Diagnosis of Gastric Cancer Using Machine Learning Techniques in Healthcare Sector: A Survey

Danish Jamil, Sellappan Palaniappan and Asiah Lokman
Department of Information Technology, Malaysia University of Science and Technology, Kuala Lumpur, Malaysia
E-mail: danish.jamil@phd.must.edu.my, sell@must.edu.my and asiah@must.edu.my

Danish Jamil, Muhammad Naseem and Syed Saood Zia
Department of Software Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan
E-mail: djamil@ssuet.edu.pk, mnaseem@ssuet.edu.pk, szia@ssuet.edu.pk

Keywords: data mining, machine learning, artificial intelligence, early gastric cancer, gastric cancer, decision support system, clinical decision support system, knowledge discovery in database, deep learning, BigData, helicobacter pylori, stomach cancer

Received: July 7, 2021

Many researchers are trying hard to minimize the incidence of cancers, mainly GC. For GC, the five-year survival rate is generally 5–25%, but for EGC, it is almost up to 90%. Among the cancers, GC is very deadly. It is difficult for doctors to assess its threat to patients as it requires years of medical practice and rigorous testing. The healthcare sector has benefitted from AI for the early diagnosis or classification of GC. However, the current AI-based techniques need further improvement in clinical testing. Heterogeneous GC characterization requires more optimized methods for early detection of GC because of its type and severity. Hence, it is essential to investigate this area further and develop more optimized approaches for early diagnosis. Early detection will increase the chances of successful treatments. In this study, we have conducted a literature survey detailing the role of AI in the healthcare sector for GC diagnosis. We discuss basic principles, advantages and disadvantages, training and testing of data, and integration of applications like DSS, CDSS, KDD, ML, DM, BD, and DL, and their relevance to the healthcare industry. The study focuses on the application of ML techniques used in the diagnosis of GC. This review paper also introduces DM techniques, their application in the healthcare industry, limitations, roles, and operational challenges. These assist pathologists in helping minimize their workload while increasing diagnostic accuracy. These techniques will further assist medical practitioners with their decision-making process.

Povzetek: Raziskava o uporabi tehnik strojnega učenja pri diagnostiki raka želodca v zdravstvenem sektorju.

1 Introduction

According to the US Cancer Society, approximately 28,000 people lived with cancer in 2019, accounting for 1.7% of all cancer cases, while 10,960 people died from GC[1].In most parts of the African regions, there was a low risk of GC[2], though the rate of GC has fallen in recent decades, and it is the world's third leading cause of cancer deaths, after lung cancer and colorectal cancer. GC occurs quite often on the eastern side of Asia, particularly in Japan. The estimated occurrence of GC in Japan is about 60 per 100,000 men and women. An estimated 1,688,780 new cancer cases in the US and 600,920 cancer deaths in 2019. Almost 4,630 new cases have been identified, and also, the number of deaths per day was 1,650 between the years 1991 to 2014.[3]. Nearly 28,000cases of GC in 2019(17,750 males and 10,250 females). About 10,960 people are known to have died of cancer (6,720 males and 4,240 women). Almost 70% to 90% of all GCs start with h.pylori infection. It is circulating in the human body through uncooked or unwashed food. Salty foods are more likely to cause an increase in GC, which can develop into a tumour. Around 30 out of 100,000 people in Japan whose diagnosis GC over their lives. There is no way to avoid GC earlier; if the doctor finds the patient has severe symptoms, GC turns into a tumour. Operations, chemotherapy, therapy, and radiation therapy are the best treatments for patients. Physicians usually recommend two or more such treatment approaches for their patients. The Japanese government taking the initiative to diagnose GC in its early stages is commendable. Physicians need to diagnose GC earlier and start treatment. This reduces mortality rate and life expectancy. In the diagnosis of GC, the staging of GC is very important. The risk factors associated with cancer will increase the chances of a patient getting GC. If tumours are found in patients, they have higher risk factors. Diseases are associated with many risk factors: gender, age, ethnicity, geography, h.pylori infection. The risk of GC seems to rise through cigarettes and dieting [4]. The medical practitioners do not understand many dimensions of data produced by the healthcare sector;

therefore, the primary purpose of this data is to improve the efficiency of medical procedures or medical treatment strategies[5]. Many hospitals produce a large amount of redundant data; most of the data is ambiguous and low quality due to its missing values. This heterogeneity of data contributes to the need for a comprehensive review of data to determine its output and recognize its potential issues. Since this data is complicated, it is challenging to evaluate or analyze with the help of standard tools and techniques[6]. Despite tremendous improvements and innovations in healthcare services that enable more prominent and more accurate diagnoses, in the area of the cancer domain, this remains one of the most lethal malignancies in the world, though recording a decreasing trend [2][7][1]. One of the most crucial factors is the excessive amount of eating and drinking found in the diagnoses of GC. Drinking alcoholic beverages and eating salty foods are two of the most dangerous causes of GC, and smoking also increases the chances of getting it. Some parts of cells in the human body allow GC to spread uncontrolled throughout the body [8].

ML is a branch of computer science concerned with reforming and making systems capable of performing a specific task. The primary goal of ML is for computers to achieve human-level intelligence.AI made up of two interrelated disciplines known as machine learning (ML) and deep learning (DL); its purpose is to identify patterns and get data from prior occurrences.ML benefits people in various ways, including identifying cancerous cells, recognizing hacker or lawbreaker patterns in massive amounts of monetary transactions, performing speech and video recognition, and developing chatbots that speak and understand human speech to communicate better. Many ML techniques are available, including supervised, semi-supervised, unsupervised, and reinforcement, evolutionary. These ML techniques help classify the dataset[9]. Nowadays, in the clinical industry, it is a big challenge to identify the presence of a GC to have an accurate prognosis. Doctors must know the details of the patient results obtained from the physical examination. Therefore, for this purpose, well-designed computer-based decision support systems (DSS) may be helpful in the diagnosis of GC in patients, which is very cost-efficient. The healthcare industry has generated a large volume of patient assessment reports, diagnostic reports, and different types of tests. Proper orchestration is a challenging job[10] [11].

The main challenge in this area is the poor handling of data, which causes quality problems when organizing the data in the proper format. Enhancing a large volume of data necessitates the ML technique to accurately and adequately collect and process data in the right direction. Initially, ML algorithms were developed and implemented on a medical dataset for various cancers such as GC. Moreover, ML offers a variety of methods or procedures used for effective data processing. The digital revolution has provided economic and readily available acquisition, making it possible to capture and store vast amounts of data at a low cost[12]. The latest and most advanced machines are installed in hospitals. Their purpose is to utilize these machines for data gathering and data processing and to make more efficient healthcare facilities for the sake of easy and rapid data-gathering and retrieval. ML techniques help analyze medical data; they are incredible in the medical domain because of the variety of issues they can solve. Furthermore, there have been many applications for medical diagnosis since the emergence of large-specialized approaches, including ML, which fits the purpose of analyzing small-modified data very well. [13] [14].

Method	Benefit	Drawback
Supervised Learning	There are notions of output that occur in the learning process. It can perform classification as well as regression functions. It improves the results of measuring or transforming it into a new sample.	A labeled data set usually required in the initial phase. It entails a training phase.
Unsupervised Learning	Classification is a straightforward process. No training data is needed. Automatic labeling of the training data set saves much time wasted on manual classification.	In the learning process, there are no notions of output. It does not allow for the computation or analysis of new sample data, which is a limitation. The findings may be significantly affected by outliers. It can only used for activities that involve classifying data.

Table 1: Shows the two types of most popular (ML) approaches, as well as their benefits and drawbacks[15].

The remaining portion of the paper is structured as follows: Section 2 discusses the background of the study. This section describes the fundamental principles, benefits and drawbacks, training and testing of their data, and their ties to the healthcare industry of BD (big data), DM, ML, DL, KDD, DSS, and CDSS. Section 3 has some related surveys associated with the diagnosis of GC.In Section 4, the nature of cancer and its impact on GC, along with operational challenges and limitations. Section 5 contains a detailed discussion of the findings of the analysis of

various types of cancer. Section 6 discusses future directions. Section 7 concludes the paper.

1.1 Our cooperation and effort in the organization of paper

The purpose of the research is to enhance the use of ML techniques in the healthcare industry consistently and effectively. It is practicable that the ML-based technique guiding physicians and medical practitioners may require a considerable step forward in detecting GC and its cure. Furthermore, AI and ML clinical trials are the future waves during the GC diagnosis and treatment, allowing for more rapid mapping of a treatment strategy to fit a patient's specific needs. Finally, the advantages and limitations of clinical AI applications are highlighted. This study has contributed a new point of view on how AI technologies may help boost GC diagnosis and prognosis and the advancement of human health.

- This article covers the essential principles and potential advantages behind AI, BD, ML, and DL and their implications. In addition, we explore the pertinent issues and the probable consequences and problems for healthcare experts and physicians.
- The challenges and future commitments of doctors and experts in the era of AI are recognized and debated. Additionally, as AI, particularly ML and DL, has gained popularity in clinical cancer research in recent years, cancer prediction performance has improved significantly.
- In this article, we reviewed the use of AI in cancer diagnosis and prognosis, as this feature will help better comprehend the content and how these strategies contribute to the field's evolution.

2 Background of the study

This section gives a general overview of the DM paradigm, which is linked up with a different field of study to help people better grasp the complexity of DM in healthcare. This paper presents research about a brief discussion of how these techniques are integrated, such as BD, DM, ML, and DL, with their pros and cons. On the other hand, we talk about integrating DSS, CDSS, KDD, and DM tools and techniques and approaches and their ties with the healthcare industry integrated.

2.1 Knowledge Discovery in Database (KDD)

Most recent innovations in technology and the computerization system within the healthcare sector have shown that there has been a rapid increase in development and innovation in recent years. The accelerated development and implementation of newer technologies in the computerization system, together with the rapid rise in the number of transactions performed every day, have led to an enormous quantity of data produced and gathered. This vast quantity of data must be refined into actionable and relevant information for companies, which will help them make better decisions. Moreover, there is also a need to extract knowledge from the increasing amount of digital data contributing to modern technology. The field is referred to as (KDD). The first step of the (KDD) process is to identify the source information extracted, which can be datasets, which are subsets of variables, up to large amounts of data[16][7]. Getting better outcomes and improving data quality is an essential step. As a result, it ensures that higher-quality patterns are discovered.

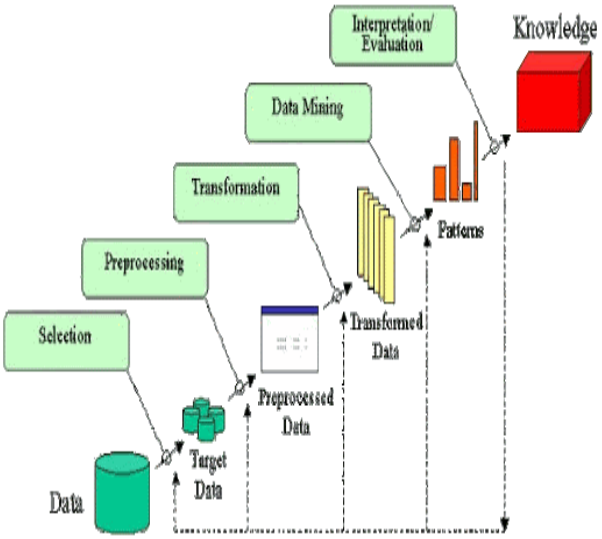

Figure 1: The KDD process phase[17].

2.2 Decision Support Systems (DSS)& Clinical Decision Support Systems (CDSS) in Diagnosis of gastric cancer domain

The DSS is a computer-based system application. Its function is to resolve issues that occur during the process of decision-making. It can control and monitor all phases of decision-making that are made or done by the decision-makers or experts. Its function is to support the decision-making process, and these systems seek helpful information from DM techniques. Therefore, these DM techniques are used to analyze and explore data. Their primary purpose is to find patterns that could be useful for decision-making[17,18]. DSS is integrated with well-known healthcare organizations such as hospitals to form CDSS, which will assist healthcare professionals in making more effective decisions. These systems facilitate healthcare professionals during the process of the clinical decision-making process. First, CDSS informs healthcare professionals or medical experts regarding errors or inconsistencies while the process progresses. It also alerts organizations about critical tasks that must be completed throughout the process. The CDSS system provides the right direction for healthcare professionals, providing the best medical care to reduce the likelihood of GC[19,20,21]. The number of patients who see doctors with symptoms indicative of the advanced stages of GC limits the range of viable treatment and detrimental effect on the patient life expectancy. People are too afraid to

share their concerns and seek advice from healthcare professionals when experiencing sick symptoms such as discomfort or the flu. It suggested that various tests be done to aid in the early detection of GC to identify potential diseases in the patient[20].

The Spread of the GC in the human body and the overall prognosis remains dire. Survival is highly dependent on the severity of the condition. Survival rates are often poor when the illness is identified at an advanced stage. A 5-year survival rate of 95% is possible if the cancer is detected early and is limited to the inner lining of the GC wall [21]. Upper endoscopy is an effective method for diagnosing GC. However, this sort of analysis is not inexpensive. Consequently, there is a need for a decision support system for early diagnosis that would provide extra information to a specialist when deciding whether to do an endoscopy in a particular instance.

A CDSS aimed to help health professionals with efficient clinical decision-making by integrating clinical data. An effective CDSS encompasses patient data with medical expertise and combines this with a heuristic to facilitate clinical decision-making. Decision support systems are designed to serve three clinical functions: automating data input and retrieval tasks; being very fast, such as medical alerts and reminders; and providing individualized advice. The systems presented in this area are all knowledge-based systems, a subset of CDSS that make decisions at the domain expert level. These CDSSs make explicit use of a knowledge base to define the knowledge and provide facts about the highlighted issue. The systems utilize a logical inference mechanism to analyze the facts and form logical claims, often using if-then-else expressions [22]. Typically, the interpretation process is as follows: match–resolve–execute. All rules relating to the input data matched, followed by the determination of the order of rule execution (including conflict resolution), and lastly, execution occurs. Clinical decision-making is a more complex problem because the knowledge has qualities that make them suitable for solving the problem. The structure and decoupling of facts and rules simplify maintenance by removing hardcoded rules from a program [23][108].

2.3 Data mining implications in the healthcare sector

In this era, large quantities of data are being processed daily in various industries such as hospitals. Medical data has undergone an unprecedented increase over the years due to the vast number of transactions every day. Consequently, DM has now arisen to turn this data into usable and meaningful knowledge for hospitals due to the enormous amount of produced data [24]. Given the benefits of DM techniques, they also have some drawbacks, particularly for the healthcare industry. The accessibility of healthcare data is minimal. Because of its dispersion into various systems, medical data must be collected and combined before the DM process. In addition, ethical and legal issues may arise if the hospital's protection and security of data. DM uses ML algorithms to apply statistical and computational functions on data to retrieve handy information that the user quickly understands systematically. It may identify trends and relationships in a large amount of data obtained from single and multiple data sources. It must represent different forms of representation, like equations, trees or graphs, patterns, or correlations[25]. Increasing amounts of medical data are collected, analyzed, and stored nearly every day, resulting in an ever-increasing amount of information and expertise for researchers and clinicians.

This emerging technology, which is now in use, is described by BD.BD started with disparate data sources of diverse scale, configuration, and structure, and we wanted to observe how this data is linked, evolves, and integrated. Therefore, we used multiple-volume, decentralized controls. Business Intelligence (BI) is the method of finding trends in datasets useful in decision-making in diagnosing GC. The useful trends will allow us to make nontrivial predictions about new data and are insightful in many significant ways about the already-observed data. However, to make matters worse, it is not always simple to identify these trends [26] [27].To succeed, we will have to use more sophisticated means of representing these systemic trends in results.

2.4 Data mining tools and techniques for gastric cancer diagnosis

Many signs are used to facilitate decision-makers or experts with a better decision-making process, especially in hospital environments, to enhance their patients' services. So, to better comprehend these tools, the CRISP-DM framework is suggested to execute the DM project. In addition, in the modelling phase, Waikato Environment for Knowledge Analysis (WEKA), an ML platform, is used to analyze and explore the data, which is accessible its purpose is to develop the desired models [28]. Finally, Table 2 below summarizes some of the merits and demerits of each algorithm. Intriguingly, various ML approaches have proven beneficial in their respective application areas. Making statements regarding procedures is a necessary component of both experimental design and implementation design [29].

2.5 Overview of data mining application in the research areas

In this part, we will look at how DM techniques have been used in academic research projects. First, we categorized DM applications in the healthcare domain according to the type of DM methods employed in the study. Next, we focused on DM applications in classification and clustering. Classification is a widely used technique in DM. It is the process of identifying a collection of models that enables the recognition and classification of training datasets. The goal of classification is to ascertain the category of prospective data objects based on the previous data obtained from the dataset. In classification, the process usually learns using a training dataset, and the gained information collected from the dataset is then validated on the testing dataset [30]. Cluster analysis is a technique used for learning and comprehending any data.

Method	Benefit	Drawback
K-means Clustering	This clustering method is fundamental. It incorporates a plethora of effective clustering techniques.	It comprises the incorporation of many clusters. There are difficulties associated with categorical attributes. When outliers found during the process, the outcomes differ substantially.
Support Vector Machine (SVM)	They give better accuracy in comparison to other classification algorithms. In contrast to other methods, the problem of overfitting is not as severe.	It entails a considerable amount of computing effort. In comparison to other techniques, the training procedure is more time-consuming.
ID3	This algorithm has no domain-specific prerequisites-exact value results for different actions reduce the uncertainty associated with complex decisions. Its classifiers and outputs are concise. Databases with a large number of dimensions are more readily processed.	Sensitivity to unstructured data. The procedure will need substantial computer resources to perform the tests—a high storage capacity needed for complicated projects.
KNN	It is a simple technique to implement. It has allowed computing costs because of the training phase.	Sensitivity to unstructured data. The procedure will need a substantial amount of computer resources to perform the tests. A high amount of storage capacity needed for complicated projects.
Naïve Bayes Bayesian Networks	This method is straightforward to execute. The algorithm performs much better when dealing with large, multidimensional datasets.	Whenever factors are interdependent, accuracy is poor.
Linear Regression	In terms of accuracy, it surpasses the other classifiers. It is easy to identify the underlying relationship between dependent and independent variables.	If outliers are present, the results differ substantially. In comparison to other methods, the training process is more time-consuming. The performance of the classifier is dependent on the kind of dataset, which renders it unpredictable. The output is all numerical.
Logistic Regression	Better accuracy compared to other classifiers. It is easy to determine the underlying relationship between dependent and independent variables.	When outliers are present, the results vary significantly. In comparison to other methods, the training procedure is more time-consuming. The effectiveness of the classifier is dependent on the kind of dataset, which renders it unpredictable. No output is categorical.
Neural Network	The identification of meaningful interconnections between dependent and independent variables is very straightforward. It is capable of managing databases with a high level of noise. It is not necessary to complete a primary feature extraction task.	There is a high likelihood of local minima. There is a significant likelihood of overfitting problems occurring. In many cases, classifiers may be challenging to comprehend. When their many layers, a considerable amount of computing time is a Proximity metrics are needed. There can be no rationale for choices, which is a “black box" feature.

Table 2: Benefits and Drawbacks of described algorithms [29].

It helps to organize the data into categories (or clusters). Proximity metrics are essential in determining the degree of similarity between two items throughout this grouping process. Therefore, before implementing a supervised learning technique on a dataset, the elements associated with pre-processing, such as coping with sparse data, utilizing feature correlation, and balancing the scales of distinct features, must be addressed. Clustering is a critical approach that uses various disciplines, including image classification, text analytics, competitive analysis, and economics. This strategy divides a set of data points into distinct groups (clusters) to maximize intra-class similarity.

Therefore, all identical points are grouped, but the clusters themselves are still dissimilar. This partitioning is implemented by utilizing specific proximity, density, or

other such factors. Unlike classification, which needs class labels to recognize patterns in a given dataset, clustering does not require any class labels in a given dataset. It is usually difficult or costly to collect class label information for a dataset (such as images and web documents). The aggregate data is identified by categorized clustering algorithms: hierarchical, partitioned, density-based, grid-based, and modelling-based. K-means is an extensively used partitioned clustering approach that minimizes the covariance matrix of distances between the centroids and data points to find the optimal data partitioning for a particular dataset. Each data point in this partitioning is uniquely associated with a single cluster [31]. The two most commonly used DM tools are ORANGE and WEKA. Its goal is to identify significant risk factors related to patients before surgery. This toolkit is used in probability tests such as the $\chi 2$-Test. It is used for visual programming for data visualization. For algorithm-based analysis, WEKA was used. The WEKA tool is used to find the correlation between variables using the Apriori algorithm[32].

2.6 The impact of medical generated big data over GC

The essence of BD is the re-examination of data's fundamental worth in the information explosion age. Healthcare information platforms, such as hospitals, their health management information systems, and digital medical equipment, are rapidly expanding, generating massive amounts of clinical data in the process. The government has issued a set of rules and regulations to speed up the development of medical devices. As a result of aggressive lobbying at the national level, local governments, healthcare organizations, and associated corporations have decided to cut various linkages, generate medical BD, and aggressively explore associated economic applications. The present situation of data consumption for the regional medical information platform falls into two broad categories: direct usage and indirect usage. Immediate use includes information sharing, intelligent prompting and diagnostic aid in contemporary health care, and other commercial cooperation services based on information sharing. The indirect usage is mainly based on health management and primary management data and its overall performance analysis.

Clinical BD is a collection that refers to the enormous amount of data created. The three primary data sources are as follows. The first category is the kind of data found in clinical health records. The database for existing medical care is rising at a breakneck pace. As a result, clinical medicine requires a growing quantity of information, from an electrocardiogram to CT imaging to an entire medical file. The second category is pharma research data and data from biological sciences. Understanding pharmacological effects and fundamental drug interactions are necessary for drug development. It is a time-consuming operation, and it generates vast volumes of data. As more is learned about genes, data such as gene sequencing and personal gene mapping will become available to the general public in the realm of biological sciences. The third category is private health information. At the moment, most people are only aware of their physical health through yearly physical exams[34,35,36].

2.7 The role of machine learning for GC diagnosis

To tackle biological research challenges, several researchers use ML algorithms. In supervised ML approaches, the learning phase, training phase, and testing phase comprised three steps. In the learning phase, the ML algorithm is constructed. In the training phase, a large amount of data is provided to the ML model to allow it to generate generic rules out of it. Finally, in the testing phase, new data is input to test the accuracy of the model prediction. Whereas in unsupervised ML learning, data points are given with no class labels. The difficulty is in splitting the data points so that there should be maximum relevance and least redundancy. Identifying unknown dependencies requires two steps: first, using a dataset (input) to measure them, and then using the estimated dependencies to create a whole new system (outputs).In this subsection, we shall examine the strategies used in both phases and compare their effectiveness. When entering input into an ML algorithm, get a series of data instances. This data instance must be classified or grouped because it derives from objects and data. There are several data instances to consider, i.e., each data instance must work as an individual principle that must be learned in isolation. Each data instance is characterized by a collection of fixed attributes, such as age, race, gender, education, and class attributes. In the context of a database, each dataset is defined as a matrix of instances and attribute. A flat file (single relationship) is an entity with many dimensions [34].

The most commonly used ML types for training approaches are supervised and unsupervised learning approaches. This approach illustrates an unsupervised learning technique in which unlabeled or novel examples are given, and there is no notion of the outcome. Such a technique aims to obtain several categories or groups, which will help us organize the information. In supervised learning, labelled data is used to estimate or map the model output[35]. The amount of labelling needed using the completely labelled data set can be reduced by around 30-40% when using our new techniques[39,40]. This limitation, however, can be addressed by Active Learning (AL), which learns incrementally through beginning with a few examples and then telling the medical expert to mark only the instances that the algorithm judges to be the most insightful in each iteration. The ML techniques are widely used in the medical industry. However, controlled instruction also involves grouping and regression. A classification method's critical problems include: Determining the number of groups, naming the various characteristics of the records, and learning from them.

This research establishes that each new sample is associated with one of the existing groups. Regression tasks are applied using the learning technique, which translates the raw data into a fundamental variable in the

model. For each new piece of research, this method may use to calculate the predictive variable.

However, classification and regression are two tasks of supervised learning. In the classification method, the learning mechanism assigns data to a finite number of groups, a new sample classified into one of the current groups using this method. While in a regression method, the mechanism converts data to a real variable. The predictive variable value for each new sample was calculated using this method [37]. A wide range of medical data requires better extraction and overall treatment methods, accurate diagnosis by GC through ML approaches, prediction of patients with GCs or lesions, and stability among domain expert knowledge in the associated areas and capability of professional data analysis and data processing that result in the best outcomes of GC diagnosis. Medical Decision Support systems combine ML and medical care as the fundamental technology of intelligent medicine, with enormous advantages for early prevention of different GCs and patient care [38].

2.8 The role of deep learning for GC diagnosis

The domain of DL offers an effective platform for supervised learning. Standard ML models such as SVM, ID3, KNN, and Naive Bayes have shallow architectures, whereas other linear and logistic regressions do not. DL is changing this. DL is a step above artificial neural networks as it incorporates additional layers that allow higher levels of abstraction and more accurate predictions from data [39].The DL model train using a variety of approaches and algorithms. Thus, the DL architecture is a multilayer stack of basic modules subject to learning; many apply computational models with nonlinear input-output mappings. At each level of the pile, each module also expands the range of inputs. Its purpose is to maximize the representation selectivity and invariance. A deep network may reflect functions of increasing complexity by adding more layers or additional units that can be added inside [40]—DL techniques used in the medical industry in some applications. DL is a type of ML, so we first introduce the basic concepts of ML. ML is one approach to data analysis that detects patterns in the data and then uses these patterns to predict future outcomes. When considering medical datasets, two main types of ML techniques: the first one has supervised learning, and the second one is unsupervised learning [41].

The first approach to ML is the supervised learning technique, which involves mapping from inputs to outputs based on labels assigned to the input-output pairs. This mapping job is characterized as a classification task when the findings are represented as categorical data. For example, a supervised learning challenge is when a machine learning agent learns to differentiate between normal and tumour tissues on pathologic slides based on labels such as "normal" or "tumour. "On the other hand, if the results are real scalars, this is known as a problem of regression[42]. The second approach to ML is unsupervised learning, which involves learning without particular labels. Clustering is one of the most common examples of unsupervised learning. Take a closer look at a circumstance in which thousands of nuclear photos were obtained from diverse cells on histopathological slides. These pictures can be grouped automatically into a set number of groups based on the similarity measurement of the ML agent. Because this activity does not contain a unique title, it is regarded as unsupervised learning. Taken together, two sample forms of ML algorithms allow medical data sets to be analyzed: supervised and unsupervised. Which approach is more appropriate is determined by the types of questions asked and the various properties of the data.[43]. AI-based deep learning is particularly well-suited for examining structured data and addressing the classification problem associated with structured data. DL is a field of ML. It is a mathematical tool utilized in research applications such as healthcare image analysis, object detection, speech analysis, and natural language processing [44]. DL has gained significant interest over the last six years as computational power has risen, system expenses have dropped, and many new datasets have been generated. DL algorithms are particularly effective in diagnosing and classifying GC and its many subtypes and segmenting tumours. DL techniques can provide superior information regarding specific types of cancers, their symptoms, locations,

Method	Benefit	Drawback
ML	These algorithms are often straightforward to implement. Algorithms are sufficiently adaptive to complex situations involving a large number of interdependent variables. Variations in the input and output may seen.	In high-dimensional databases, complex relationships between dependent and independent variables are challenging to determine. Its computational cost is very high.
DL	In high-dimensional databases, complex interactions between dependent and independent variables are effectively-recognized. It has the potential to maintain databases with a high level of noise.	Both the input and output are identical. The risk of an overfitting problem is very significant. When compared to machine learning, implementation is more complex. The training phase requires much more computing power than machine learning.

Table 3: Benefits & Drawbacks of ML and DL.

stages, aggressiveness, and metastases. Physicians can benefit from DL approaches by providing supplementary thoughts and identifying regions connected to pictures. Furthermore, a single DL model has been demonstrated to be helpful in diagnosis when compared to conventional medical procedures[45]. Table 3 shows some of the benefits and drawbacks of some commonly mentioned techniques, ML and DL [46].

One of the primary problems of ML is predicting results from the new data. Frequency analysis of the data performed using training data. A training data set is a set of input variables; corresponding outcomes are selected at random for each case, including positive and negative examples. Often, a subset of the initial unlabeled data set, referred to as the testing set, is used to evaluate the model consistency [47]. We are subject to estimation faults during the training phase. In other words, if supervised learning determines how well it has performed, the smaller the gap between training and testing errors, the better the algorithm will perform on testing data. These two principles describe the problems of overfitting and under-fitting in ML, and both trigger algorithm output reduction. If supervised learning incorrectly models the underlying dataset random fluctuations, it may become overly focused on the latter rather than describing it as an error. Under-fitting occurs when the algorithm fails to identify a feasible solution because the observed data does not explain consciousness. Thus, the least amount of overfitting and under-fitting occurs where a model encompasses both the training and testing sets since the training and testing data are distinct but have the same underlying distribution. As it stands, the method capacity (for fitting the training samples) is close to 100% in this case[48]—criteria for preventing and discovering the pitfall of overfitting. The most popular approach to overfitting a model is to restrict its complexity rather than minimize the number of features. The most common technique of under-fitting is to maximize the number of features. At around the same time, the number of samples used in the training and testing of a model should be of adequate size to obtain reliable results. What is done but is not always practical. In the third stage, multiple sources for evaluating the algorithm performance on a large and smaller set of data are outlined[49].

The development of Electronic Health Records (EHRs), which are permanent records of individual health records, is a priority for global healthcare systems. The exponential growth in the quantity of digitized clinical data has facilitated the development of data-driven healthcare, which integrates intelligent data analytics to improve decision-making and individualized treatment based on complex, diverse, and restricted data. The usage of ML and DL is required for thoughtful data analysis. ML combines methods from AI that allow the machine to learn from data to uncover complex and hidden patterns automatically. In the healthcare system, ML models are currently being used to evaluate data. In general, predictive analytics makes use of supervised ML algorithms. These strategies either categorize data into discrete categories (classification methods) or predict value (regression methodology)[50][51]. The DL approach progressively extracts higher-level information from the input image using several processing layers of linear and nonlinear transformations. The majority of DL techniques are built on a neural network framework. The word "deep" is categorized and frequently used to indicate several layers buried inside neural networks. DL methods learn features directly from the data, eliminating the need for human feature extraction. DL has been used for medical images only in the healthcare environment. Nevertheless, novel applications for EHR and bio-signal research have emerged lately.

AI undoubtedly has a wide range of applications in clinical practice. Using a range of clinical diversity characteristics and the current lack of objectivity and universality in expert systems improves patient usability while helping with existing subjectivity and objectivity problems[52]. In addition, ML may assist healthcare institutions in educating young physicians about clinical diagnosis and decision-making. A rising number of research publications show the outstanding diagnostic and prognostic performance of ML-based computer systems. In particular, DL algorithms are revolutionizing our capacity to analyses imaging data. These findings may increase sensitivity and guarantee that radiologists will have fewer false positives. They do, however, risk overfitting the training data, leading to brittle performance degradation in some scenarios. As a result, ML often entails a trade-off between accuracy and intelligibility. More precise models, like boosted trees, random forests, and neural networks, are often incomprehensible understandable models, such as logistic regression, Naive-Bayes, and single decision trees, often perform worse [53].

3 Related work

The author[54] summarizes the epidemiology and management of GC and gastroesophageal junction cancers (GEJC) and estimates their global economic and humanistic burdens. GC has a significant impact on patient health since it is associated with severe symptoms, a shortage of effective treatments, and economic costs that are expected to continue rising worldwide over the next decade. Nevertheless, there is still significant room for advancement in early detection and intervention and the discovery of new life-extending medications. In addition, predictive biomarkers may be used to identify patients who are most likely to benefit from a particular medicine, which can help maximize treatment efficiency and patient success. Unfortunately, although several studies have examined the epidemiology and treatment of GC and gastroesophageal junction cancers (GEJC), no worldwide estimates of their economic impact have been published to our knowledge.

In this study, the author [59] highlights how ML may support cancer detection and therapy via supervised, unsupervised, and DL techniques. Current technology approaches are grouped under one cluster for accuracy, sensitivity, specificity, and fake positive metrics compared with benchmark data sets.

In this paper[55], the author proposed a prediction method for survival, distant metastases, and peritoneal

metastases in GC using Gaussian Naive Bayes (GNB), XGBoost, and random forest algorithms. The study has observed the most successful models in OS prediction of distant metastases. Further, the peritoneal metastases were identified to be GNB with 81% accuracy.

In this study, the author [56] discusses how ML with supervised, unsupervised, and DL techniques can help with cancer diagnosis and treatment. Many state-of-the-art methods are classified under the same cluster for the accuracy, sensitivity, specificity, and false-positive metrics and the results compared to benchmark datasets. This study also examines, categorizes, and exposes current limits on methods for distinct forms of cancer. In addition, several obstacles to prospective future work. The primary goal of this study is to provide new researchers with an intellectual foundation for them to begin their research in medicine. The challenges in cancer detection and treatment are redesigning the research pipeline, understanding cancer development phenomena, developing preclinical models, accurately managing complex cures, treating earlier, developing and delivering innovative clinical trial methods, and improving trial accuracy.

According to the author [57], many classifiers have been used in cancer diagnosis, with a decision tree, neural network, and support vector machine outperforming others. While these classifiers have been the most accurate at predicting, their findings may include datasets that differ. In real-time, the gastric cancer detection process requires high prediction accuracy with less training time. Also, adopting optimization algorithms for improving the network architecture based on the dataset availability can improve the detection rate.

In this research, the author [58] states that due to the almost total absence of symptoms in the early stages of GC, it is difficult to diagnose the exact form of cancer in the earliest accounts. Endoscopy is a very accurate and precise diagnostic technique. While the processes perform under the supervision of a physician, malignant patches may be omitted or not diagnosed effectively. Due to the inability to completely identify the malignant spot, cancer may reappear after invasive surgery. To reduce this issue, a computerized decision support system, CDS, was developed with the help of expert physicians and image processing techniques. The CDS approach adopted in this study serves as a guide for gastrointestinal physicians, guiding them to identify cancerous patches in endoscopy images of the framework, collecting samples from these spots, and providing a more accurate diagnosis. The region will be determined with the use of biopsy samples taken from the patient. As a result, it is regarded as a model. Thus, this study would have prevented patients' mental health from deteriorating, as well as the complications connected with multiple biopsies and the resultant loss of faith in clinicians.

Even though the decrease in the prevalence of GC, the author [59]claims that it is still a fatal disease, with more than a quarter of a million individuals dying each year. H.pylori is the primary cause, accounting for 60–70% of all cases. In a clinical investigation done in China, the incidence of GC reduced up to 39% following h. Pylori eradication during 15 years. Combining h. pylori eradication with endoscopic screening has shown promise in GC incidence in high-risk groups. In Japan, this risk-based approach to GC prevention can pave the way for eliminating this, especially deadly illnesses.

In this study, the author [60] explains recent advances in GC prevention measures and recommends international cooperation ability to achieve population-based helicobacter pylori treatment programs, as the most evidence-based strategy currently available for GC prevention, in the context of demonstration projects in selected populations, to increase. In order to reduce the tremendous loss of life and productivity caused by this avoidable cancer, we must act quickly. To substantially decrease the enormous loss of life and productivity caused by this avoidable malignancy, researchers propose a nationally coordinated effort to adopt a population-based h.pylori treatment program, the most robust evidence-based approach presently available for GC prevention, in the context of demonstration projects in selected groups that will be scaled up later.

4 Nature of cancer and its impact

Before implementing an application, let get some background information on the classification, grouping, and impact of cancer. Cancer is caused by the growth of abnormal cells in the human body. It is generally a tumour type characterized by its size, shape, type of tumour, pattern of growth, and location [61]. Tumours are therefore divided into three different categories: benign, premalignant, and malignant. The term "benign tumour" refers to tumours that pose no danger to the patient's health (i.e., they do not cause an invasion of the surrounding tissue and do not spread to any other part of the human body). Although premalignant tumours are not yet cancerous, it has been shown that cancer can develop and form tumours. This tumour wants close surveillance of the patient. Malignant tumours, in the end, primarily affect the human body. In the case of malignant patients, the tumours rapidly split and spread over remote sites, thereby inducing metastases and other organ infestations [62]. Various therapies, including surgery, immunotherapy, and radiation therapy, are used because of the unparalleled complexity of cancer. Surgical procedures are frequently performed on patients to prevent tumours from spreading further into the human body. Chemotherapy is a method of diagnosing and treating hormone-related tumours.

4.1 Contribution of data mining in cancer domain

In the field of the medical domain, healthcare-associated DM is one of the most valuable and demanding applications of KDD. The problem is by data sets consisting of volume, value, velocity, variety, and veracity. Furthermore, the medical records data set is stored and distributed in multiple locations to integrate various sources. Moreover, the other issue faced by data miners is political, legal, and social problems with sensitive medical data. The reality is that data analysts do

not have domain knowledge. Therefore, it requires active collaboration between domain experts and data miners[63]. There are various DM techniques, particularly those used to identify issues in the medical domain, and the problems that people face are as follows. When speaking with different people, the term DM takes on a new meaning. However, large amounts of data to predict future events are analyzed as the true basic definition. DM currently plays a significant role in the health industry by improving the efficiency of the health care industry. DM addresses several real-life issues at the moment. Because raw data is mainly transformed into more meaningful information by the DM technique[3], criticize the lack of DM compliance with all statistical requirements[64]. For example, many data extraction tools use the same sample of training and testing[65]. DM may remain a medical resource, given this criticism.DM may help physicians identify cancer by giving them a better perspective of the disorder and generating a wealth of information that may be examined in many medical areas, where vast quantities of data may overlap[65]. The statistical accuracy of the models included is just one indication of the significance of DM in the medical field.

4.1.1 Heterogeneity of data

In the field of the medical domain, the data set contains dissimilarities in its data type and is very large in volume. Images, patient interviews, physician notes, and explanations are all used to collect clinical data. These factors are critical for DM scholars, as the doctor identifies and guesses the best treatment for patients. Physicians use unstructured, data-free texts in English, images, and other clinical data for software normalization and processing. In contrast to other branches of science, the underlying medical data set. Because medical data lacks a formal structure, the research organizes the information gathering [63].

4.1.2 Moral and social issues

Because human data is involved in medical data mining, they have related issues, as there may be some legal, ethical, and social concerns. Prevention of patient data and sensitive information handled very carefully. The range of available human medical data knowledge for DM is enormous. Data ownership can stop attempts to acquire the necessary data or to link other data sets. There are serious issues concerning the possession of data by patients, including ongoing high-profile hearings and court investigations. Concerns about privacy and security are another distinguishing feature of medical data, particularly in diabetes. The physician has easy access to the patient's data; there is no need for such data to be published. When transferring data to other servers, data security is a concern[63].

4.2 Operational challenges in the cancer domain

A cancer forecasting support system enables doctors to make appropriate, accurate, and timely decisions that reduce overall treatment costs. For the conduct of experiments, different classifiers were used. The KNN classifier provides the most predictive variables with the highest classification accuracy. The summary results showed that the Decision Tree Inducing methodology (C5) had an accuracy of 93.6%. The second most efficient model to be used with precision to classify it is 91.2%. The worst-case logistic regression model has a classification accuracy of 89.2%[64]. With so many risk factors linked with heart disease, a solid model is required to estimate cancer's likelihood. The time of the GC is precious for the heart. The correct risk assessment can save the lives of many patients[109][65]. ML algorithms have played a critical role in categorizing cervical cancer to diagnose its early[4]. ML is now one of the most promising and fast-growing sectors for medical data diagnostics. The different ML algorithms, the functions, data sets, and the exactness used. Track the process of correlation. Moreover, to predict the depth of cervical cancer through machine learning [4].

In the last decade, physicians could not say which of their patients could lead to cancer GC. Most patients do not know how long they will have any abnormalities in their stomachs. Doctors make decisions based on their experience and skills, not on the rich data but the patient database. This practice induces undesirable prejudices, mistakes, and high medical costs that affect the quality of patient service. The integration of clinical decision support (CDSS) and the patient database might lead to intelligent decision-making. In this context, DM is a fantastic approach to making better clinical decisions [66]. Early diagnosis of cancer prolongs human lives and is vital in fighting against GC. Medical imaging data is another factor in the early detection of GC diagnosis. Despite the increase in medical imaging data, the interpretation of the data concerning or compared to the speed of progression of cancer GC is time-consuming and difficult. In addition, if physicians misinterpret data in detection GCs, this will decline the accuracy rate sharply. ML is a sub-branch of artificial intelligence, widely used in medical image processing for cancer detection, classification, and tumour segmentation diagnosis [67]. Another prominent application of ML in the healthcare field is cancer, such as breast cancer, heart cancer, cervical cancer, and tuberculosis cancer. It is often done by applying ML methods to images of different organs or tissues suspected [56]. ML models have been proposed in multiple research works in the literature to detect cancer based on tissue images due to their success with image classification problems in general.

4.3 Limitations of data mining in healthcare sector

Although the DM provides information and support to paramedic staff by identifying patterns hidden in the dataset, the DM capabilities remain limited. It is worth noting that not all hidden patterns are stored in a dataset and can only use the DM technique. It should be reasonable and feasible to make the pattern interesting. DM is, therefore, for manual intervention to benefit from

the knowledge extracted. DM, for instance, could help diagnose, prescribe or replenish the intuition and skills of the doctor [68]. This approach argued that medical data is often composed of heterogeneous variables such as ethnicity, a family history of cancer, medicines, allergic responses, metabolic problems, and imaging tests. Each gives a partial representation of a patient's health. Moreover, the statistical properties of the above outlets are intrinsically distinct. As many scholars and doctors analyze this data, they address two challenges: the curse of dimensionality (the feature space grows exponentially in terms of dimension and sample size) and the variability of the feature sources and their statistical properties. These factors contribute to cancer GC diagnosis delays and inconsistencies, preventing patients from receiving timely treatment [110]. Therefore, there is a strong need for a comprehensive approach that enables early cancer GC diagnosis and can be utilized as a physician's decision. As a result, physicians and advanced statistical domain experts are overwhelmed with the dilemma of identifying innovative techniques for predicting cancer GC prognosis and diagnosis, as existing paradigms are incapable of handling any of this knowledge. This prerequisite is intimately associated with innovations in other sectors, such as BD, DM, and AI [10]. While an increasing amount of information is available in today's healthcare sector, doctors and nursing staff have difficulties performing time-consuming manual data analysis to make the best medical decisions while reducing uncertainty, patient risks, and costs. It has the unintended effect of resulting in substandard patient care, which is unacceptable. According to one study, 44,000 to 98,000 people die in the United States each year due to preventable medical errors (which account for 2–4% of all medical fatalities). Separate research found that up to 40% of patients were not receiving the appropriate treatment for chronic or acute illnesses [26].

5 Result analysis and discussion of the finding

The author [69] explains the four ML algorithms to identify GC patients in their study. Their research sought to find patterns using customized ML algorithms to assess gastrointestinal problems and mobile algorithms to predict GC cancers. The findings showed that 95% of the improved performance in the data set through an algorithm. In their studies [5] show that hybrid ML models may improve sensitivity and general accuracy. The neural network has categorized cancer cells with 100 percent sensitivity and 99.66 percent accuracy.

This study [70] was used to evaluate medical information using hybrid ML algorithms. After addressing the constraints of prior architectures, they have implemented a new computational intelligence architecture. This new computing architecture utilizes the SVM to carry on the medical data set. The researchers also pointed out, when we integrated it with the feature selection algorithm, the Genetic Algorithm (GA), the SVM outperformed the other classification algorithms by attaining greater accuracy and sensitivity.

The author's goal in this study[71]is to address some hybrid approaches, specifically DM and optimization techniques. The author explores and applies several strategies for classification and prediction to diagnose heart GC in earlier stages. It also builds up the structure for decision-making and prediction. According to the author of this study[72], cervical cancer is the fourth leading cause of death in women. In this paper, the author suggests three approaches—the SVM-based approach for the diagnosis of cervical cancer. The second author proposed two improved SVM approaches for diagnosing cancer samples: machine-recursive removal vector and support machine-main component analysis (SVM-PCA).

This paper [73]tested several hybrid ML models and discovered that the SVM hybrid model and simulated annealing provided 96% predictive precision when classifying hepatitis patients. These previous studies inspired the following studies to implement several hybrid combinations of ML algorithms, such as classification algorithms integrated with feature selection algorithms, and then use error optimization algorithms to perform hyper-parameter tuning. Removing unnecessary data from a dataset improves prediction model performance and avoids algorithm misdirection.[74].It is evident from Table 5 that several studies have been done in the past to find the DM technique and its use in various types of cancer, such as GC, breast, heart, GC, and cervical. DM techniques like SVM (Support Vector Machine), KNN, Decision Tree, and Naive Bayes have shown the most significant results in terms of accuracy when compared to other techniques. GC and different DM methods have been the subject of extensive study efforts in the last several decades. Many new ML techniques, including Artificial Neural Network (ANN), Bayesian networks (BN), Random Forest (RF), Support Vector Machines (SVM), Decision Trees (DT), and multilayer perceptron (MLP). Although ML algorithms have been widely used in GC and ultimately delivered high classifications, an appropriate level of validation is required in daily clinical and practice to take these methods into account.

Globally, GC continues to be a leading cause of cancer death, with a high death rate, attributable to the fact that the vast majority of GC patients are at an advanced stage of cancer, and prognoses are bleak treatment preferences are minimal. Certain types of cancer, such as GC, are hard to identify early on owing to their non-specific symptoms and ambiguous tell-tale that are difficult to identify at an early stage. As a result, improved prediction models based on multivariate data and high-resolution diagnostic tools are vital in clinical cancer research. Because of a large, curved organ with blind spots in the stomach, it is pretty challenging to inspect the whole stomach[75] thoroughly. If AI can identify the anatomical parts of the stomach, it may be sure that the whole stomach has been thoroughly checked. Because GC occurs in individuals with chronic gastritis, certain EGC seem to be similar to gastritis and are thus difficult to distinguish. The prognosis of GC varies according to the stage at the diagnosis; the prognosis is terrible when detected at an advanced stage. Nevertheless, the 5-year survival rate for early gastric cancer approaches 90 percent. A gastric

cancer diagnosis is an interdisciplinary research method that is identical to the endoscope [76]. Each year, about 80 million deaths occur due to misdiagnosis of cancer. A large number of cases and a patient's short prior medical history have resulted in fatal errors committed. These factors do not affect ML.ML algorithms can anticipate and diagnose cancers at a quicker rate in the healthcare industry when compared to medical experts, which is a significant advantage. The significance of early diagnosis and prognosis in increasing the survival rate of GC patients has been well documented [77]. Aside from that, the tumour in the digestive system in GC patients directly affects digestion and nutrient absorption and the adverse effects associated with chemical treatment of the GI tract. GC may be severe if not treated immediately. The reduced gastric digesting function may harm the nutritional condition of patients. Cancer is a chronic disease that affects the start site and can spread to other sites, resulting in a cascade of adverse effects on the patient's health and nutritional status. The presence of malnutrition as the first symptom indicates the occurrence of this cancer. It has a significant impact on cancer patients' nutritional and health conditions due to the adverse effects of cancer treatment[78][79]. The integration of genetic and histological components has brought progress in our knowledge of gastric cancer at a pathological level. One of the areas of pathological diagnosis might greatly benefit from methods that support DL integration. Many patients are hopeful as to how their diagnosis will be affected according to the present mood. DL may help increase the range of treatment choices and methods. In the world of cancer, having more detailed knowledge about the particular causes of stomach cancer provides it with an advantage over other forms of cancer since the origins of these cancers are not known fully or decoded. [80][81].

Many technical difficulties are still to be faced before AI can significantly impact the medical profession. It is essential to ensure that the training data has adequate high-quality data since these approaches rely heavily on massive amounts of high-quality data. Some diverse healthcare systems may collect data with biases and noise, which can negatively affect a model's training to perform well in one environment, but it does not work well elsewhere [111]. When diagnostic tasks reveal poor inter-expert agreement, machine learning models trained on the data may enhance their performance. Comprehensive data curation is needed to handle a range of data sources; extensive data curation is necessary. Comprehensive data curation is required. Many successful machine-learning models cannot be understood by those who are not involved in their development. These computer models can perform better than humans, is difficult to articulate the concepts in the models, pinpoint weaknesses in the models, or uncover new biological insights when analyzing these computational "black boxes" [82][83]. The healthcare data collection, storage, and sharing problem persist with electronic health records (EHRs). Safe data exchange via cloud services is possible with privacy-preserving techniques (such as third-party-hosted computing environments). The development of interoperable apps that satisfy the standard for clinical information is needed to make this infrastructure widely available. There is limited, sluggish, and inconsistent integration of health-related information across healthcare apps and locales [84].

To sum up, the research on AI in stomach cancer in the current context focuses on diagnosing cancer. At this point, the real benefits of AI are more impressive than broadly applicable therapeutic benefits. It may explain why there are so few AI researchers in the medical field. It may be feasible to extend the fundamental concepts of minimally invasive laparoscopic surgery (as they now exist) and apply them to the development of new computer-aided techniques, such as AI. Clinical issues that we see approaching are, on the surface, quite similar to AI-based problems [85]. Access to global medical resources may slow the adoption of AI in well-resourced areas, but more sufficient resources may have advanced the development of AI faster. Clinicians with AI expertise may be the crucial figures as far as AI advancement is concerned. There is a sizeable degree of acceptability and feasibility to the negative effect of accident situations, which may reduce by AI-assisted regurgitation, digestion and distribution of standardized therapeutic administration. Medical professionals nowadays have limited time to keep up with the newest advances in the digital patient care system in today's healthcare system. As a result, health expenditures remain high; regrettably, a significant portion of the population does not have access to quality medical treatment[86].

It is common to see physicians making extensive use of clinical trial data throughout utilizing vast amounts of patient data to grow and enhance their practice.ML may lead to clinical discoveries by uncovering previously unnoticed patterns in large data sets. It is clear that AI is a part of this area, but training and cooperation between specialists in computer science experts and medical professions are even more critical. This new technology implementation in an affordable manner by the medical personnel before it affects patient care [84]. Cancer treatment may be affected by AI invention because it may impact many aspects of it. These include predicting, screening, understanding large data sets, and interpreting imaging tests in the clinic. Early detection of tumour targets in both healthy and high-risk populations affords the potential to find cancer before it spreads, providing patients with a chance of successful treatment and more rapid recovery for a cure. A rise in AI, ML, and DL is speeding up, and soon these advances will revolutionize cancer screening and diagnosis. While we are eager to use cutting-edge AI technology to enhance cancer prediction, we must also work to educate our cutting-edge AI technologies on the nature of cancer early on. While AI applications are currently restricted, the potential for AI to play a significant role in cancer early detection is enormous. It is possible to determine diagnosis, prognosis, and therapy response by extracting information from the results. [85][57].

TYPE	MACHINE LEARNING TECHNIQUE	REFERENCES	ACCURACY
BREAST CANCER	Neural Network(NN), DBN, and backward-propagation.	Abdel-Zaher and Eldeib (2016)[5]	99.68%
	Support Vector Machine, K-Nearest Neighbors, multilayer perceptron, Decision Trees, Random Forest, Logistic Regression, Adaptive boosting, Gradient Boosting.	Turgut et al. (2018)[87]	95%
	SVM,Decision Tree, C4.5,Naive Bayes, classification algorithms.	Pritom et al. (2016)[74]	76.26%
	Comparison of LR and RF.	R. Kannan et al.(2018)[88]	87%
	Decision Tree.	Rajesh Jangade et al. (2018)[89]	75.10%
	Genetic Algorithm(GA) application.	KaanUyar (2017)[90]	97.78%
HEART ATTACK	Random Forest, Decision Tree, and Naïve Bayes.	H. Benjamin et al.,(2018)[91]	81%
	DecisionTree(DT), Genetic Algorithm (GA), Artificial Neural Network (ANN), naive Bayes algorithms.	Hilal et al. (2017).[92]	69.5%
	Neural Network Naïve Bayes classifier, and J48 decision tree.	Noreen Akhtar et al. (2018)[93]	80%
	Principal Component Analysis, decision tree, and SVM.	Dey, A et al. (2016)[94]	70%
	(CANFIS) coactive neuro-fuzzy inference system, and GA.	Parthiban L et al. (2018)[95]	-
	Hybrid DM model.	Shrivastava A al (2016)[71] Zriqat I al (2016)[96]	99.0%
	Support Vector Machine technique discovered to be attractive with high accuracy in the model.	Assari R et al. (2017)[6]	84.33%
CERVICAL CANCER	Support Vector Machine (SVM) with Genetic Algorithm (GA).	Kalantari et al. (2017)[70]	97.88%
	(SVM) algorithms Support Vector Machine with PCA.	Wu and Zhou (2017)[72]	93.97%
	Radial kernel support vector classifier (SVM Radial), Bayesian Optimization, and GB Machines.	Nishio et al. (2018) [97]	-
	Grid Optimization algorithm and (SVM).	Zhao et al. (2018)[98]	85.56%
GASTRIC CANCER	Apriori, CN2 Rules, C4.5, and Naive Bayes (NB).	Mahmoodi et al. (2016)[99]	87.2%
	Logistic regression (LR), C5.0, Decision Tree (DT), multilayer perceptron (MLP), and tree augmented naive Bayesian network.	Liu, M et al. (2018)[100]	77.84%
	Different techniques such as support vector machine (SVM), decision tree (DT), naïve Bayesian model, and k-nearest neighbour used to find the closest neighbour in a dataset (KNN).	AsgharMortezagholi et al. (2019)[69]	90.08%
	Logistic Regression Algorithm, Genetic Algorithm & MICE Algorithm.	LadanGoshayeshi et al. (2017)[66]	72.57%
	K-Nearest Neighbour, XGBoost, and LightGBM.	Amirgaliyev Y et al. (2019)[101]	95%

Table 4: DM techniques & their uses in various cancer.

5.1 Applied excellence in machine learning and deep learning

Medical experts find ML a helpful asset in patient care, prevention, and identification of infectious GCs. Effective use of these strategies is almost entirely missing in the hospital environment. A more natural way of thinking about ML algorithms trained on a task and then learned to do that job. What is true if the training data set was manually selected and labelled under the supervision of someone who favoured the techniques, parameters, and processes used to construct the task. Through their interpretation, analysis, and optimization components, ML algorithms leverage the accessibility of massive quantities of data and higher computational architectures to represent more multivariate analysis processes than traditional approaches; DL facilitates the detection of previously hidden patterns, extrapolation of trends, and prediction of outcomes in a wide variety of problems, all while seeking to "learn". Now, ML algorithms are an initiation into the medical reports of patients. The aim is to determine, for example, which patients are more likely to need readmission to the hospital or who are unable to adhere to prescription medications. The applications in diagnosis, testing, drug development, and clinical trials are nearly limitless [102]. Despite the abundance of digitized evidence, predictive methodologies constructed from hospital records remain predominantly on a linear model and only have more than 20 or 30 variables. However, one tangible advantage of ML is that experts are not required to determine which parameters to consider

and in what variations. An essential aspect to remember when implementing ML in medical care is the accuracy of evidence from different sources. Each healthcare system can collect patient information in unique ways to attain mutual objectives. As a result, before implementing ML, it is necessary to match the results. Avoids data overfitting, which makes it more challenging to extend the methodology to other data sets. The issue of racism is often crucial. This challenge arises where training data coverage is inadequate and inaccurate when referring to minority groups. In particular, it is valuable in medical care to provide a variety of broad datasets to reflect the distinctive features of each group of patients. In general, having a variety of significant datasets in a hospital is desirable, intending to emphasize the distinguishing features of each patient group. Therefore, the algorithm's intelligibility is of great importance. It is essential to maintain a balance between performance and accessibility. Due to the (variable) complexity of higher-performance models (for example, DL), it is much more difficult to identify them. By contrast, models (regression models or decision trees) are considerably more precisely defined[10]. Finally, another area where ML has shown great promise is in the field of the healthcare sector. Modern medical organizations frequently use EHR due to the widespread use of electronic health records (and many heterogeneous data components) [103]. It includes patient demographics, diagnoses, laboratory test results, medicines patients have already taken, and medical records. In addition to medical imaging, sensor, and text data, patient data includes imaging, sensor, and text data[103].

Although it was previously believed that having access to more information on each patient would lead to better-informed medical choices, this has yet to be proven true. While this remark highlights that medical practitioners are constantly bombarded with information, it glosses over the massive amount of data available to medical specialists. The use of ML algorithms is highlighted in this part by applications of these techniques in several research areas, including predicting individual patient responses to cancer medicines, diabetes research, retinopathy detection, and cancer detection [10].

6 Future research and developmental challenges

Marketing and manufacturing are only two industries that have made extensive use of (ML). Its use in the healthcare sector is gaining traction. Difficulties such as irrelevant characteristics, uncertainty, computational difficulty, dynamic nature, and computational time are becoming more widely studied, thanks to the increase in recent research investigating the ML complexities. This section discusses potential research applications, such as personalized treatment, data loss during pre-processing, clinical data collection for scientific purposes, automation for junior expert users, collaborative study and domain expert experience, integration into the healthcare sector, and prediction-specific to DM application and integration with the healthcare system. There are some challenges mentioned below for the diagnosis of GC using the ML technique [104].

6.1 Information loss during data pre-processing

Information gathering or data pre-processing is the most time-consuming & cost-effective component in the DM. Missing values accounted for roughly 46.5 percent of the data and 363 out of 410 attributes in one study. We lose a large amount of information when we filter out missing value instances and outliers. Future studies should develop a more accurate way of determining missing values rather than when they were removed. Additionally, new or modified data gathering procedures were modified to circumvent this problem. If there is any missing data, a strategy is to deal with outliers. However, as shown in one of the studies we reviewed, outliers may be utilized to learn uncommon disorders. Instead of leaving out the oddities, future studies should seek them out and discover what they can teach[105].

6.2 DM process automation for junior expert users

Physicians, nurses, and other paramedic staff with insufficient data analytics knowledge or training are the beneficiaries of DM in the healthcare sector. One way to address this issue is to establish an automated (i.e., not supervised by humans) system for end-users. A cloud-based framework for preventing medical errors was also designed, although the work would be difficult due to the variety of application areas and that no single algorithm would be equally accurate for all applications [30].

6.3 The study's interdisciplinary approach and domain technical expertise

Health informatics is a field of study that encompasses multiple disciplines. In certain domains of healthcare issues (for example, oncologists for cancer studies), DM is used in conjunction with expert opinion. Approximately 32% of publications in analytics did not include any professional guidance. Deeper analysis should involve representatives from a variety of fields, including healthcare [106].

6.4 Incorporated into the healthcare sector

Several articles reviewed attempted to incorporate the DM technique into the decision-making mechanism itself. The effect of knowledge discovery via DM on the workload and time of medical practitioners is questionable. Future research should investigate system integration and its influence on professional environments[103]. The research findings provide valuable information on GC patients' nutritional state. As a result, clinicians and nutritionists have a solid foundation to build treatments that are helpful to patients. Patients suffering from nausea and vomiting due to their cancer treatment may compromise their nutritional needs. When attempting to improve a patient's general health, especially those with

GC, it is critical to ensure that the patient consumes food before, during, and after treatment. In light of these facts, scientists conducted a study to see whether there was a link between dietary circumstances and GC patients' quality of life. The nutritional status of GC patient survivors impacts their quality of life and correlates strongly with obesity, overweight, and underweight. Quality of life is significantly impacting on receiving regular medication, and patients should thus get it. While patients with eating disorders are identified, they also get dietary guidance to help them develop appropriate treatment plans [107]. As a further point, there are so many challenges that occur during the diagnosis of GC. These issues are possibly the challenging obstacles to resolve which raised. Although it is still too early to establish general guidelines to deal with this challenge, I believe it is feasible to concentrate on these topics; I want future researchers to consider the suggestions. As a result, this is an open area for future researchers to explore.

- It identifies the critical risk factors, such as nutrients, that contribute to the development of GC.
- What are the interventions? How can we determine the interventions for the factors that contribute to GC?
- To obtain an accurate prognosis, use different ML techniques to detect the presence of GC.

7 Conclusion

The role of BD in the healthcare industry is to empower us to develop more detailed health profiles of patients and predictive models for individuals, to improve patient care and prognosis in GC. In this paper, we cover BD, emphasizing its use in the healthcare sector. Additionally, the primary challenge with BD in healthcare is making the data simple to interpret, which is beneficial for medical practitioners since it is a tool for detecting significant patterns in complex data. ML offers a solution to this challenge. The main focus of big data analytics in healthcare is integrating and analyzing massive volumes of complex and heterogeneous data from various sources, including biomedical data and electronic health information data. ML assists physicians in diagnosing effectively and predicting prognostic outcomes using multiple techniques that are beneficial to the patient's health with the support of ML. Furthermore, ML promotes the integration of models by utilizing various algorithms to support and predict factors with perfect precision.

The rapid implementation in the medical industry and the use of BD are still challenging tasks for researchers in this domain. This article discusses the ML, DL, and BD paradigms and the different applications and their associated possible limitations, challenges, advantages, and drawbacks in GC. In addition, this research incorporates the KDD process, which shows how medical practitioners will extract knowledge from datasets and integrate (DSS) into the medical data set, which will assist in producing outcomes utilizing DM algorithms. This paper also gives information about DM's classification algorithms that will be implemented for various cancers.

References

[1] M. Mahmood, B. Al-Khateeb, and W. M. Alwash, "A review on neural networks approach on classifying cancers," Int J Artif Intell, vol. 9, no. 2, pp. 317–326, 2020. https://doi.org/10.11591/ijai.v9.i2.pp317-326

[2] A. M. Brushfield, T. T. Luu, B. D. Callahan, and P. E. Gilbert, "A comparison of discrimination and reversal learning for olfactory and visual stimuli in aged rats.," Behav. Neurosci., vol. 122, no. 1, p. 54, 2008. https://doi.org/10.1037/0735-7044.122.1.54

[3] R. A. Smith et al., "Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening," CA. Cancer J. Clin., vol. 69, no. 3, pp. 184–210, 2019. https://doi.org/10.3322/caac.21557

[4] A. Shetty and V. Shah, "Survey of Cervical Cancer Prediction Using Machine Learning: A Comparative Approach," in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1–6. https://doi.org/10.1109/icccnt.2018.8494169

[5] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," Expert Syst. Appl., vol. 46, pp. 139–144, 2016. https://doi.org/10.1016/j.eswa.2015.10.015

[6] R. Assari, P. Azimi, and M. R. Taghva, "Heart Disease Diagnosis Using Data Mining Techniques," Int. J. Econ. Manag. Sci., vol. 6, no. 3, 2017. https://doi.org/10.4172/2162-6359.1000415

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., vol. 17, no. 3, p. 37, 1996. https://doi.org/10.1145/240455.240464

[8] P. Bertuccio et al., "Citrus fruit intake and gastric cancer: The stomach cancer pooling (StoP) project consortium," Int. J. cancer, vol. 144, no. 12, pp. 2936–2944, 2019. https://doi.org/10.1002/ijc.32046

[9] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, "The role of artificial intelligence in healthcare: a structured literature review," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–23, 2021. https://doi.org/10.1186/s12911-021-01488-9

[10] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Med., vol. 17, no. 1, pp. 1–9, 2019. https://doi.org/10.1186/s12916-019-1426-2

[11] S. M. Jameel, M. A. Hashmani, M. Rehman, and A. Budiman, "An adaptive deep learning framework for dynamic image classification in the internet of things environment," Sensors, vol. 20, no. 20, p. 5811, 2020.

[12] S. van Baalen, M. Boon, and P. Verhoef, "From clinical decision support to clinical reasoning support systems.," Authorea Prepr., 2020. https://doi.org/10.22541/au.159986468.80473725.

[13] S. Akundi, R. Soujanya, and P. M. Madhuri, "Big Data Analytics in Healthcare Using Machine Learning Algorithms: A Comparative Study," 2020. https://doi.org/10.3991/ijoe.v16i13.18609.

[14] M. A. Hashmani, S. M. Jameel, H. Al-Hussain, M. Rehman, and A. Budiman, "Accuracy performance degradation in image classification models due to concept drift," Int. J. Adv. Comput. Sci. Appl, vol. 10, 2019. https://doi.org/10.14569/ijacsa.2019.0100552.

[15] P. Acharya and M. Mathur, "Artificial intelligence in dermatology: the 'unsupervised'learning," Br. J. Dermatol., vol. 182, no. 6, pp. 1507–1508, 2020. https://doi.org/10.1111/bjd.18955

[16] T. Silwattananusarn and K. Tuamsuk, "Data mining and its applications for knowledge management: a literature review from 2007 to 2012," arXiv Prepr. arXiv1210.2872, 2012. https://doi.org/10.5121/ijdkp.2012.2502

[17] S. S. ZIA, P. AKHTAR, and T. J. A. MUGHAL, "Case Retrieval Process of CBR Technique Implements on Knowledge-Based Clinical Decision Support Systems (KBCDSS) for Diagnosis of Breast Cancer Disease," Sindh Univ. Res. Journal-SURJ (Science Ser., vol. 47, no. 2, 2015. https://doi.org/10.26692/sujo/2019.01.22

[18] A. Karahoca, Advances in data mining knowledge discovery and applications. BoD--Books on Demand, 2012. https://doi.org/10.5772/3349

[19] P. E. Beeler, D. W. Bates, and B. L. Hug, "Clinical decision support systems," Swiss Med. Wkly., vol. 144, p. w14073, 2014. https://doi.org/10.4414/smw.2014.14073

[20] C. Schuh, J. S. de Bruin, and W. Seeling, "Clinical decision support systems at the Vienna General Hospital using Arden Syntax: Design, implementation, and integration," Artif. Intell. Med., vol. 92, pp. 24–33, 2018. https://doi.org/10.1016/j.artmed.2015.11.002

[21] J. Ferlay et al., "Cancer statistics for the year 2020: An overview," Int. J. Cancer,2021 https://doi.org/10.1002/ijc.33588

[22] S. S. ZIA, P. Akhtar, and T. J. A. MUGHAL, "Schematic Cycle of Case-Based Reasoning Technique Implements in Clinical Decision Support Systems Used for Diagnosis of Liver Disease," Sindh Univ. Res. Journal-SURJ (Science Ser., vol. 47, no. 2, 2015.

[23] T. Lysaght, H. Y. Lim, V. Xafis, and K. Y. Ngiam, "AI-assisted decision-making in healthcare," Asian Bioeth. Rev., vol. 11, no. 3, pp. 299–314, 2019. https://doi.org/10.1007/s41649-019-00096-0

[24] H. C. Koh, G. Tan, and others, "Data mining applications in healthcare," J. Healthc. Inf. Manag., vol. 19, no. 2, p. 65, 2011. https://doi.org/10.1109/icdmw.2011.202

[25] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017. https://doi.org/10.1016/j.csbj.2016.12.005

[26] C. Neto, M. Brito, V. Lopes, H. Peixoto, A. Abelha, and J. Machado, "Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients," Entropy, vol. 21, no. 12, p. 1163, 2019, https://doi: 10.3390/e21121163. https://doi.org/10.3390/e21121163

[27] S. M. Jameel, M. A. Hashmani, H. Alhussain, M. Rehman, and A. Budiman, "An optimized deep convolutional neural network architecture for concept drifted image classification," in Proceedings of SAI Intelligent Systems Conference, 2019, pp. 932–942. https://doi.org/10.1007/978-3-030-29516-5_70

[28] F. M. Couto, Data and text processing for health and life sciences. Springer Nature, 2019. https://doi.org/10.1007/978-3-030-13845-5

[29] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," J. Glob. Health, vol. 8, no. 2, 2018. https://doi.org/10.7189/jogh.08.020303

[30] S. R. Kumar, N. Gayathri, S. Muthuramalingam, B. Balamurugan, C. Ramesh, and M. K. Nallakaruppan, "Medical big data mining and processing in e-healthcare," in Internet of Things in Biomedical Engineering, Elsevier, 2019, pp. 323–339. https://doi.org/10.1016/b978-0-12-817356-5.00016-4

[31] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high-dimensional databases: A review," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 3, p. e1300, 2019. https://doi.org/10.1002/widm.1300

[32] H. W. Ian and F. Eibe, "Data mining: Practical machine learning tools and techniques." Morgan Kaufmann Publishers, 2005. https://doi.org/10.1145/2020976.2021004

[33] L. Wang and C. A. Alexander, "Big data analytics in medical engineering and healthcare: methods, advances and challenges," J. Med. Eng. \& Technol., vol. 44, no. 6, pp. 267–283, 2020. https://doi.org/10.1080/03091902.2020.1769758

[34] I. H. Witten, E. Frank, and M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition. Morgan Kaufmann, 2017. https://doi.org/10.1016/b978-0-12-374856-0.00015-8

[35] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, 2015. https://doi.org/10.1016/j.csbj.2014.11.005

[36] N. Nissim et al., "Improving condition severity classification with an efficient active learning based framework," J. Biomed. Inform., vol. 61, pp. 44–54, 2016. https://doi.org/10.1016/j.jbi.2016.03.016

[37] N. Iqbal and M. Islam, "Machine learning for dengue outbreak prediction: A performance evaluation of

different prominent classifiers," Informatica, vol. 43, no. 3, 2019. https://doi.org/10.31449/inf.v43i3.1548

[38] N. Nissim, Y. Shahar, Y. Elovici, G. Hripcsak, and R. Moskovitch, "Inter-labeler and intra-labeler variability of condition severity classification models using active and passive learning methods," Artif. Intell. Med., vol. 81, pp. 12–32, 2017. https://doi.org/10.1016/j.artmed.2017.03.003

[39] S. K. Zhou et al., "A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises," arXiv Prepr. arXiv2008.09104, 2020. https://doi.org/10.1109/jproc.2021.3054390

[40] T. Lu, Y. Du, L. Ouyang, Q. Chen, and X. Wang, "Android malware detection based on a hybrid deep learning model," Secur. Commun. Networks, vol. 2020, 2020. https://doi.org/10.1155/2020/8863617

[41] T. J. Saleem and M. A. Chishti, "Exploring the applications of Machine Learning in Healthcare," Int. J. Sensors Wirel. Commun. Control, vol. 10, no. 4, pp. 458–472, 2020. https://doi.org/10.2174/2210327910666191220103417

[42] S. Mittal and Y. Hasija, "Applications of deep learning in healthcare and biomedicine," in Deep Learning Techniques for Biomedical and Health Informatics, Springer, 2020, pp. 57–77. https://doi.org/10.1007/978-3-030-33966-1_4

[43] Y.-W. Chen and L. C. Jain, "Deep Learning in Healthcare." Springer, 2020. https://doi.org/10.1155/2020/8863617

[44] S. Pitoglou, "Machine Learning in Healthcare: Introduction and Real-World Application Considerations," in Quality Assurance in the Era of Individualized Medicine, IGI Global, 2020, pp. 92–109. https://doi.org/10.4018/978-1-7998-2390-2.ch004

[45] A. Mustafa and M. Rahimi Azghadi, "Automated Machine Learning for Healthcare and Clinical Notes Analysis," Computers, vol. 10, no. 2, p. 24, 2021. https://doi.org/10.3390/computers10020024

[46] H. Wang and B. Raj, "A survey: Time travel in deep learning space: An introduction to deep learning models and how deep learning models evolved from the initial ideas," arXiv Prepr. arXiv1510.04781, 2015. https://doi.org/10.1007/978-1-4842-2766-4_14

[47] A. Crippa et al., "Use of machine learning to identify children with autism and their motor abnormalities," J. Autism Dev. Disord., vol. 45, no. 7, pp. 2146–2156, 2015. https://doi.org/10.1007/s10803-015-2379-8

[48] A. D. Gavrilov, A. Jordache, M. Vasdani, and J. Deng, "Preventing model overfitting and underfitting in convolutional neural networks," Int. J. Softw. Sci. Comput. Intell., vol. 10, no. 4, pp. 19–28, 2018. https://doi.org/10.4018/ijssci.2018100102

[49] P. Samui, Handbook of research on advanced computational techniques for simulation-based engineering. IGI Global, 2015. https://doi.org/10.4018/978-1-4666-9479-8

[50] X.-Y. Wang and J. M. Garibaldi, "Simulated annealing fuzzy clustering in cancer diagnosis," Informatica, vol. 29, no. 1, 2005.

[51] S. Sunarti, F. F. Rahman, M. Naufal, M. Risky, K. Febriyanto, and R. Masnina, "Artificial intelligence in healthcare: opportunities and risk for future," Gac. Sanit., vol. 35, pp. S67--S70, 2021. https://doi.org/10.1016/j.gaceta.2020.12.019

[52] M. Masmoudi, B. Jarboui, and P. Siarry, "Artificial Intelligence and Data Mining in Healthcare." Springer, 2020. https://doi.org/10.1007/978-3-030-45240-7

[53] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny, "A Review on Deep Learning Approaches in Healthcare Systems: Taxonomies, Challenges, and Open Issues," J. Biomed. Inform., p. 103627, 2020. https://doi.org/10.1016/j.jbi.2020.103627

[54] M. Casamayor, R. Morlock, H. Maeda, and J. Ajani, "Targeted literature review of the global burden of gastric cancer," Ecancermedicalscience, vol. 12, 2018. https://doi.org/10.3332/ecancer.2018.883

[55] M. Akcay, D. Etiz, and O. Celik, "Prediction of Survival and Recurrence Patterns by Machine Learning in Gastric Cancer Cases Undergoing Radiation Therapy and Chemotherapy," Adv. Radiat. Oncol., vol. 5, no. 6, pp. 1179–1187, 2020. https://doi.org/10.1016/j.adro.2020.07.007

[56] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges," J. Infect. Public Health, vol. 13, no. 9, pp. 1274–1289, 2020. https://doi.org/10.1016/j.jiph.2020.06.033

[57] S.-L. Zhu, J. Dong, C. Zhang, Y.-B. Huang, and W. Pan, "Application of machine learning in the diagnosis of gastric cancer based on noninvasive characteristics," PLoS One, vol. 15, no. 12, p. e0244869, 2020. https://doi.org/10.1371/journal.pone.0244869

[58] A. N. Richter and T. M. Khoshgoftaar, "A review of statistical and machine learning methods for modeling cancer risk using structured clinical data," Artif. Intell. Med., vol. 90, pp. 1–14, 2018. https://doi.org/10.1016/j.artmed.2018.06.002

[59] A. Yasar, I. Saritas, and H. Korkmaz, "Computer-aided diagnosis system for detection of stomach cancer with image processing techniques," J. Med. Syst., vol. 43, no. 4, pp. 1–11, 2019. https://doi.org/10.1007/s10916-019-1203-y

[60] J. Y. Park and R. Herrero, "Recent progress in gastric cancer prevention," Best Pract. \& Res. Clin. Gastroenterol., p. 101733, 2021. https://doi.org/10.1016/j.bpg.2021.101733

[61] A. Onasanya and M. Elshakankiri, "Smart integrated IoT healthcare system for cancer care," Wirel. Networks, pp. 1–16, 2019. https://doi.org/10.1007/s11276-018-01932-1

[62] M. D. Islam, W. A. Kaplan, D. Trachtenberg, R. Thrasher, K. P. Gallagher, and V. J. Wirtz, "Impacts of intellectual property provisions in trade treaties on access to medicine in low and middle income

countries: a systematic review," Global. Health, vol. 15, no. 1, p. 88, 2019. https://doi.org/10.1186/s12992-019-0528-0

[63] K. J. Cios, B. Krawczyk, J. Cios, and K. J. Staley, "Uniqueness of Medical Data Mining: How the new technologies and data they generate are transforming medicine," arXiv Prepr. arXiv1905.09203, 2019. https://doi.org/10.1016/s0933-3657(02)00049-0

[64] M. Kumari and V. Singh, "Breast cancer prediction system," Procedia Comput. Sci., vol. 132, pp. 371–376, 2018. https://doi.org/10.1016/j.procs.2018.05.197

[65] G. Purusothaman and P. Krishnakumari, "A survey of data mining techniques on risk prediction: Heart disease," Indian J. Sci. Technol., vol. 8, no. 12, p. 1, 2015. https://doi.org/10.17485/ijst/2015/v8i12/58385

[66] L. Goshayeshi et al., "Predictive model for survival in patients with gastric cancer," Electron. physician, vol. 9, no. 12, p. 6035, 2017. https://doi.org/10.19082/6035

[67] N. C. Caballé, J. L. Castillo-Sequera, J. A. Gómez-Pulido, and M. L. Polo-Luque, "Machine learning applied to diagnosis of human diseases: A systematic review," 2020. https://doi.org/10.3390/app10155135

[68] W. H. Organization and others, "Cancer. 2018," World Heal. Organ. Available http//www. who. int/mediacentre/factsheets/fs297/en, 2017. https://doi.org/10.23846/ow3.ie71

[69] A. Mortezagholi, O. Khosravizadehorcid, M. B. Menhaj, Y. Shafigh, and R. Kalhor, "Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: Using data mining method," Asian Pacific J. Cancer Prev., vol. 20, no. 9, pp. 2607–2610, 2019, doi: 10.31557/APJCP.2019.20.9.2607. https://doi.org/10.31557/apjcp.2019.20.9.2607

[70] A. Kalantari, A. Kamsin, S. Shamshirband, A. Gani, H. Alinejad-Rokny, and A. T. Chronopoulos, "Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions," Neurocomputing, vol. 276, pp. 2–22, 2018. https://doi.org/10.1016/j.neucom.2017.01.126

[71] A. Shrivastava and S. S. Tomar, "A hybrid framework for heart disease prediction: review and analysis," Int. J. Adv. Technol. Eng. Explor., vol. 3, no. 15, p. 21, 2016. https://doi.org/10.19101/ijatee.2016.315003

[72] W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," IEEE Access, vol. 5, pp. 25189–25195, 2017. https://doi.org/10.1109/access.2017.2763984

[73] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," Comput. Methods Programs Biomed., vol. 108, no. 2, pp. 570–579, 2012. https://doi.org/10.1016/j.cmpb.2011.08.003

[74] A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in 2016 19th International Conference on Computer and Information Technology (ICCIT), 2016, pp. 310–314. https://doi.org/10.1109/iccitechn.2016.7860215

[75] M. A. de Brito, C. Neto, A. Abelha, and J. Machado, "Prediction of mortality and occurrence of complications for gastric cancer patients," in 2019 International Conference in Engineering Applications (ICEA), 2019, pp. 1–6. https://doi.org/10.1109/ceap.2019.8883494

[76] P. Jin et al., "Artificial intelligence in gastric cancer: a systematic review," J. Cancer Res. Clin. Oncol., pp. 1–12, 2020. https://doi.org/10.1007/s00432-020-03304-9

[77] H. Nakashima, H. Kawahira, H. Kawachi, and N. Sakaki, "Artificial intelligence diagnosis of Helicobacter pylori infection using blue laser imaging-bright and linked color imaging: a single-center prospective study," Ann. Gastroenterol., vol. 31, no. 4, p. 462, 2018. https://doi.org/10.20524/aog.2018.0269

[78] M. Venerito, A. C. Ford, T. Rokkas, and P. Malfertheiner, "Prevention and management of gastric cancer," Helicobacter, vol. 25, p. e12740, 2020. https://doi.org/10.1111/hel.12740

[79] B. Kramer et al., "Long-term quality of life and nutritional status of patients with head and neck cancer," Nutr. Cancer, vol. 71, no. 3, pp. 424–437, 2019. https://doi.org/10.1080/01635581.2018.1506492

[80] K. Togashi, "Applications of artificial intelligence to endoscopy practice: The view from Japan Digestive Disease Week 2018." Wiley Online Library, 2019. https://doi.org/10.1111/den.13354

[81] S. Yalcin et al., "Nutritional aspect of cancer care in medical oncology patients," Clin. Ther., vol. 41, no. 11, pp. 2382–2396, 2019. https://doi.org/10.1016/j.clinthera.2019.09.006

[82] S. Vollmer et al., "Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness," bmj, vol. 368, 2020. https://doi.org/10.1136/bmj.m1312

[83] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," Cancer Lett., vol. 471, pp. 61–71, 2020. https://doi.org/10.1016/j.canlet.2019.12.007

[84] A. S. Ahuja, "The impact of artificial intelligence in medicine on the future role of the physician," PeerJ, vol. 7, p. e7702, 2019. https://doi.org/10.7717/peerj.7702

[85] T. M. Noguerol, F. Paulano-Godino, M. T. Mart\'\in-Valdivia, C. O. Menias, and A. Luna, "Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology," J. Am. Coll. Radiol., vol. 16, no. 9, pp. 1239–1247, 2019. https://doi.org/10.1016/j.jacr.2019.05.047

[86] S. Hamid, "The opportunities and risks of artificial intelligence in medicine and healthcare," 2016. https://doi.org/10.1201/b19187-4

[87] S. Turgut, M. Da\ugtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018, pp. 1–3. https://doi.org/10.1109/ebbt.2018.8391468

[88] R. Kannan and V. Vasanthi, "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease," in Soft Computing and Medical Bioinformatics, Springer, 2019, pp. 63–72. https://doi.org/10.1007/978-981-13-0059-2_8.

[89] R. Chauhan, R. Jangade, and R. Rekapally, "Classification model for prediction of heart disease," in Soft Computing: Theories and Applications, Springer, 2018, pp. 707–714. https://doi.org/10.1007/978-981-10-5699-4_67

[90] K. Uyar and A. \.Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," Procedia Comput. Sci., vol. 120, pp. 588–593, 2017. https://doi.org/10.1016/j.procs.2017.11.283

[91] H. David and S. A. Belcy, "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES.," ICTACT J. Soft Comput., vol. 9, no. 1, 2018. https://doi.org/10.21917/ijsc.2017.0202

[92] H. Almarabeh and E. Amer, "A study of data mining techniques accuracy for healthcare," Int. J. Comput. Appl., vol. 168, no. 3, pp. 12–17, 2017. https://doi.org/10.5120/ijca2017914338

[93] N. Akhtar, M. R. Talib, and N. Kanwal, "Data Mining Techniques to Construct a Model: Cardiac Diseases," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, 2018. https://doi.org/10.14569/ijacsa.2018.090173

[94] A. Dey, J. Singh, and N. Singh, "Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis," Int. J. Comput. Appl., vol. 140, no. 2, pp. 27–31, 2016. https://doi.org/10.5120/ijca2016909231
L. Parthiban and R. Subramanian, "Intelligent heart disease prediction system using CANFIS and genetic algorithm," Int. J. Biol. Biomed. Med. Sci., vol. 3, no. 3, 2008. https://doi.org/10.1109/iama.2009.5228016

[95] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," arXiv Prepr. arXiv1704.02799, 2017. https://doi.org/10.21884/ijmter.2017.4211.vxayk

[96] M. Nishio et al., "Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization," PLoS One, vol. 13, no. 4, p. e0195875, 2018. https://doi.org/10.1371/journal.pone.0195875

[97] Y. Zhao, Y. Liu, and W. Huang, "Prediction model of HBV reactivation in primary liver cancer.Based on NCA feature selection and SVM classifier with Bayesian and grid optimization," in 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, pp. 547–551. https://doi.org/10.1109/icccbda.2018.8386576

[98] S. A. Mahmoodi, K. Mirzaie, and S. M. Mahmoudi, "A new algorithm to extract hidden rules of gastric cancer data based on ontology," Springerplus, vol. 5, no. 1, p. 312, 2016. https://doi.org/10.1186/s40064-016-1943-9

[99] M.-M. Liu, L. Wen, Y.-J. Liu, Q. Cai, L.-T. Li, and Y.-M. Cai, "Application of data mining methods to improve screening for the risk of early gastric cancer," BMC Med. Inform. Decis. Mak., vol. 18, no. 5, p. 121, 2018. https://doi.org/10.1186/s12911-018-0689-4

[100] Y. Amirgaliyev, S. Shamiluulu, T. Merembayev, and D. Yedilkhan, "Using Machine Learning Algorithm for Diagnosis of Stomach Disorders," in International Conference on Mathematical Optimization Theory and Operations Research, 2019, pp. 343–355. https://doi.org/10.1007/978-3-030-33394-2_27

[101] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," NPJ Digit. Med., vol. 1, no. 1, pp. 1–10, 2018. https://doi.org/10.3410/f.733181042.793560090

[102] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (EHRs) A survey," ACM Comput. Surv., vol. 50, no. 6, pp. 1–40, 2018. https://doi.org/10.1145/3127881

[103] S. Shirazi, H. Baziyad, and H. Karimi, "An Application-Based Review of Recent Advances of Data Mining in Healthcare," J. Biostat. Epidemiol., 2019. https://doi.org/10.18502/jbe.v5i4.3864

[104] V. V Petrov, O. P. Mintser, A. A. Kryuchyn, and Y. A. Kryuchyna, "Big Data in medicine: promise and challenges," 2019. https://doi.org/10.11603/mie.1996-1960.2019.3.10429

[105] D. Cirillo and A. Valencia, "Big data analytics for personalized medicine," Curr. Opin. Biotechnol., vol. 58, pp. 161–167, 2019. https://doi.org/10.1016/j.copbio.2019.03.004

[106] G. Torbahn, T. Strauss, C. C. Sieber, E. Kiesswetter, and D. Volkert, "Nutritional status according to the mini nutritional assessment (MNA)®as potential prognostic factor for health and treatment outcomes in patients with cancer--a systematic review," BMC Cancer, vol. 20, no. 1, pp. 1–18, 2020. https://doi.org/10.1186/s12885-020-07052-4

[107] K. Farooq, B. S. Khan, M. A. Niazi, S. J.Leslie, and A.Hussain, "Clinical decision support systems: A visual survey," arXiv Prepr. arXiv1708.09734, 2017. https://doi.org/10.31449/inf.v42i4.1571.

[108] G. Veselov, A. Tselykh, A. Sharma, and R. Huang, Applications of Artificial Intelligence in Evolution of Smart Cities and Societies, Informatica, vol. 45, no. 5, 2021. https://doi.org/10.31449/inf.v45i5.3600

[109] M. Možina,"Arguments in interactive machine learning," Informatica, vol. 42, no. 1, 2018.

[110] A. A. Abaker and F. A. Saeed, "A Comparative Analysis of Machine Learning Algorithms to Build a Predictive Model for Detecting Diabetes Complications," Informatica, vol. 45, no. 1, 2021. https://doi.org/10.31449/inf.v45i1.3111