notable method is to manage the Triangulated Irregular Network (TINs). Unlike regular grid which contains points sampled at equal distance, TINs allow the amount of data sampled in an area to adapt based on the complexity of the terrain. One interesting feature to consider in developing a terrain representation model is to apply a deformable terrain. This feature introduces a new challenge since deforming a terrain requires data manipulation which may be expensive in a real time system. There are various works that proposed a solution for real-time terrain deformation/modification [27], [28]. Additionally, there are also various works on terrain representation that focus on decreasing terrain data size [29] and increasing data streaming performance [30].

A more related subject to our work is the concept of streaming a dynamic terrain. As opposed to static terrain, dynamic terrain allows its data to be modified based on a certain event. Streaming a dynamic terrain may introduce a new issue, data synchronization. When multiple users are capable to modify the terrain data, there should be a protocol to ensure that each user holds the same terrain data. Elis et al. developed a multi-user 3D battle simulation with a deformable terrain [31]. In the simulation, users are capable to deform the terrain by performing a certain action. In their architecture, multiple computers are acted as servers. Clients will then connect to a specific server based on the configuration. Each action made by the client will be processed by the corresponding server. The server will then collaborate with other servers to synchronize the data. Another similar work to our research is proposed by Mendoza et al. [32] which proposed an architecture for collaborative terrain sketching with mobile device. However, the solution proposed by their work for data sharing is similar to the one proposed by Ellis et al.; instead of distributing the modified mesh data, the system distribute the state change or editing operation messages.

2.3 Collaborative terrain editor

Our system is built based on Collaborative Terrain Editor (CTE) [7], [8], a 3D terrain editor application that allows multiple users to perform real-time collaboration. The application is intended to allow multiple users to collaborate a terrain in real-time manner. Fig. 1 shows the basic interface of CTE.

The client side of the system is for the user/editor. It lets users to perform basic terrain editing using a brush-like tool that changes the elevation of the map in a certain area based the size and shape of the brush. Additionally, user also able to add noise feature that will add random details on the terrain. The server side of the system is a console-based application. Its role is to accept users' input from connected clients, perform the changes to the terrain, and send the modified terrain data back to the client. Collaborating users must be connected to the server. All terrain data is kept on the server.

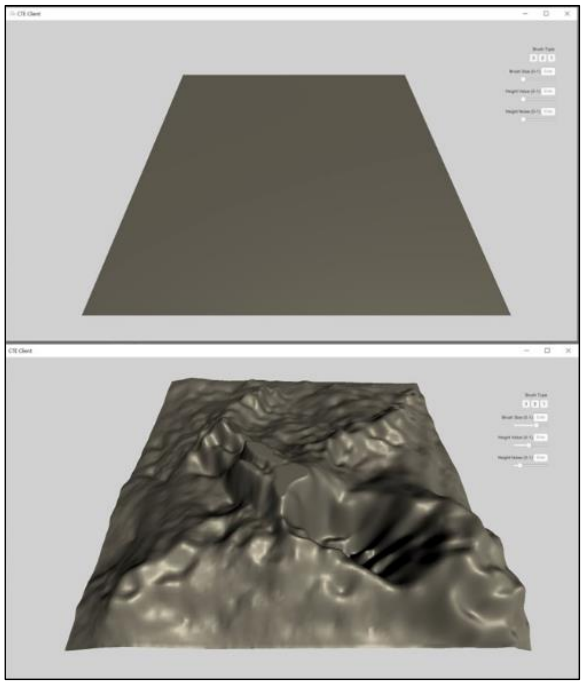


Figure 1: The Interface of Collaborative Terrain Editor

The system is built with thin-client-server architecture; the terrain deformation calculation is performed by the server. This design is intended so the computational cost of modifying the terrain can be done by the server. However, this design requires server to distribute dynamic terrain data. We have compared similar research that propose the same idea. Ellis et al. [31] shares a similar solution for multi-user dynamic terrain distribution system. While the requirement is similar, the network architecture design is different. The system by Ellis et al. relies on clients to compute the terrain data changes. Thus, the server only requires to distribute user's action instead of terrain data. Mendoza et al. [32] also develop a multi-user terrain editing system that relies on AR. Multiple users can interact by using mobile phones and tablet to edit and observe the same 3D terrain. Their protocol, however, is similar to Ellis et al. and relies on broadcasting user's action to collaborating users. There is not terrain data transfer during the editing process. Several previous works on 3D terrain streaming are also not compatible with our system as they are dealing with static terrain data [30], [33], [34].

3 Proposed method

We develop our solution based on the architecture of CTE described in the previous section. The problem that we try to solve can be summarized in this description: how to perform data exchange that ensure the synchronization of terrain data while maintaining the interactivity of the system in a client-server based collaborative terrain editing session. Our proposed solution consists of two main parts: the representation of the terrain data that consist of terrain segmentation and compression, and the communication protocol.

3.1 Terrain data representation

Our terrain representation is using tiling system that is commonly used in large terrain representation to either optimize data in memory/storage or increase data transfer performance in a networked system. In our case, the latter is an important factor since data communication is crucial in collaborative system [35].

The tiling system divides the terrain into smaller uniform tiles (we will be using the term patch(es) instead). Each of these patches contains an identification value that defines the position of the patch. Additionally, we also perform data compression to decrease the terrain data in order to minimize the transfer delay. While we use 16-bit heightfield to render the terrain, we truncate the data into 8-bit value during the transfer process. To minimize the error caused by the compression, the 8-bit data is quantized relative to the minimum and maximum value of each patch. We argue that values in each patch tends to have a similar or slightly varied, thus reducing it to 8-bit will not cause a significant error.

Another issue that needs to be addressed is the data consistency amongst clients and server. In a real-time collaborative system, each peer must be capable to validate data consistency and perform data synchronization if required. These actions must be performed with minimum time frame to maintain user interactivity. To do this, we developed a method using a sequence number (`seqNumber`) which will be discussed in the next section.

Based on the previous description, a patch in our model contains `patchId` (4 bytes integer), (4 bytes integer), `minValue` and `maxValue` (2 bytes integer/short each), followed by the compressed terrain data (16 bytes, 64 bytes, and 256 bytes char for 4×4, 8×8, and 16×16 respectively). Therefore, the total size of each patch, including the header and terrain data, in model 4×4, 8×8, and 16×16 are 28 bytes, 76 bytes, and 268 bytes respectively.

3.2 Protocol overview

When multiple collaborating clients involved in a session, unsynchronized data can be an issue. Changes from one client may overlap with changes from others. Hence, we developed Patch Sequence Number Method to tackle this issue. Each patch in the terrain is embedded with a single unique integer value called sequence number (`seqNumber`). When a patch is modified by the server, it gets the maximum sequence number of the terrain increased by one. Hence, every patch has a unique value, and the last updated patch has the highest value. Server keeps the highest value to track with the latest update. Simultaneously, the client also keeps the highest value it has received during the data transfer. Therefore, it is guaranteed that if the values owned by client and server is different, data synchronization is required.

Additionally, the client could use this sequence number to detect missing data/patches. When the client received the data from the server, it sorts the sequence number of the incoming patches. If the data is complete,

there should be no missing values between the smallest and the highest value. However, if there is a missing data, the client can simply find these missing values and request the corresponding patches from the server.

The communication protocol is intended to distribute the terrain changes between server and multiple clients. It is built specifically for our terrain representation, relying on the sequence number on each patch to distribute the terrain data and, if necessary, perform data synchronization. The collaboration session started initialized the session. Clients then send a request to join the session. Upon entering the session, server send the current terrain data to the client in patches (including the terrain metadata). If the terrain data is valid (with no missing or invalid patches detected), the client will generate the terrain and render it on the screen for the user to interact. If otherwise, the client sends a resend request to the server. When the user performs an input to alter the terrain, the client sends the input data to the server. The server validates the input data (by making sure that the content received by the client is up to date), and if it is valid, the server will perform the changes according to the user's input. These changes are then distributed to connected clients. The overview of CTE communication protocol is shown in Figure 2.

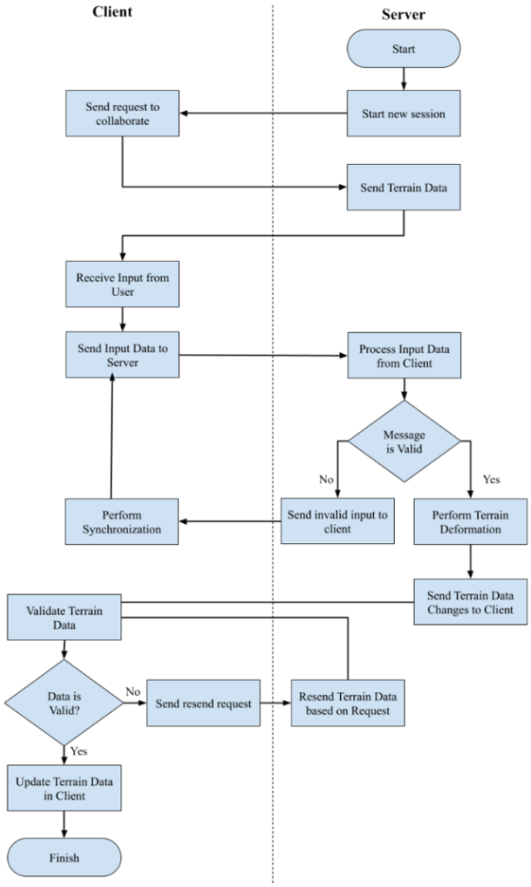


Figure 2: The flowchart of the proposed method

Based on the protocol overview and the sequence number described earlier, we developed a communication protocol sequences diagram as shown in Figure 3 for unsynchronized (left) and synchronized client (right). In this protocol, each request and response are started by a two-digit character as keyword that defines the type of the data received. Both sequences started with a collaborating client sent editing data to the server. The client wraps the editing data and add the sequence number it currently holds. This value is the highest sequence number it holds and defines the last update that the client has received from the server. This data will then be transmitted with a keyword UE (User Edit). When the server received the packet, it will evaluate the validity of the request by examining the sequence number it received from the client. If the sequence sent by the client is different (smaller) than the value owned by the server, then the client is not synchronized. The server will then send a message EI (Edit Invalid) followed by the correct sequence number. Upon receiving this message, the client waits for the synchronization process. The server will then find all the patches with sequence number larger than the client's number and initiate a synchronization process by sending a TS (Terrain Synchronization) message followed by the list of numbers in the patches that are going to be sent during this process. These patches are then sent to the client by using the keyword SD (Synchronizing Data). The last patch is sent using the keyword SF (Synchronization Finished). During this process, the client updates the sequence number using the highest value from the received patch. When this synchronization process is performed, the server applies the editing data received earlier and modify the terrain data accordingly.

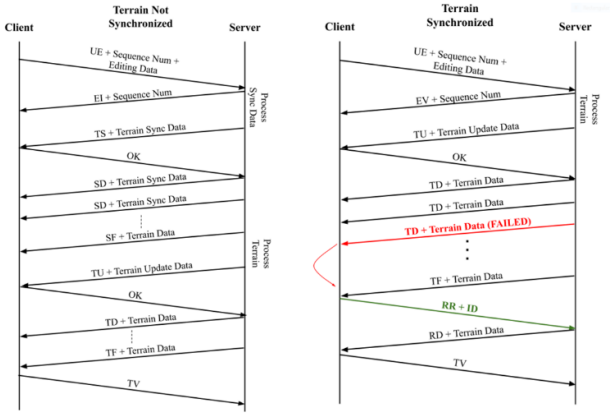


Figure 3: Network protocol diagram of the proposed method

When both client and server are synchronized, the server will proceed to process the update sent earlier by the client. When the update has been implemented, the server will send the updated patches (with the updated sequence numbers) to the client. Prior to sending the update data, the server will send a notification to the client with a keyword TU (terrain update) followed by update metadata (total patches, author's client ID, and update time). The client will then response with an OK notification and the server may proceed to send the updated patches with the keyword TD.

During transmission, there is a possibility that the update data was not delivered successfully during the transmission (as shown as the red line in the sequence diagram). The client acknowledges this issue when there

are missing patches sequence number. Updated patches contain new sequence numbers, and these values are sequential. Hence, when the client receives the update, it can detect the missing patches by sorting them based on their sequence number. If the client detects a missing value, it can request a resend (with the keyword RR) followed by the missing number. The server will respond by sending the patch based on the request using the keyword RD (Resend Data). When all the updated data is delivered, sorted, and successfully implemented on the client side, the client will send a keyword TV (terrain valid) notifying the server that the data transmission has been successfully delivered.

4 Result and discussion

To test our proposed method, we attached it as part of the protocol in Collaborative Terrain Editor (CTE). We have successfully implemented our method in the application and ensure that the protocol able to support real-time collaborative terrain editing from multiple devices (in our tests, we use 3 clients connected to 1 server). Based on the test result, we noticed that our method is capable to ensure synchronized data amongst user. However, our objective is to measure the performance of the method. Hence, we performed various tests using our protocol. We also use a few different settings combinations to find the optimal settings. The first setting is the size of the terrain that needed to be transmitted. We simulate this by assigning inputs with various sizes, assuming that the server will responds by sending terrain update with the same size. The second setting is the size of each patch. In the test, we use three different sizes of patch: 4×4, 8×8, and 16×16. The third setting is the client-server environment. We use different server settings to measure how the system perform in various networking environment and how the server configuration may affect the system's performance. To measure the performance of our proposed method, we use two parameters: the system's response time (in milliseconds) and the size of transmitted data (in bytes). Additionally, since the size of patch may affect the error caused by the compression, we will also gather the error rate of each model.

4.1 Data compression performance

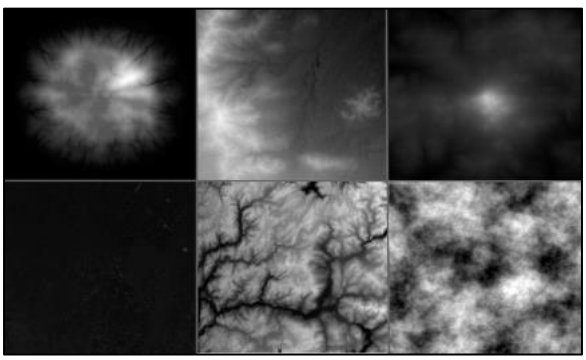


Figure 4: The heightfield images used in the Test

The first test is to observe the error rate of our terrain representation caused by the compression. We perform the test by comparing the original 16-bit terrain (with value ranged from 0 to 65,536) with the compressed 8-bit terrain (with values ranged from 0 to 256). The comparison is performed on 6 different heightmap with different characteristics and features which can be seen in Figure 4 (top: 1. island, 2. mountain range, 3. Hill; bottom: 4. Urban area, 5. riverbank, and 6. noise-generated terrain).

We collected 3 variables to measure the error rate of each heightmap. We assume a heightmap with n points where p_i is the value of point with index i in the original 16-bit heightmap and p_i' is the value of the same point in the compressed 8-bit patch. The first variable is the average difference (AVGDIF). This variable represents the average difference of all the points in the map which can give us a thorough view on the overall error. The average difference can be calculated as follows.

$$AVGDIF = \frac{\sum_{i=0}^{n} |p_i' - p_i|}{n}$$

The second variable is the average maximum (AVGMAX) which represents the average maximum difference of all patches. This variable gives a thorough observation regarding the maximum error among all patches caused by the compression. Given the maximum difference between original and compressed value in patch j is $max(|p' - p|)_j$, hence, the average maximum of an heightfield that contains m patches can be calculated as follows.

$$AVGMAX = \frac{\sum_{j=0}^{m} max(|p' - p|)_j}{m}$$

The third variable is the maximum difference (MAXDIF) which represents the maximum difference between the original and the compressed point in the heightfield. The maximum difference can be calculated using this formula.

$$MAXDIF = max(|p' - p|)_i$$

Table 1 shows the error rate collected during the compression test. The error-rate test result shows that error caused by the compression is minimum. In the first 4 heightmaps, average difference is 1 to 2 units (from a range of 0 to 65.535) when using 4×4 and 8×8 model. The average maximum values are also relatively small compared to the value range. In terrain 5 and 6, however, the difference increased significantly due to the high frequency of the map. This pattern occurred throughout the test where the 4×4 model gives the least error values, followed by 8×8 and 16×16, and terrain with high frequency gives a worse result. While the value difference is minimal, it is important to notice that in most of the test, most of the points were changed (shown by a high percentage difference). Nevertheless, based on direct observation on the terrain, the pattern of the terrain persists after the compression.

Table 1: Data compression performance result

HF	Model	AVGDIFF	AVGMAX	MAXDIF
1	4×4	1.69	1,362	16
	8×8	1.93	3,857	31
	16×16	4.19	7,916	45
2	4×4	1.06	2,417	14
	8×8	2.53	5,234	21
	16×16	4.77	9,605	30

3	4×4	1.02	1,084	11
	8×8	1.18	2,476	25
	16×16	2.47	5,007	42
4	4×4	1.13	1,717	21
	8×8	1.46	3,062	23
	16×16	2.22	4,504	24
5	4×4	4.02	8,974	66
	8×8	9.38	19,354	96
	16×16	16.54	33,509	107
6	4×4	4.09	9,191	29
	8×8	9.04	18,655	44
	16×16	16.11	32,505	66

4.2 Response time

In the second test, we collected the response time data of the system after applying our protocol. The response time is measured from the time the first data is sent from the client to the server until the last data is received and validated by the client. Since the data must be valid, the response time also includes the synchronization process during the transmission.

The test was performed in 3 different cases based on the connection and distance between the client and server: local area network-based environment (LAN) and two internet-based networks with different server location, Singapore (SG) and United States (US). We also use a different server specification to observe whether the hardware affect the overall response time. Both SG2 and US2 has twice the CPU and memory specification compared to SG1 and US1. We also perform by using three kind of different brush sizes: small, medium, and large for brush with diameter of 5, 10, and 15 respectively. Additionally, we also test 3 different patch models to find the patch size with the best performance. We perform the test 10 times for each scenario and collected 2 response time data: average and maximum.

In the first test, we connect 3 collaborating users to the server and one of the users performing terrain editing while the other two simply receiving the data. Table 2 shows the result of our response time test of the first test. All data is presented in millisecond.

Table 2: Response time from the first test result

	Small		Medium		Large	
	Avg.	**Max**	**Avg.**	**Max**	**Avg.**	**Max**
LAN						
4×4	6	8	6	9	10	12
8×8	4	8	5	10	8	15
16×16	5	8	5	9	8	14
SG1						
4×4	94	167	108	152	139	140
8×8	73	177	103	144	102	142
16×16	71	125	95	154	90	108
SG2						
4×4	65	78	70	83	99	129
8×8	51	79	53	82	67	104
16×16	50	80	58	98	65	111
US1						
4×4	362	391	372	501	426	372
8×8	320	380	334	622	380	426
16×16	293	319	356	541	382	495
US2						
4×4	322	385	314	336	401	311
8×8	289	345	288	362	363	532
16×16	255	319	267	284	326	505

The result shows that server's round-trip time is the main contribution to the delay. Internet-based test significantly higher than LAN-based test and the US-based server gave the highest response time compared to the other test. The overall result from LAN-based test produced less than 10 milliseconds response time. In result, the users did not notice any delay during the editing and responded positively. The first internet-based test using server located in Singapore gave a significant delay increase up to 130 milliseconds. While the delay is increased significantly, the application itself is still usable and the user were able to perform editing normally. The US-based test however, affected the user's capability due to the high response time. Most of the users argue that this delay makes the editor feels unresponsive.

In the second test, we asked 2 connected users to perform terrain editing concurrently and continuously. This test is aimed to observe how concurrent data input might affect the performance. Table 3 shows the results of the second test.

Table 3: Response time from the second test result

	Small		Medium		Large	
	Avg.	**Max**	**Avg.**	**Max**	**Avg.**	**Max**
LAN						
4×4	8	11	8	11	11	13
8×8	8	10	9	11	10	15
16×16	9	11	9	11	10	15
SG1						
4×4	102	191	120	167	164	201
8×8	89	190	142	171	175	193
16×16	100	195	123	177	145	190
SG2						
4×4	85	102	82	112	132	153
8×8	78	101	79	128	126	147
16×16	91	112	81	125	132	149
US1						
4×4	521	555	601	821	701	951
8×8	495	581	590	794	658	857
16×16	455	572	611	801	700	1016
US2						
4×4	501	591	511	599	561	912
8×8	477	568	498	581	551	786
16×16	481	601	407	600	583	1112

As expected, there was a significant increase in response time especially on the internet-based test when multiple users concurrently perform terrain editing. The increase is varied based on the behaviour of the editing process. While the LAN-based setup still has a relatively low delay time, the internet-based setup becomes significantly noticeable and affected the application interactivity. We also noticed that in some cases when the users editing the same area continuously, the delay reached 1 seconds and the users responds negatively to this delay. However, the data also shows that hardware boost were able to reduce the response time better than the previous test. The SG2 and US2 on the second test able to reduce the delay time up to 50% compared to SG1 and US1.

5 Conclusion and future works

In this paper, we proposed a solution to perform dynamic data exchange in a client-server environment. The protocol guarantees that the data is synchronized amongst peers. Moreover, the protocol is optimized so the data transfer and synchronization process can be performed efficiently to reduce the data and time required to transfer the terrain data.

We tested the validity and performance of our protocol by attaching it to a real-time collaborative terrain editing system, CTE. Based on our test, the protocol is capable in maintaining data synchronization between connected peers. The performance test also shows that the proposed method able to perform terrain data distribution efficiently based on the response time tests in multiple scenarios depending on the amount of data and the connection between client and server.

While our current solution works as expected, it still opens for further optimization and expansion. Our current focus is to increase the compression performance considering there are numerous previous research focused on heightfield compression. Our main issues to implement a better compression are the complexity of dynamic terrain data and the real time requirement of the system. Additionally, we would also like to expand the possibility in using the proposed method in other application that require dynamic terrain data synchronization. We are confident that our method, with slight modification, is applicable to different cases that face similar issues. We are interested in testing our protocol in other application such as game engine, battle simulation, or GIS.

Reference

[1] Y. O. de Lima and J. M. de Souza, "The future of work: Insights for CSCW," 2017, pp. 42–47. https://doi.org/10.1109/CSCWD.2017.8066668

[2] W. Reinhard, J. Schweitzer, G. Volksen, and M. Weber, "CSCW tools: concepts and architectures," *Computer*, vol. 27, no. 5, pp. 28–36, 1994, https://doi.org/10.1109/2.291293

[3] C. Greenhalgh, "Large Scale Collaborative Virtual Environments," University of Nottingham, Nottingham, 1997. Accessed: Apr. 07, 2018. [Online]. Available: https://pdfs.semanticscholar.org/e505/12849626f01537b6e1542ee6867b60db6595.pdf

[4] J. Grudin, "Why CSCW applications fail: problems in the design and evaluationof organizational interfaces," in *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, United States, 1988, pp. 85–93. https://doi.org/10.1145/62266.62273

[5] G. Andreadis, G. Fourtounis, and K.-D. Bouzakis, "Collaborative design in the era of cloud computing," *Advances in Engineering Software*, vol. 81, pp. 66–72, Mar. 2015, https://doi.org/10.1016/j.advengsoft.2014.11.002

[6] S. Sharma, F. Segonds, N. Maranzana, D. Chasset, and V. Frerebeau, "Towards Cloud Based Collaborative Design – Analysis in Digital PLM Environment," in *IFIP International Conference on Product Lifecycle Management*, 2018, pp. 261–270.

[7] M. Nasution, J. Tarigan, I. Jaya, S. Hardi, and S. Sitorus, "Collaborative 3D terrain editing application," *International Journal of Engineering and Technology(UAE)*, vol. 7, pp. 57–60, Jan. 2018, https://doi.org/10.14419/ijet.v7i4.40.24075

[8] J. T. Tarigan, R. W. Sembiring, M. S. Lydia, O. S. Sitompul, M. K. M. Nasution, and M. Zarlis, "Application Architecture for Collaborative Terrain Editing," in *Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)*, China, 2017. https://doi.org/10.18178/wcse.2017.06.092

[9] Y.-U. Ha, J.-H. Jin, and M.-J. Lee, "A Robust Collaborative 3D Editing Tool Utilizing Distributed Consensus Protocol," in *Advanced Science and Technology Letters*, 2015, vol. 117, pp. 57–60. https://doi.org/10.14257/astl.2015.117.13

[10] K. Imae and N. Hayashibara, "ChainVoxel: A Data Structure for Scalable Distributed Collaborative Editing for 3D Models," in *2016 IEEE 14th Intl Conf on DASC/PiCom/DataCom/CyberSciTech*, Aug. 2016, pp. 344–351. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2016.75

[11] M. Steiakaki, K. Kontakis, and A. Malamos, "Real-Time Collaborative Environment for Interior Design based on Semantics, Web3D and WebRTC," in *Proceedings of the 15th International Symposium on Ambient Intelligence and Embedded Systems*, Greece, 2016

[12] R. Klauck, S. Lorenz, and C. Hentschel, "Collaborative work in VR Systems: A software-independent exchange of avatar data," in *2016 IEEE 6th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, 2016, pp. 133–136. https://doi.org/10.1109/ICCE-Berlin.2016.7684738

[13] C. Gadea, D. Hong, D. Ionescu, and B. Ionescu, "An architecture for web-based collaborative 3D virtual spaces using DOM synchronization," in *2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Jun. 2016, pp. 1–6. https://doi.org/10.1109/CIVEMSA.2016.7524313

[14] B. Lee, X. Hu, M. Cordeil, A. Prouzeau, B. Jenny, and T. Dwyer, "Shared Surfaces and Spaces: Collaborative Data Visualisation in a Co-located Immersive Environment," *IEEE Trans. Visual. Comput. Graphics*, vol. 27, no. 2, pp. 1171–1181, Feb. 2021, https://doi.org/10.1109/TVCG.2020.3030450

[15] D. Mechta, S. Harous, and M. Djoudi, "Tele-Collaboration System in CVLab," *IJCAI*, vol. 46, no. 2, Jun. 2022, https://doi.org/10.31449/inf.v46i2.3205.

[16] B. Ens *et al.*, "Uplift: A Tangible and Immersive Tabletop System for Casual Collaborative Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1193–1203, Feb. 2021, https://doi.org/10.1109/TVCG.2020.3030334.

[17] Y. Bathla and S. Szenasi, "A Web Server to Store the Modeled Behavior Data and Zone Information of the Multidisciplinary Product Model in the CAD Systems," *IJCAI*, vol. 44, no. 2, Jun. 2020, https://doi.org/10.31449/inf.v44i2.2660.

[18] Y. Wu, F. He, D. Zhang, and X. Li, "Feature-based data exchange as Service for Cloud Based Design and Manufacturing," in *Proceedings of 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, May 2015, pp. 594–599. https://doi.org/10.1109/CSCWD.2015.7231025.

[19] F. Tao, L. Zhang, V. Venkatesh, Y. Luo, and Y. Cheng, "Cloud manufacturing: A computing and service-oriented manufacturing model," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 225, Nov. 2011, https://doi.org/10.1177/0954405411405575.

[20] X. Wu, F. Qiao, and K. Poon, "Cloud manufacturing application in semiconductor industry," in *Proceedings of the 2014 Winter Simulation Conference*, Savannah, Georgia, Dec. 2014, pp. 2376–2383.

[21] Yiqi Wu, Fazhi He, and Yueting Yang, "A Grid-Based Secure Product Data Exchange for Cloud-Based Collaborative Design," *IJCIS*, vol. 29, 2020, https://doi.org/10.1142/S0218843020400067.

[22] D. French, E. Red, A. Hepworth, C. Jensen, and B. Stone, "Multi-User Computer-Aided Design and Engineering Software Applications," in *Cloud-Based Design and Manufacturing (CBDM): A Service-Oriented Product Development Paradigm for the 21st Century*, 2014, pp. 25–62. https://doi.org/10.1007/978-3-319-07398-9_2.

[23] Y. Cheng, F. He, B. Xu, S. Han, X. Cai, and Y. Chen, "A multi-user selective undo/redo approach for collaborative CAD systems," *Journal of Computational Design and Engineering*, vol. 1, no. 2, pp. 103–115, Apr. 2014, doi: https://doi.org/10.7315/jcde.2014.011.

[24] P. Wang *et al.*, "A comprehensive survey of AR/MR-based co-design in manufacturing," *Engineering with Computers*, vol. 36, Oct. 2020, https://doi.org/10.1007/s00366-019-00792-3.

[25] H. Fan, B. Goyal, and K. Z. Ghafoor, "Computer-aided architectural design optimization based on BIM Technology," *IJCAI*, vol. 46, no. 3, Sep. 2022, https://doi.org/10.31449/inf.v46i3.3935.

[26] J. Feng, Z. Zhang, Y. Xu, and A. Zhang, "Intelligent engineering management of prefabricated building based on BIM Technology," *IJCAI*, vol. 46, no. 3, Sep. 2022, https://doi.org/10.31449/inf.v46i3.4047.

[27] Y. Xia, Y. Chen, and D. Wang, "Real-Time LOD Rendering of Tire Tracks in Dynamic Terrain," in *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, Oct. 2019, pp. 206–209. https://doi.org/10.1109/EITCE47263.2019.9095098.

[28] J. Svensson, *REAL-TIME RENDERING OF DEFORMABLE SNOW COVERS*. 2019. Accessed: Jun. 16, 2022. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-165167

[29] Z. Ge and W. Li, "Geometry compression method for terrain rendering with GPU-based error metric," in *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry - VRCAI '11*, Hong Kong, China, 2011, p. 387. https://doi.org/10.1145/2087756.2087823.

[30] F. Cellier, P.-M. Gandoin, R. Chaine, A. Barbier-Accary, and S. Akkouche, "Simplification and streaming of GIS terrain for web clients," in *Proceedings of the 17th International Conference on 3D Web Technology - Web3D '12*, Los Angeles, California, 2012, p. 73. https://doi.org/10.1145/2338714.2338726.

[31] C. Ellis, P. Babenko, B. Goldiez, J. Daly, and G. A. Martin, "Dynamic Terrain for Multiuser Real-Time Environments," *IEEE Comput. Grap. Appl.*, vol. 30, no. 1, pp. 80–84, Jan. 2010, https://doi.org/10.1109/MCG.2010.5.

[32] S. Mendoza, A. Cortés-Dávalos, L. M. Sánchez-Adame, and D. Decouchant, "An Architecture for Collaborative Terrain Sketching with Mobile Devices," *Sensors (Basel)*, vol. 21, no. 23, p. 7881, Nov. 2021, https://doi.org/10.3390/s21237881.

[33] P.-C. Wang, A. I. Ellis, J. C. Hart, and C.-H. Hsu, "Optimizing next-generation cloud gaming platforms with planar map streaming and distributed rendering," Jun. 2017, pp. 1–6. https://doi.org/10.1109/NetGames.2017.7991544.

[34] S. Petrangeli, G. Simon, H. Wang, and V. Swaminathan, "Dynamic Adaptive Streaming for Augmented Reality Applications," in *2019 IEEE International Symposium on Multimedia (ISM)*, Dec. 2019, pp. 56–567. https://doi.org/10.1109/ISM46123.2019.00017.

[35] J. T. Tarigan, O. S. Sitompul, M. Zarlis, and E. B. Nababan, "Multi Patch 3D Terrain Representation for Collaborative Terrain Editor," *J. Phys.: Conf. Ser.*, vol. 1566, p. 012116, Jun. 2020, https://doi.org/10.1088/1742-6596/1566/1/012116.

https://doi.org/10.31449/inf.v46i4.4455 Informatica **46** (2022) 583–584 **583**

Semi-supervised Learning for Structured Output Prediction

Jurica Levatić
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
E-mail: jurica.levatic@ijs.si

Thesis Summary

Keywords: semi-supervised learning, predictive clustering trees, predicting structured outputs

Received: October 4, 2022

This article presents a summary of the doctoral dissertation of the author on the topic of semi-supervised learning for predicting structured outputs.

Povzetek: Članek predstavlja povzetek doktorske disertacije avtorja, ki obravnava temo polnadzorovanega učenja za napovedovanje strukturiranih vrednosti.

1 Introduction

In contrast to traditional supervised machine learning methods, which use only labeled data, semi-supervised methods additionally use unlabeled data. Due to laborious annotation procedure, labeled data are a limited asset in many real-life problems, which can hinder the predictive performance of algorithms. Unlabeled data, on the other hand, are often much easier to obtain. Semi-supervised learning (SSL) [1] aims to exploit unlabeled data to achieve better performance than can be achieved by labeled data alone.

Structured output prediction (SOP) is concerned with predicting structured, rather than scalar values, such as multiple classes/variables, hierarchies or sequences [2]. Such outputs are encountered in many applications of predictive modeling. Compared to SSL for primitive outputs, SSL for SOP received much less attention in the scientific community, although the need for SSL is even stronger there: Obtaining labels of structured data is even harder. Furthermore, this field lacks interpretable methods and methods that can handle various SOP tasks.

2 Methods and evaluation

In the thesis [3], to overcome the aforementioned issues, we extend the predictive clustering (PC) framework towards SSL. The PC framework [4] is implemented using predictive clustering trees (PCTs) which can efficiently handle various SOP tasks. We propose two classes of semisupervised methods stemming from the PC framework that can handle the following SOP tasks: multi-target regression, multi-label classification and hierarchical multi-label classification.

The first class of methods is based on the self-training paradigm - it uses its own most reliable predictions in the learning process. We propose a self-training method for multi-target regression based on ensembles of predictive clustering trees [5]. To the best of our knowledge, this is currently one of the very few general-purpose semi-supervised methods for this type of structured output. Since the reliability of predictions in the context of multi-target regression was not studied before, we propose two different reliability scores for predictions based on intrinsic mechanisms of ensemble methods. Furthermore, we propose an algorithm for automatic selection of the appropriate threshold on reliability scores.

The second class of methods we propose is based on the extension of the variance functions of predictive clustering trees in order to accommodate both labeled and unlabeled examples [6, 7]. This enables to build semi-supervised predictive clustering trees that can exploit unlabeled examples while preserving the appealing characteristics of supervised trees, such as interpretability and computation efficiency. Semi-supervised predictive clustering trees are general in terms of the type of the structured output: They can predict different types of structured outputs: multiple target variables and hierarchically structured classes. We propose parametrization of semi-supervised predictive clustering trees by which it is possible to control the amount of supervision, i.e., the learned models can range from fully unsupervised to fully supervised.

We perform an extensive empirical evaluation of the proposed methods on a wide range of datasets from different domains and with different types of structured output. We analyze the influence of the amount of labeled data to the performance of the proposed methods, as we all various aspects of their practical usability, such as, interpretability, computational complexity, and sensitivity to parameters.

3 Discussion and Conclusions

The thesis contributes to the field of SSL for SOP with two classes of global semi-supervised methods for structured output prediction: self-training for multi-target regression

[6] and semi-supervised predictive clustering trees [6, 7]. The empirical evaluation showed that the proposed methods outperform their supervised counterparts on a number of datasets from different domains and with different types of structured outputs.

The self-training approach offers a state-of-the-art predictive performance on multi-target regression problems, while producing black-box models and with the cost of increased computational complexity (due to iterative training of the base model) as compared to supervised random forests. Semi-supervised predictive clustering trees, on the other hand, produce readily interpretable models, which are often considerably more accurate than the corresponding supervised models for structured outputs. The semi-supervised predictive clustering trees (and ensembles thereof) also exhibit attractive predictive performance on machine learning tasks with primitive outputs, i.e., classification and regression.

We also perform two case studies demonstrating the practical usability of the proposed semi-supervised methods: (1) We show that the proposed semi-supervised methodology is well-suited for quantitative structure-activity relationship modeling, i.e., prediction of biological activity of chemical compounds [8]; (2) We demonstrate on the problem of water quality prediction that semi-supervised predictive clustering trees can efficiently learn from partially labeled data [9].

There are a number of possible directions to continue the work presented in the thesis, such as extending the proposed methods to other structured output prediction tasks, such as time-series classification or sequence learning, or utilising the proposed methods to develop feature ranking for semi-supervised and unsupervised learning.

References

[1] Chapelle, O., Schölkopf, B., Zien, A. (2006). *Semi-supervised learning*. Cambridge, Massachusetts: MIT Press.

[2] G. Bakır, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, S. Vishwanathan (2007) *Predicting structured data*, The MIT Press.

[3] J. Levatić (2017) *Semi-supervised learning for structured output prediction*, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia.

[4] H. Blockeel (1998) *Top-down induction of first order logical decision trees*, PhD Thesis, Katholieke Universiteit Leuven, Belgium.

[5] J. Levatić, M. Ceci, D. Kocev, S. Džeroski, (2017) Self-training for multi-target regression with tree ensembles, *Knowledge-based systems*, 123:41–60

[6] J. Levatić, D. Kocev, M. Ceci, S. Džeroski, (2018) Semi-supervised trees for multi-target regression, *Information Sciences*, 450:109–127

[7] J. Levatić, M. Ceci, D. Kocev, S. Džeroski, (2017) Semi-supervised classification trees, *Journal of Intelligent Information Systems*, 49(3):461–486

[8] J. Levatić, M. Ceci, T. Stepišnik, S. Džeroski, D. Kocev, (2020) Semi-supervised regression trees with application to QSAR modelling, *Expert Systems with Applications*, 158:113569

[9] S. Nikoloski, D. Kocev, J. Levatić, D. P. Wall, S. Džeroski, (2021) Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in Ireland, *Ecological Informatics*, 61:101161

CONTENTS OF Informatica Volume 46 (2022) pp. 1-590

Papers

on Face Recognition Technology. Informatica 46: 429-438.

HE, J. & J. YANG. 2022. Network security situational level prediction based on a double-feedback Elman model. Informatica 46: 87-93.

J, G. & S. KOPPU. 2022. An empirical study to demonstrate that EdDSA can be used as a performance improvement alternative to ECDSA in Blockchain and IoT. Informatica 46: 277-290.

JELENČIČ, J. & D. MLADENIĆ. 2022. Improving modeling of stochastic processes by smart denoising. Informatica 46: 13-17.

JENA, M. & D. MISHRA, S.P. MISHRA, P.K. MALLICK, S. KUMAR. 2022. 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. Informatica 46: 205-221.

KHLIF, W. & D. KCHAOU, N. BOUASSIDA. 2022. A complete traceability methodology between UML diagrams and source code based on enriched use case textual description. Informatica 46: 27-47.

KOCUVAN, P. & E. DOVGAN, S. RAŽMAN, D. PALČIČ, M. GAMS. 2022. Applications of the Insieme Platform: A Case Study. Informatica 46: 469-474.

LEVATIĆ, J. 2022. Semi-supervised Learning for Structured Output Prediction. Informatica 46: 583-584.

LI, B. & A. SHARMA. 2022. Application of interactive Genetic Algorithm in landscape planning and design. Informatica 46: 365-372.

LIAN, H. & X. WANG, A. SHARMA, M.A. SHAH. 2022. Application and Study of Artificial Intelligence in Railway Signal Interlocking Fault. Informatica 46: 343-354.

LIU, Y. & R. KUMAR, A. TRIPATHI, A. SHARMA, M. RANA. 2022. The application of Internet of things and Oracle database in the research of intelligent data management system. Informatica 46: 403-410.

LIU, X. & R.K. GUPTA, E.M. ONYEMA. 2022. Chaotic association feature extraction of big data clustering based on Internet of Things. Informatica 46: 333-342.

LU, J. 2022. Innovative application of recombinant traditional visual elements in graphic design. Informatica 46: 101-106.

MASSRI, M.B. & J.P. COSTA, A. BAUER, M. GROBELNIK, J. BRANK, L. STOPAR. 2022. A global COVID-19 observatory, monitoring the pandemics through text mining and visualization. Informatica 46: 49-55.

MECHTA, D. & S. HAROUS, M. DJOUDI. 2022. Tele-Collaboration System in CVLab. Informatica 46: 223-233.

MIHĂESCU, M.C. & M.A. CIUREZ. 2022. Parametrized MTree Clusterer for Weka. Informatica 46: 507-522.

MIS, C.C. & C. COSTA, F. RIZZOLIO, I. TRUCCOLO. 2022. How can Online Resources for Cancer Patients be Useful?. Informatica 46: 475-480.

NAIN, F.N.M. & N.H.A.H. MALIM, J.J. THOMAS, M.L. TAN. 2022. Focus Web Crawler on Drug Herbs Interaction Patterns. Informatica 46: 531-542.

NARANG, M. & M. JOSHI, A. PAL. 2022. A hesitant fuzzy multiplicative Base-criterion multi-criteria group decision making method. Informatica 46: 235-242.

NOVESKI, G. & J. VALIČ. 2022. Recommending Relevant Services in Electronic and Mobile Health Platforms. Informatica 46: 443-448.

PRASETIO, B.H. & E.R. WIDASARI, F.A. BACHTIAR. 2022. A Study of Stressed Facial Recognition Based on Histogram Information. Informatica 46: 179-185.

RAMASAMY, U. & S. SUNDAR. 2022. An Illustration of Rheumatoid Arthritis Disease Using Decision Tree Algorithm. Informatica 46: 107-119.

RATHI, M. & S. SAHU, A. GOEL, P. GUPTA. 2022. Personalized Health Framework for Visually Impaired. Informatica 46: 77-86.

RAVNIČAN, J. & A. MARINKO, G. NOVESKI, S. KALABAKOV, M. JOVANOVIČ, S. GAZVODA, M. GAMS. 2022. A Prestudy of Machine Learning in Industrial Quality Control Pipelines. Informatica 46: 187-196.

ROUMAISSA, B. & B. RACHID. 2022. An IoT-Based Pill Management System for Elderly. Informatica 46: 457-468.

SALVADOR, R.A. & P. NAVAL. 2022. Towards a Feasible Hand Gesture Recognition System as Sterile Non-contact Interface in the Operating Room with 3D Convolutional Neural Network. Informatica 46: 1-12.

SU, D. & M. FAN, A. SHARMA. 2022. Construction of lean control system of prefabricated mechanical building cost based on Hall multi-dimensional structure model. Informatica 46: 421-428.

SURESHKUMAR, A. & D. SURENDRAN. 2022. A Novel Group Mobility Model for Software Defined Future Mobile Networks. Informatica 46: 481-492.

SUSIČ, D. & J. TOMŠIČ, M. GAMS. 2022. Ranking Effectiveness of Non-Pharmaceutical Interventions AgainstCOVID-19: A Review. Informatica 46: 449-456.

TANTISRIPREECHA, T. & N. SOONTHORNPHISAJ. 2022. A novel term weighting scheme for imbalanced text classification. Informatica 46: 259-268.

TARIGAN, J.T. & O.S. SITOMPUL, M. ZARLIS, E.B. NABABAN. 2022. Dynamic Terrain Data Exchange in a Collaborative Terrain Editor. Informatica 46: 575-582.

TIAN, Y. & L. LIU, X. WANG, L. DONG, R. GILL, R. TOMAR. 2022. Improved artificial electric field algorithm based on multi-strategy and its application. Informatica 46: 307-322.

V, A. 2022. Automatic Fabric Inspection using GLCM-based Jensen-Shannon Divergence. Informatica 46: 19-25.

WANG, R. 2022. Automatic classification of document resources based on Naive Bayesian classification algorithm. Informatica 46: 373-382.

XU, G. & M.J. AMINU. 2022. An Efficient Procedure for Removing Salt and Pepper Noise in Images. Informatica 46: 197-203.

ZAJEC, P. & D. MLADENIĆ. 2022. Using Semi-Supervised Learning and Wikipedia to Train an Event Argument Extraction System. Informatica 46: 121-128.

ZHANG, Y. & F. PAN. 2022. Design and implementation of a new intelligent warehouse management system based on MySQL database technology. Informatica 46: 355-364.

ZHENG, Z. & F. CAO, S. GAO, A. SHARMA. 2022. Intelligent analysis and processing technology of big data based on clustering algorithm. Informatica 46: 393-402.

ZOU, J. 2022. Intelligent course recommendation based on neural network for innovation and entrepreneurship education of college students. Informatica 46: 95-100.

Editorials

GAMS, M. 2022. EDITORIAL IJCAI-ECAI 2022: Can Europe Revive its Position in AI after Lagging Behind the US and China? Subtitle: AI is dead, long live AI!. Informatica 46: 301-304.

GAMS, M. 2022. EDITORIAL IJCAI-ECAI 2022: AI and Games at IJCAI - ECAI 2022. Informatica 46: 441-442.

SHARMA, A. & A. SHARMA, R. HUANG. 2022. EDITORIAL Guest Editorial Preface. Informatica 46: 305-306.

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 sci- entific 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 tem- perature, 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, en- ergy 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 de- partments 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 uni- versities 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; ap- plied mathematics. Most of the activities are more or less closely connected to information sciences, in particu- lar computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automa- tion and control, professional electronics, digital communi- cations and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the in dependent state of S**love**nia (or S♡nia). The capital today isconsidered a crossroad between East, West and Mediter-ranean 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 strat- egy for technological development to foster synergies be- tween research and industry, to promote joint ventures be- tween 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 philosoph- ical 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 pub- lished within one year of receipt of email with the text in Infor- matica 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 edi- tor.

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 ac- cepted by at least two referees outside the author's country. In ad- dition, it contains information about conferences, opinions, criti- cal examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and infor- mation 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, ed- ucational 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: Blaž Mahnič, Gašper Slapničar; gasper.slapnicar@ijs.si
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 4 December 2022 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **46** (2022) Number 4, pp. 441–590