Implication of Data Mining in Healthcare Field

Suzan Katamoura[a], Layla Hajr[b], Ahmad Alhamed[c]

a King Saud University, IS Department, Imam Road, Riyadh, Saudi Arabia, katamoura@hotmail.com

b King Saud University, MIS Department, Imam Road, Riyadh, Saudi Arabia, lhajr@ksu.edu.sa

c King Saud University, IS Department, Imam Road, Riyadh, Saudi Arabia, aalhamed@ksu.edu.sa

Abstract: The Healthcare domain has very critical data that affects lives. Therefore, this sector needs to utilize this data properly to improve its services, and patients care. Data mining techniques offer excellent analysis methods that can benefit healthcare service providers and patients. There are many available algorithms with various performances, each with pros and cons. However, every algorithm's usage depends on the application, data type, and other features involved. Moreover, vital healthcare prediction applications like heart disease detection, obesity prediction, and cesarean delivery operations forecasting need to be implemented with the support of the different DM approaches. Furthermore, using data mining tools for simple and better analysis results is of great support. Even with the great benefits it introduced, there are still many challenges facing data mining accomplishments. This paper provides an extensive overview of data mining in the healthcare sector. The researchers' contribution are: (1) the comprehensive overview of related elements that work as a basic introduction, (2) structure summary for the advantages and disadvantages of the common techniques, (3) the thoroughly description of the most used tools in healthcare, (4) the detailed comparison of the reviewed literature, and (5) the discussion and recommendations of future directions to benefit from data mining analytics in this field optimally.

Keywords: Big Data, Data Mining, Healthcare, Prediction, Classifications, Deep Learning

1. Introduction (Times New Roman 10 Bold)

The medical field generates a massive amount of data from various sources, such as patients' personal information, medical history, genetic data, pathogen genomics, medical imaging, clinical data, pharmacokinetics, digital epidemiology, and course assessment. This generated data is to be stored and then analyzed to extract valuable knowledge to help understand the sickness and achieve significant leaps in treatments (Suzan Katamoura, 2022). In addition, it can be utilized for medicinal services and medication regulatory purposes. Moreover, technological development allows for making significant and dependable inferences regarding well-being. Since the data recorded in healthcare is large in volume coming in high speeds consisting of varying datasets, mostly with high accuracy from trustworthy data sources, it is considered as big data due to the similarities they share with respect to volume, velocity, variety, value, and veracity. However, the data in healthcare applications has some characteristics as part of the domain requirements like it is huge amount , it is comes from multiple sources , it is represented in different types (e.g text, numeric, document, graph) and it can be dirty and need cleaning and correct (Santos-Pereira et al., 2022), so big medical data differ from other fields' data for the following reasons:

(1) It is hard to frequently access because of its privacy,

(2) It is generally structured as a result of the raw data simplification that occurs during the extraction process,

(3) It is costly since the data collection is conducted through expensive instrumentation,

(4) It has legal issues associated with its use, and

(5) It is affected by several uncertainty factors resulting from measurement or human errors, missing data, and coding mistakes for the information written in reports. Consequently, dealing with big data has many issues related to data integrity, security, and inconsistency that need to be considered while working (Elezabeth et al., 2018).

Furthermore, big healthcare data is a multilayered system that needs automated applications utilizing advanced technologies to best extract the value out of data. Also, it requires Artificial Intelligence (AI) algorithms to analyze the data and Machine learning (ML) methods to facilitate automatic decision-making. Additionally, big data requires efficient strategies for data handling, smart cloud applications, adequate storage, and proper visualization to gain practical understanding (Subrahmanya et al., 2022).

Analyzing big data to extract proper knowledge that helps to make a better decision is called Data Mining (DM), which is a powerful tool for identifying patterns and discovering hidden relationships. Data mining is considered a major and the most time-consuming stage in a bigger process for Knowledge Discovery in Database (KDD process) (Dwivedi et al., 2018). Nevertheless, some literature uses KDD and DM interchangeably. KDD process consists of five main steps, which are (i) data selection from different sources, (ii) pre-processing for the selected data, (iii) transformations in which the data is converted to a suitable format for further processing, (iv) data mining where appropriate technique is implemented for information extraction, and (v) interpretation and evaluation of the resulted knowledge (Tomar & Agarwal, 2013).

The selection step aims to choose the most relevant data for the topic of interest. Then, at the data preprocessing stage, the data preparation, integration from different sources, cleaning, normalization, and reduction occur. Then, the resulting dataset will be input into the next phase, data mining, which is the most time-consuming and problem-solving step. The output from data mining is the discovered knowledge that is needed for application in new situations, which is the ultimate goal of data mining usage. The last step is assessing and interpreting the captured knowledge and then presenting it using appropriate visualization tools (Shirazi et al., 2019).

Additionally, it is worth mentioning that some literature broke down the KDD process into seven phases; data cleaning, integration, selection, transformation, mining, pattern evaluation, and knowledge presentation. (Amin & Ali, 2017), in any case, some data mining tools perform all methodology steps which are called end-to-end analytical tools (Santos-Pereira et al., 2022).

Data mining methods and their applications in the medical field are considered a new concept. Hence, implementing data mining in healthcare has practicality issues, i.e., if the medical assumptions deduced from the data are wrong, all the work would be unsuccessful. Therefore, science and technology must go hand in hand. Moreover, various methods are used in data mining, such as classification, clustering, and deep learning. The use of each technique depends on many factors depending on the dataset and the objective of the work. Despite the issues above related to big and healthcare data, massive benefits will result from capturing the knowledge in this data and analyzing it. These benefits cover a wide dimension, from administration and planning to the quality of healthcare services. The benefits include supporting strategy planning, optimizing facility performance, offering easy access to information, reducing resource wastage, improving proactive equipment maintenance, and reducing the costs associated with research, healthcare services, and Insurance fraud. Other vital benefits are improving healthcare and life quality, predicting epidemics, avoiding preventable deaths, building better predictive models, and collecting warning signs of serious diseases at an early stage for faster and cheaper treatment (Elezabeth et al., 2018).

The following sections elaborate more on the importance of data mining in the healthcare field, the most common data mining techniques and applications, some available tools, the challenges facing data mining implementation in the healthcare industry, discussion, and the proposed future work to increase the advantage of data mining.

This work aims to offer a comprehensive review of the implementation and applications of data mining in the healthcare sector. The authors will highlight the importance of data mining, explain popular techniques, summarize the advantages and disadvantages of these techniques, list vital applications, present common tools, explain challenges faced data mining field, and propose future work to maximize the benefit of data mining analytics in the medical field.

2. Importance of Data Mining in Healthcare Field

In general, all healthcare institutions worldwide store their data in electronic format. Nevertheless, healthcare-generated data are massive and very complex to handle and analyze using traditional methods. Thus, the Data needs to be transformed into useful information for decision-making. Nevertheless, big-data analytics become possible with the advancement of features like flexibility and scalability of information technology (IT) infrastructure, including the centralized repository for structured and unstructured data at any scale (data lakes), cloud data storage, and data warehouse structures, along with management solutions. However, the healthcare industry needs to select the proper IT infrastructure, analytic tools, and visualization approaches to overcome any limitations to the insights provided by big data. Moreover, it is required to change some current policies to balance the benefits of data mining approaches and the protection of patients’ privacy, including data use, access, sharing, and privacy policies.

Since data mining algorithms can deal with large datasets, they can create substantial value in the medical field by improving outcomes and reducing costs since it possesses the ability to handle big data. Furthermore, Data mining can facilitate knowledge discovery across many scientific disciplines. This multidisciplinary characteristic allows for identifying relationships among health phenomena. For instance, a data mining model can utilize geospatial data related to locations of hospitals and patients' houses, and other healthcare data in order to determine and address the complex reasons that lead a set of community members to obtain routine medical care through the emergency unit causing unnecessary expenses. Furthermore, data mining techniques use inductive reasoning and empirical data analysis that allows researchers to detect interesting patterns independent of hypotheses (Roski et al., 2014).

Therefore, data mining is becoming increasingly essential in healthcare to provide successful analyses. For instance, data mining can help detect fraud and abuse in medical insurance, make informed decisions about customer relationships by management healthcare organizations, identify appropriate treatments and best practices for doctors, and offer excellent and affordable healthcare services. Moreover, data mining algorithms show better performance, particularly with the new studies to develop frameworks to handle health big data analytics to understand the data better, overcome the challenges, and provide higher quality results. Also, a few concepts need to be considered to support healthcare data analysis, including data aggregation, maintenance, and integration.

Another evolving area of data mining is descriptive data mining. Its importance arises from the fact that exploratory approaches support the researchers significantly in identifying the relations between the different attributes and categorizing the groups of contributing data. This area of data mining aims to move from specific, i.e., data to general, i.e., knowledge about the effects of a specific attribute, i.e., a gene, on health conditions (Tekieh & Raahemi, 2015).

Another importance of data mining techniques in the healthcare industry is developing intelligent healthcare decision support systems, in addition to supporting the organizations in different medical management decisions making. Examples of such decisions are staying days in a hospital, ranking hospitals, fraud insurance claims by patients and providers, identifying better treatment methods for a particular group of patients, and constructing effective drug recommendation systems (Ahmad et al., 2015).

Fichman identifies six healthcare factors and discusses how they motivate the research's results reported in the literature. These factors describe the importance of data mining in the healthcare field, including the stakes are life and death, healthcare information is highly personal, healthcare is highly influenced by regulation and competition, healthcare is professionally driven and hierarchical, healthcare is multidisciplinary, healthcare IS implementation is complex with important implications for learning and adaptation (Fichman et al., 2011).

Data mining algorithms are essential for understanding sickness and well-being ideas to achieve significant leaps in treatment progress, mainly in disease analysis and aversion. Computers and technology will support Data mining techniques to examine this big data faster with the intelligent mechanism. Such huge information can support making substantial and dependable inferences concerning the well-being of a human (Elezabeth et al., 2018).

3. Data Mining Techniques:

Data mining process includes the selection of the most suitable technique to be embedded to the problem's dataset to identify any patterns or relations in the data and generate results. The algorithm refines crucial parameters to create the mining model, induce actionable patterns based on specific measures, calculate statistics, validate procedures, and then, interprets the captured patterns and combines and exploits the outcomes using other systems. To define the mining model, mostly a composition of multiple algorithms can be used (Sen & Khandelwal, 2018).

In general, data mining analysis is divided into two main categories: supervised and unsupervised analytics. Nonetheless, there is a tendency to apply supervised methods in the healthcare field. In addition, using classification and clustering techniques is most common in medical analysis (Shirazi et al., 2019). Since Deep Learning (DP) is a relatively new data mining technique that can be supervised, unsupervised, or semi-supervised, it will be introduced in a separate section in this paper.

3.1 Supervised Models

Supervised analytics use training sets consisting of input and output pairs to train the system to learn generating knowledge (Dwivedi et al., 2018). They describe the qualitative or quantitative relations among variables (input and output). The supervised methods are powerful tools for predictive modeling using a backward approach in data analysis that require pre-defined model output. Furthermore, their success depends on two factors. The first element is domain expertise, which is crucial for developing efficient models concerning specific architecture, selected inputs, and fine-tuned parameters. However, the drawback of this involvement is reducing the value of big data since only a small subset of variables is used to develop the model. The second factor is the training dataset based on the availability of input and output variables. Moreover, to increase the model efficiency and reliability of the predictive models, the quality of the training dataset must be considered despite the difficulties associated with collecting high-quality data (time-consuming, cost, and impracticality sometimes) (Shirazi et al., 2019).

3.1.1 Classification Algorithms:

Classification is a supervised learning method with predefined class categories. The classification approach is a key and most common data mining technique. It reveals the cohesions and differences between different elements, i.e., detects the common features in a group of data in a class and differentiates them from those belonging to other predefined classes. Moreover, the learning process of the classification model input a set of training data sample where the output is known (Ma et al., 2019). Therefore, it can build models or functions that predict typical characterizations of datasets and then assign unknown data points to the target category. For instance, classify tumors into either benign or malignant classes (Dwivedi et al., 2018) Another example is to associate a risk factor with patients by studying their diseases patterns (Ahmad et al., 2015). Thus, classification analysis is used actively in medical research predictive analysis (Raju et al., 2020). To evaluate the correctness of the classification model, the testing dataset is used for testing (Tomar & Agarwal, 2013).

There are two main methods of classification: Binary and multilevel. While the binary approach has only two possible classes (Female and Male), multiclass classification has three or more target classes (Tumor, Ulcer, Infection) (Ahmad et al., 2015). Generally, there are several algorithms to construct the classification model, including Statistical, K- Nearest Neighbors (K-NN), Bayesian method, Decision Tree (DT) method, Support Vector Machine (SVM), Artificial Neural Network (ANN) (Jothi et al., 2015; Ma et al., 2019).

- Statistical

There are several statistical algorithms used as classifiers, such as Multivariate Time Series (MTS) and its variants: Mahalanobis Distance (MD) and Mahalanobis Space (MS). The Multivariate Time Series (MTS) algorithm consists of multiple time-dependent variables, where each variable depends on its previous values as well as other variables. Therefore, it is useful to use this dependency for future values forecasting. This algorithm is used widely in multivariable statistical analysis. The Mahalanobis Distance (MD) is a multivariate distance measure to calculate the distance between a point and a distribution. It constructs statistical judgments and is excellent for classification when datasets are skewed or highly imbalanced. In addition, the Mahalanobis Distance algorithm has good performance concerning measurement scaling; demonstrates better sensitivity in the testing stage. The Mahalanobis Space (MS) is a database comprising the statistical values, i.e., means, standard deviations, and the variables correlation structure, in the reference group. Thus, it is used to represent the observations' abnormality degree within a known reference group (Jothi et al., 2015).

- K-Nearest Neighbours (K-NN)

K-Nearest Neighbor (K-NN) is the simplest classification method. It is efficient, accurate, and competitive with other classifiers. It classifies unknown data using established data points, detecting an item's properties, and predicting its closest neighbors with comparable attribution. Furthermore, most K-NN votes classifies each new instance depends on, where k is a small positive number that needs to be defined by experiment. However, training examples must remain in memory during categorization. Thus, it is also called Memory-based classification. While its memory demand is a drawback, the training data is never lost. Additionally, Euclidean distance must be normalized for continuous attributes. This is to overcome its major fault, which is that large values affect distance difference measures more than the small ones. Moreover, it manages training data noise and works better for huge datasets. Nevertheless, processing all training set examples and classifying new data will slow down the computer (Ahmad et al., 2015). Healthcare prediction models often use k-nearest neighbor because the algorithm's classification accuracy is more essential in medical diagnosis than classification time (Jothi et al., 2015). For example, K-NN can classify chronic diseases to create an early warning system, such as studying the relationship between cardiovascular disease and hypertension to reduce complications from these diseases (Tomar & Agarwal, 2013).

- Bayesian Methods

The Bayesian learning Method is based on Bayes theory. Bayes theorem focuses on discrete probability distributions, including prior and posterior probabilities of the dataset. It is a simple and efficient classifier (Tomar & Agarwal, 2013). In general, the Bayesian method is computationally efficient and has a natural ability to handle missing data efficiently. Moreover, its models have good prediction accuracy. An additional feature is its ability to avoid overfitting when classifying the data (Jothi et al., 2015).Two standard methods use Bayes's theorem basis: Naive Bayesian and Bayesian Belief Networks (BBN).

- Naïve Bayes (NB) Classifier:

It assumes independence between all attributes, which is not sustained in the medical domain since patients' symptoms and health states are mostly related. However, despite this significant disadvantage, Naïve Bayesian algorithm proved efficient and accurate if attributes are independent (Tomar & Agarwal, 2013).

- Bayesian Belief Network (BBN):

It is more common to be used in the healthcare field for analytics models. For example, it is used to develop decision support systems for health risk analysis (Tomar & Agarwal, 2013). The global interest in BBN is due to the ease it adds to the computation process and the better speed and accuracy it offers when dealing with massive datasets (Ahmad et al., 2015).

- Decision Trees (DT)

The Decision Tree is a very popular approach for representing the classifier. The Decision Tree represents a directed acyclic graph (DAG). The structure of the DAG comprises nodes and edges. There are different types of nodes within the structure of DAG; the root node is the topmost node in the tree without input edges, the internal nodes are the following nodes with input and output edges, the leaf nodes (end nodes) are the nodes with input edges only (Ma et al., 2019).The Decision Tree structure is constructed by dividing the dataset of a specific problem into smaller subsets. Then, the complete tree forms when obtaining the leaf nodes after a specific number of divisions (Raju et al., 2020).

Moreover, the Decision tree classifier is an inductive learning method based on cases (abnormal instances). Thus, the model works recursively in a top-down manner to compare the internal nodes' data values of the decision tree (test on a specific attribute) to decide the branches (outcome of that test) moving down until the leaf nodes, where the conclusions (class label) are represented. Selecting attributes in the training dataset to generate classification rules in the right order improves the final decision (Ma et al., 2019). The Decision Tree classifier helps forecast class labels by identifying the major elements impacting them (Raju et al., 2020). Also, it calculates operational research analysis conditional probabilities. The classifier's ability to accept nominal and numeric input characteristics allows discrete-value representation. This approach is useful in healthcare because domain expertise is not needed to make judgments, its representation is self-explanatory, compact, and simple to manage missing or incorrect information. It helps determine if a patient needs readmission (Ahmad et al., 2015). Predicting breast cancer survival, identifying chronic illness patients' activities, describing skin disorders in various ages, and exhibiting adult smoking habits are further uses (Tomar & Agarwal, 2013). Despite these benefits, decision trees have several drawbacks. It overreacts to the training dataset and noise and cannot handle continuous variables and complicated attribute interactions (Ahmad et al., 2015).

- ID3 algorithm

The ID3 algorithm is one of the popular models used in Decision Tree classification analysis uses a greedy non-backtracking technique. This classification algorithm ranks the attributes' selection on all levels of nodes by information gain and tests each internal node to obtain the maximum class information about that instance. Then, if this attribute divides the sample set into smaller subsets, the entropy value of the model is minimized. Furthermore, the shortest path from an internal node to each child leaf node reduces the average depth of the decision tree, making classification quick and accurate. One good use of this algorithm is mining daily physical assessment data to classify patients and anticipate their health state to recommend nutrition, rest, exercise, and treatment (Ma et al., 2019).

- Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm's concept is based on statistical learning theory. Its ability to be extended for multiclass situations makes this technique appealing. The SVM classifies efficiently by creating single or multiple hyperplanes to segregate attribute values in the input space. Support vectors optimize the data point-hyperplane separation. In addition, the classifier translates the input space into a higher-dimensional space to make data point separation simpler. Polynomial, Gaussian, and sigmoid kernel functions translate non-linear training samples to high-dimensional space. Additionally, it can map using mathematical programming. Due to its many attractive features and excellent experimental performance, particularly in the healthcare field, the SVM classifier is growing in popularity (Tomar & Agarwal, 2013). Some of SVM algorithms' advantages: (1) it provides very accurate results, (2) it can be easily extended to multiclass problems, (3) it offers a generalization of excellent performance and low error, (4) it performs well in high-dimensional spaces even when the dimensions exceed the number of samples, (5) It implements the decision function using a subset of the training dataset, saving memory, and (6) Its capacity to define different kernel functions for the decision function shows flexibility.

However, the classifier's performance degrades as the number of attributes far exceeds the number of examples, and it cannot directly deliver probability estimates (calculates estimates using a costly validation as a five-fold cross) (Ahmad et al., 2015) (Jothi et al., 2015). Examples of SVM analytics are the classification of patients into a high or low risk for diabetes, the classification of microarray (genetic) data for diseases detection, and the prediction model for breast cancer (Tomar & Agarwal, 2013).

- Artificial Neural Network (ANN)

Neural Network Model simulates human neurons with multiple linked nodes. Edges transport data using a transfer function. These networked nodes create output functions in parallel. Then, they make new observations from existing examples, even with network failure nodes. Moreover, a neural network assigns weights to edges and activation numbers to nodes. Consequently, this artificial brain's power varies depending on the quantity of data nodes where data passes through. Finally, the Neural Network's training rules improve the learned network's interoperability (Raju et al., 2020).

The Neural Network algorithm has several core features that make it popular for classification tasks such as, the model can change structure and weight to reduce error, it accurately anticipates outcomes, and it classifies new data types and trains noisy data. In contrast, the Neural Network Technique has drawbacks, including the need for experimentally derived parameters like the optimal number of hidden layer nodes. Hence, its performance is parameter dependent. Further, its training process is time-consuming and computationally expensive. Additionally, the approach is considered a “black-box” method since it doesn't reveal its inner workings. Nevertheless, healthcare uses the Neural Network classification algorithm for its benefits. For example, the algorithm predicts different cancer diagnoses (Ahmad et al., 2015). However, Multilayer feedforward addresses the challenge for multiclass Neural Networks. For instance, several existing models use Artificial Neural Networks to discover lung diseases and analyze chest and heart diseases for developing effective decision support systems (Tomar & Agarwal, 2013). There are two common concepts of the Neural Network classifier:

- Multilayer Perceptron (MLP) Network:

MLP is a feedforward Neural Network that uses a backpropagation algorithm. A few parameters describe its architecture; the number of layers, the number of nodes, the transfer function used in each layer, and the connection between nodes in each layer and those in adjacent layers. Multilayer Perceptron Network has one or multiple hidden layers between the input and output main layers. Each network layer is constructed of units fed with network inputs simultaneously, creating the input layer, which weights them and passes them in return simultaneously to the subsequent (hidden) layers. Finally, the last weighted outputs are fed to the units of the output layer (Sondakh et al., 2017).

- Radial Basis Function (RBF) Network:

RBF is also a feedforward Neural Network that contains two main layers in addition to the input and output layers. However, it differs from MLP by its extra feature that the hidden units make computations by implementing a Gaussian radial function network. Each hidden unit characterizes a specific point in the input sample. Therefore, the output depends on the distance between the point and the example. So, as the two points are closer in the distance, the output is more robust. Nevertheless, the output layer of both algorithms is similar since it takes the linear output combination of the hidden units (Sondakh et al., 2017).

3.1.2 Regression:

Regression is one of the valuable data mining Techniques. It is identified as a mathematical function demonstrating the correlation between variables, one or more independent variables, and only one dependent variable. This technique estimates datasets' relationships and models the data with minimum errors (Dwivedi et al., 2018). Regression can be classified as linear or non-linear regression based on the number of independent variables.

- Linear Regression: is constructed between one numeric dependent variable and one or more numeric independent variable(s); it cannot be used for categorized data. The idea is to find a correlating line between these variables and try to fit the points by calculating their vertical distances from the line and then minimize the sum of the square of the distance. In the linear regression method, both dependent and independent variables are known (Ahmad et al., 2015).
- Non-linear (Logistic) Regression: logistic regression is a type of non-linear model that can be used for categorical data. Also, it does not consider the linear relationship between variables. This model uses multiple independent variables, while the dependent variable can take two values (usually 0 or 1) which are used as a basis for predicting the probability of an occurrence using the logit function (Raju et al., 2020). There are two types of logistic regression based on the number of predicted outcomes: binomial and multinomial.

Like classification, regression is used for predicting the class or outcome of a function. However, classification algorithms achieve excellent performance with categorical attributes, and the regression model is suitable for continuous attributes. This data mining technique is used widely in the medical field for many purposes, such as predicting diseases or patient survivability (Tomar & Agarwal, 2013).

3.1.3 Anomalies Detection:

Anomaly or outlier detection is used for discriminating the abnormal or significant changes in the data record, which may require correction or even more investigation. It is used mainly with other techniques, such as classification or clustering (Dwivedi et al., 2018). This analysis is used often in medical field due to the high sensitivity of the data and the impact that could affected the data accuracy. Furthermore, it is crucial to evaluate the accuracy of the used anomaly detection approach, particularly, if the obtained dataset seems unreliable. Thus, the literature presents three anomaly detection methods including, standard support vector data description, density-induced support vector data description and Gaussian mixture. The assessment resulted in an average of 93.59% accuracy for a balanced dataset with 2.63 standard deviation. Nevertheless, since only one research introduced this method, its effectiveness can't be confirmed (Jothi et al., 2015).

3.2 Unsupervised Models

In unsupervised analytics, no training set is used; instead, the system discovers hidden structures and correlations among the dataset (Dwivedi et al., 2018). Thus, the main advantage of unsupervised methods is the ability to discover previously unknown knowledge. On the other hand, unsupervised methods use a forward approach in data analysis. Thus, the entire dataset is considered input, and no explicitly predefined mining target. Therefore, it can optimally achieve the value of big data for practical applications. In addition, unsupervised analytics are more capable and practical for detecting new knowledge despite the limited background knowledge (Shirazi et al., 2019).

3.2.1 Clustering:

Clustering is partitioning a dataset into similar finite groups with unknown structures in the data (Dwivedi et al., 2018). The clustering splits the objects into (K) clusters; the objects within each cluster have remarkable similarity and significant dissimilarity to objects in other classes, so the Partitioning is based on similarity measure. The clustering methods are categorized into Partitioning Methods, Hierarchical Methos, and Density-Based Methos. It should be noted that using different clustering algorithms on the same dataset may generate different clusters.

- Partitional Clustering

The partitioning method requires defining the number of clusters before dividing the datasets. This approach uses exclusive cluster separation; thus each instance belongs to one cluster and each cluster contains at least one object. Most partitioning techniques use object distance. Thus, methods are grouped by cluster centroid selection, object relocation, and similarity measurement. K-means and K-medoids determine object-cluster centroids similarity after the user sets the number of clusters (K value). K-means is more common. Partitioning clustering (K) randomly distributes things to (k) distinct sets. Iteratively move items. Relocate items based on cluster similarity per cycle. Each cluster has a centroid. Compared to other clusters, examples in the same cluster exhibit significant intra-class similarity. K-mean and K-medoid cluster techniques define the centroid as the cluster's mean or medoid. The k-mean centroid is the mean value of the objects in each cluster, and the K-medoid utilizes the most centrally situated point (medoid), which is the point with the lowest dissimilarity to all other locations in the cluster. The new instance repeats the preceding steps. Partitional clustering approaches outperform hierarchical clustering algorithms in accuracy. Recursively moving items optimize their approach. Hierarchal algorithms cannot scale, whereas partitional algorithms can manage enormous datasets. Algorithms cluster data quicker. These approaches outperform hierarchical algorithms. However, partitional algorithms' clustering results depend mostly on the initial randomly selected centroids, so they always produce different results (Ahmad et al., 2015). Thus, many researchers develop the K-Mean clustering strategy to improve clustering results and address algorithm weaknesses (Sura I. Mohammed ALI, 2021). Clustering improves public healthcare. Using k-means clustering can detect breast cancer recurrence, Alzheimer's disease, and pathologic and non-pathologic groupings. K-means partitional clustering also classifies high cholesterol and blood pressure as high or low heart disease risks (Tomar & Agarwal, 2013).

- Hierarchical Clustering

Hierarchical Clustering separates the dataset without determining the number of clusters. Hierarchical clustering may be agglomerative or divisive depending on the decomposition method. Agglomerative algorithms first cluster each item. Then iteratively merges similar objects to form larger clusters until it meets a termination condition or all objects are merged into one group. However, divisive hierarchical clustering methods assume all objects are in one cluster and iteratively split the dataset into smaller clusters until a condition terminates the splitting or all objects are in one cluster. Hierarchical clustering algorithms might struggle with merging or splitting points. After merging or splitting objects, the following phase will act on the new clusters, making the choice crucial. It won't combine or break clusters or replace objects (Tomar & Agarwal, 2013).

whether any hierarchical algorithm can measure the distance (dissimilarity) between two clusters with more than one object (linkage) and three objects. Any hierarchical algorithm must measure linkage, or distance, between two clusters with more than one object. Single-link, complete-link, and average-link linkage measurements. These three kinds determine cluster build and dendrogram shape. The single-link clustering method groups the nearest pair of items from two separate clusters by dissimilarity. The complete-link method chooses the largest distance between two group items. The average-link method picks the average distances between each pair of items from two clusters, which often yields the greatest accuracy. Merging the shortest relationship between two groups. Hierarchical clustering methods are useful for visualizing item similarities. The researcher may also estimate clusters (Ahmad et al., 2015). Hierarchical algorithms may improve microarray data processing and forecast Rheumatoid Arthritis severity using gene representation. Grouping patients by hospitalization duration improves resource management (Tomar & Agarwal, 2013).

- Density Based

Partitioning and hierarchical methods can be used to cluster spherical-shaped, but They cannot accurately find clusters for arbitrary shapes or convex regions, such as oval clusters. The density-based clustering method evolved to overcome some limitations of partitional clustering and hierarchical clustering methods. The density-based cluster is modeled as dense regions in the data space basis of the analysis of density connectivity. The basic three approaches of density-based clustering are DBSCAN, OPTICS, and DENCLUE. There are many advantages of density-based clustering. One of them is that there is no need to specify the number of clusters at the beginning. Another is its capability to handle arbitrarily shaped clusters. The last one we mentioned here is its efficiency in working with outliers and noisy situations. However, one disadvantage of density-based clustering is that it cannot handle many variations in densities along with data points. Another disadvantage is that calculating the distance is essential for the result (Ahmad et al., 2015). Nonetheless, density-based clustering techniques are essential in biomedical research because they can handle arbitrary shape clusters. These algorithms prove their efficiency and effectiveness in obtaining interesting patterns from a voluminous database containing primarily biomedical images, i.e., using the DBSCAN algorithm to cluster wounded skin images (Tomar & Agarwal, 2013).

3.2.2 Association:

Association Rule learning detects frequent patterns and establishes a correlation between variables in the data repository. It is also called dependency modeling (Dwivedi et al., 2018). Additionally, it is known as market basket analysis because it identifies the association of purchased items or hidden patterns of customers' sales in a transactional database (Tomar & Agarwal, 2013). Association mining attracted substantial attention due to its ability to extract association rules when applied to actual databases. Thus, the association rule method remarkably detects the correlation among diseases, symptoms, and health status in the healthcare field. Also, it is heavily used to determine the affiliations between different diseases and the prescribed medicines. Similarly, medical insurance companies use this approach to detect fraud and abuse (Ahmad et al., 2015).

Furthermore, it is used to detect intermittent relations in electronic health data such as heart patients' data. When integrated with classification methods, it enhances healthcare data analysis by finding rules in the data, then using them to construct an efficient weighted associative classifier (Tomar & Agarwal, 2013). Moreover, unlike classification and clustering, the evaluation factor and the primary objective of association mining algorithms is efficiency. However, since the association method mines all association rules, the accuracy is not considered an evaluation factor. There are two common associative rule algorithms:

- Apriori Algorithm

The Apriori algorithm determines the relations among datasets by demanding two input percentage values, support and confidence, and filtering the undesired association rules based on predefined criteria. Respectively, these values refer to the user's interest in the association rules that repeatedly occur in a dataset and their accuracy. The principle of the Apriori algorithm considers a minimum support and confidence constraint. Thus, the item's descendants are not frequent if the item itself is not frequent, which considerably reduces the search and improves the algorithm's efficiency. There are various methods to improve the efficiency of the Apriori algorithm, including Hash table, transaction reduction, and partitioning (Ahmad et al., 2015). The apriori algorithm is used in literature to generate association rules, then use them to classify the patients with type-2 diabetes, detect the occurrence of diseases in specific geographical locations at certain timespan, and generate the heart disease rules by revealing the factors related to gender (Tomar & Agarwal, 2013).

- Frequent Pattern (FP) Tree Algorithm

FP-tree algorithm detects frequent datasets without making a candidate dataset. This algorithm works in two steps. First, the algorithm constructs its data structure. Then, the frequent dataset is drawn from this data structure. This association analysis finds out the unseen relationship among attributes. Thus, it is extensively used in healthcare to determine the relationship between different diseases and drugs. For instance, it is used to identify interesting patterns and build knowledge discovery models to find valuable information in audiometric datasets (Tomar & Agarwal, 2013).

3.2.3 Summarization:

Data's growth and complexity have highlighted the need to summarize the data to obtain valuable information. The data summarization technique is alternatively called visualization. It provides vital support to summarize an entire dataset using its most significant factors or attributes. Also, it helps to visualize the dataset and present the mined trends and patterns in a comprehensible and simplified manner. The presentation can be in tabular or graphical format (i.e., Tabular Summarization, Data Visualization) (Raju et al., 2020). In addition to visualization, summarization offers a compact dataset representation, such as a concise dataset description and report generation (Dwivedi et al., 2018). Summarization is a fundamental data mining concept. The appropriate statistical tests can be decided based on the general trends discovered from the summarization. Data summarization has a commonplace application in the healthcare field to deliver high-quality services, patient care, and optimal decision-making support. For example, summarize clinical information to generate discharge summary reports, daily progress notes, change of shift patient endorsement, and medical cases' presentations (Feblowitz et al., 2011).

3.3 Deep Learning

Deep learning is a raising topic in data mining field that descendants of machine learning field. It is an optimized neural network with more than two layers where the extra hidden layers refine for accuracy. Deep learning simulates human brain function to learn from massive data sets enabling accurate predictions (Abdullah et al., 2022). Depending on availability of labeled datasets, deep learning may be supervised, unsupervised, or semi-supervised. Deep learning improves analytical and physical automation, and processes even unstructured data, such as text and images. These capabilities remove some human experts’ dependency. For example, deep learning can automatically identify crucial traits to classify medical photos. Additionally, deep learning can also represent non-linearity, which enhances normal and anomalous point segregation and data variation modeling. However, this technique is costly to scale as it requires a powerful computing and hardware components, i.e., processing units, copious memory, etc. (Kaul et al., 2022). On the other hand, computational resources enable the algorithm hierarchical feature learning process, which eliminates the need to explicitly define anomalies during feature design. Deep learning may easily discover long-term dataset correlations without specifying them due to its neural network design. Nonetheless, deep learning is needed in healthcare for image recognition and medical records analysis. Anomaly detection is crucial in medical signal analysis. Recurrent architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are used in medical anomaly detection deep learning. These approaches are efficient in modeling the temporal correlations in time series data using memory. Data types, problem specifications, and learning strategies drive model selection. For example, For medical diagnostics, black-box deep models must be interpreted to comprehend the choice (Tharindu Fernando, 2020). Also, deep learning systems also detect cancer. Nowadays, the researchers developing deep learning algorithms for preventative and therapeutic healthcare (Helén, 2022).

Choosing an Algorithm

A paramount concern when using data mining algorithms is selecting the algorithm for the task. Answering this question depends on the type of data and the application or problem under analysis. For example, the classification algorithms are most suitable when the data samples have several attributes, as the classifier algorithm predicts a single or multiple discrete variables using these attributes. However, regression algorithms are appropriate mostly to predict continuous variables. On the other hand, association algorithms help determine the relationship between various characteristics in the dataset to support domain experts in considering significant association rules that are vital decisions. In comparison, classification rules focus on discovering the class of attributes without considering the relationships of attributes. In addition, the clustering technique is used when there is no or little knowledge about the dataset. Nevertheless, clustering type is also used in different contexts, i.e., hierarchical clustering is used when there is no predefined cluster number, while partitioned algorithm requires a predefined number of clusters. Furthermore, it is also worth mentioning that there is no single algorithm that can produce the best result for every dataset (Sen & Khandelwal, 2018).

Techniques’ Advantages and disadvantages:

A comparison summary of supervised and unsupervised data mining techniques is in Table (1), including a comparison of their main advantages and disadvantages from the literature.

Table 1: Advantages and Disadvantages of DM Techniques

Supervised Models		
Algorithm	**Advantage**	**Disadvantage**
Classification Methods		
Statistical	- Practical to use this dependency for future values forecasting. - Good performance concerning measurement scaling	- Skewed or highly imbalanced data affect the algorithm's performance.
K- Nearest Neighbour (K-NN)	- Easy to implement. - Analytic method. - Training is fast - Use local information.	- It is a slow process. - High storage required - Affected by dimensionality. - Sensitive to noise. - Testing is slow.
Naïve Bayes Classifier	- Fast and easy to use and predict the class of the test dataset. - When the assumption of independence holds, it works better than other classification algorithms. - Works well in the case of categorical input variables in place of numerical variables.	- Limited to the independence of the predictor's assumption. - If a categorical variable has a category (in the test dataset), which was not observed in the training dataset, then the model will assign a 0 (zero) probability and cannot make a prediction.
Decision Trees (DT)	- easy to comprehend, interpret, and decision making. - Allows for the addition of new possible scenarios. - Can be combined with other decision techniques. - No required domain knowledge. - Easy to process high-dimension data. - Handles both numerical and categorical data.	- Restricted to one output attribute. - May suffer from overfitting. - Unstable depending on the dataset type. - Relatively inaccurate. - May generate complex decision-making situations for numeric data. - Categorical data is challenging to handle.
Support Vector Machine (SVM)	- Effective in high dimensional space. - Resistant to overfitting. - Memory efficient as it uses only SV to build hyperplane. - Better accuracy than other classifiers. - Easy to handle nonlinear data points.	- Used for small datasets - Computationally expensive. - The training process takes more time. - Premature optimization - The main issue is the selection of the correct function. - Designed to solve the problem of binary class.
Artificial Neural Network (ANN):	- Ability to learn how to do tasks based on training/experience. - Can handle missing or noisy data in case of fault or partial network destruction. - Can easily work with a large number of datasets. - Easily identify complex relations between dependent and independent variables.	- Since a network cannot be retrained, it is challenging to modify an existing network. - Overfitting issue. - Local minima. - Processing is difficult to interpret and requires high processing time in large neural networks.

Regression Methods:	- Predict data with minimum errors. - Suitable for continuous attributes. - Provides information about features' statistical significance	- Performance is low with categorial attributes. - Suitable for low cardinality.
Anomalies Detection methods:	- Excellent performance for a balanced dataset.	- Depends on the algorithm used with it. - Performance is affected when dimensionality increase.
Unsupervised Models		
Algorithm	**Advantage**	**Disadvantage**
Clustering Methods		
Partition (K-means) Clustering	- Simple clustering approach. - Efficient. - Less complex method.	- Requires predefined clusters no. - Problem with handling categorical attributes. - Handle only spherical-shaped clusters and are not - Suitable for discovering arbitrary shapes. - Result varies in the presence of an outlier
Hierarchical Clustering	- Easy to implement. - Good visualization capability. - No need to predefine clusters no.	- Slow - cubic time complexity in many cases. - Once a decision regarding the selection of a merge or split point is made, it cannot be undone. - Not work well in the presence of noise and outliers, and not scalable. - Unsuitable for discovering arbitrary shapes
Density Based Clustering	- No need to predefine clusters no. - Easily handle arbitrary shape clusters. - Worked well in the presence of noise.	- Not handle the data points with varying densities. - Results depend on the distance measure.
Association Methods:	- Higher efficiency than other algorithms. - Determine the direction and strength of each relationship	- Only reveal the relations, not the reason for the existing connection. - A time-consuming process
Summarization Methods:	- Easy to understand the data - Easy to communicate - Easy to determine relationships between data points	- Gives an assessment, not explanations.
Deep Learning Methods:	- Do not require human intervention. - Strong learning ability from the results. - Good performance.	- Computationally costly - High hardware requirements - Complex model design - Time-consuming - Inability to explain the reason for reaching a specific result.

4. Healthcare Applications of Data Mining Algorithms

This section will introduce three common medical applications of data mining methods as presented by different literatures.

4.1 Heart diseases Detection

Alternatively, cardiovascular disease (CVD) impacts the heart's normal activity. In particular, cardiovascular disease concerning with blood vessels. Additionally, specific activities cause blood obesity, pressure fluctuation, and increased blood glucose level, which result in heart disorders. Such activities include physical inactivity, tobacco smoking, unhealthy diet, or other reasons. In addition, various forms of heart illness include Cardiac Arrhythmia, Angina, Congenital Heart Disease, Coronary Artery Disease, Cardiomyopathy, Myocardial Infarction, Congestive Heart Failure, Mitral Valve Disease, and Pulmonary Stenosis. Unfortunately, heart diseases take millions of lives annually, i.e., 31% of global deaths. Therefore, diagnosing heart diseases at early stages allows proper treatment and even prevents death. Thus, in the healthcare domain, there are tremendous efforts in heart disease prediction (Dwivedi et al., 2018).

The researchers used various classification algorithms to implement a heart disease prediction model: Decision Tree, Artificial Neural networks, and Bayesian Classifier. Therefore, the performance of the different algorithms was analyzed using various measures like accuracy and the time taken to construct a model. Generally, implementing algorithms with related feature selection offers better results than that considering all attributes, excluding the Naïve Bayes technique, which scores the highest accuracy in both cases using all or selected features, about 82%. In contrast, the decision tree model using all attributes gives the lowest accuracy, about 77 %. Similarly, decision trees and naïve bayes have the fastest execution time, while the neural networks algorithm spent the longest time in constructing the model. Furthermore, due to the sensitivity of heart disease, it is crucial to consider a high True Positives value and a low False Positives value. Thus, the classification algorithms are the most suitable for the early identification of disorders, considering the algorithm's accuracy in identifying heart patients as well as the rates of both True Positives and False Positives values (Shafique et al., 2015).

4.2 Predicting Obesity

Healthcare sectors are actively working to provide solutions to illnesses. With the help of evolving data mining techniques that discover interesting patterns in data, healthcare institutions offer better services. As mentioned previously, the primary data mining application in the medical domain is disease prediction, treatment analysis effectiveness, and correlation between symptoms and diseases. A primary medical condition affecting people's lives worldwide is obesity, which is excessive body weight. It is the leading cause of many other health complications like heart attacks, diabetes, and depression. A combination of an unbalanced diet and an unhealthy lifestyle can cause obesity, sometimes genetic causes. Therefore, it is crucial to predict obesity.

Many methods were used to predict obesity in patients from other factors, such as classification algorithms. The literature compares a few classification techniques, including decision trees, neural networks, and Logistic regression, to find the most accurate technique. The neural networks model with ten hidden layers was found to be slightly more accurate than decision trees with a value of 73%. In comparison, the decision tree model shows 72% accuracy. However, increasing the number of hidden layers improves the accuracy. Nevertheless, increasing hidden nodes reduces the model feasibility as the time complexity and processing power increase dramatically. Other classification techniques offer a 70–75% prediction accuracy rate, which is insufficient due to other existing issues with these models. Nevertheless, adding more attributes to the dataset can increase its accuracy (Raju et al., 2020).

4.3 Predict Caesarean Delivery Operations

The revolution of technologies with the possession of a massive amount of medical data make it critical to healthcare decisions. For instance, some health situations, such as cesarean delivery operations, require urgent decision-making. Cesarean delivery is considered an existing complication or challenge to normal delivery. Thus, it is crucial to predict the delivery type within an appropriate timespan for better preparation by medical staff and the mother. Therefore, the dataset of cesarean delivery cases analysis gains interest among researchers. Since it is vital for predicting the safest delivery type for mother and child, several data mining techniques were examined, i.e., naïve Bayesian, support vector machine, k-nearest neighbors, linear regression, and decision trees. Moreover, a cross-validation (CV) approach is used to evaluate the employed models to ensure accurate and reliable results. The results demonstrate that the naïve Bayesian algorithm was more accurate than other classifiers, averaging 65%. In contrast, the linear regression produced the lowest accuracy model with a data accuracy of 48% rate. There is a great need for additional data on the cesarean delivery operation to improve the prediction accuracy, as available data is rare. Finally, the result provides significant recommendations about the leading causes and supports in decision-making (Jamjoom, 2020).

5. Data Mining Tools

The healthcare field is one of the areas most impacted by data mining. However, there are significant gaps in this field because medical researchers often lack experience in data mining techniques and the ability to interpret the results. On the other hand, computer scientists are unaware of specific details in other domains like the medical. To fill this gap, data mining tools have become increasingly important in healthcare, especially with the growing amount of data generated from electronic health records (EHRs), medical imaging, and other sources. These tools help healthcare organizations to analyze data and gain insights to improve patient outcomes, reduce costs, and optimize workflows. This research section will discuss some of the commonly used data mining tools in healthcare. Various data mining tools are commonly used in healthcare depending on the specific task or application. Available data mining tools have different features, interfaces, methods they support, data types that can be accessed, and many capabilities. Overall, many different data mining tools are available for healthcare applications, each with strengths and weaknesses. Therefore, the choice of tool will depend on the specific data mining task, data available, and user preferences.

The most popular open-source data mining tools in healthcare:

WEKA: a popular open-source data mining software that can be used for various healthcare applications, including disease diagnosis, prediction, and treatment.

RapidMiner: a powerful and easy-to-use data mining software. It is an open-source data mining tool used in healthcare for data preprocessing, predictive modeling, and visualization. It offers a range of machine learning algorithms and statistical models. It can be used for various healthcare applications, such as disease diagnosis, drug discovery, patient monitoring, hospital readmission rates' prediction, high-risk patients' identification, and clinical decision-making improvement (RapidMiner. https://rapidminer.com/industries/healthcare/).

KNIME: an open-source data analytics platform that is widely used in healthcare data mining, such as for disease risk factors identifying, clinical pathways analyzing, patient stratification, clinical trial analysis, drug discovery, and patient outcomes prediction. In addition, KNIME offers a range of healthcare-specific extensions, such as the Healthcare Extension, which includes nodes for processing medical images and genomic data (KNIME. https://www.knime.com/solutions/healthcare).

IBM SPSS: a comprehensive statistical analysis software that can be used for healthcare data mining tasks, such as identifying patterns in patient data, predicting treatment outcomes, and analyzing clinical trial data.

IBM Watson Health: IBM Watson Health is a healthcare data mining tool suite that uses machine learning algorithms to analyze structured and unstructured data. The tools can help healthcare organizations to identify trends, predict outcomes, and personalize treatments. In addition, IBM Watson Health provides predictive modeling, data visualization, and data exploration tools and offers solutions for population health management, clinical decision support, and patient engagement (IBM Watson Health. https://www.ibm.com/watson-health/).

Orange: a user-friendly data mining software that can be used for various healthcare applications, such as predicting patient outcomes, identifying high-risk patients, and analyzing electronic health record data.

SAS: a powerful and widely used statistical analysis software that can be used for healthcare data mining tasks, such as analyzing patient data, predicting disease outbreaks, and identifying trends in healthcare utilization.

SAS Enterprise Miner: SAS Enterprise Miner is a data mining tool widely used in healthcare for predictive modeling and data analysis. It can handle large datasets and provides a range of data mining techniques such as decision trees, neural networks, and regression analysis. In addition, SAS Enterprise Miner can be used for various healthcare applications, such as fraud detection, disease surveillance, and patient risk assessment (SAS Analytics. https://www.sas.com/en_us/industry/health-care.html).

Microsoft Azure Machine Learning: Microsoft Azure Machine Learning is a cloud-based data mining tool that can be used for predictive modeling, data visualization, and machine learning. It offers a range of algorithms and supports popular programming languages such as Python and R. Microsoft Azure Machine Learning can be used in healthcare for applications such as predicting readmission rates, analyzing patient data, and identifying high-risk patients.

The above-mentioned tools are just a few examples of the many data mining tools available for healthcare applications. These tools provide a wide range of features and capabilities that can help clinicians and researchers analyze and interpret healthcare data. The selection of a data mining tool will depend on the specific needs of the healthcare organization and the types of data being analyzed.

Santos-Pereira et al., 2022 discuss KNIME, scikit-learn Spark, and RapidMiner tools. Then, compare them based on healthcare application requirements. The selected tools are compared based on different criteria selected from different perspectives like Critical Capability measured by performance and scalability, data access, data preparation, data exploration and visualization, and user experience, and based on Data characteristics extracted from the healthcare domain like a large amount of data, multiple data sources, different data types and cloud, dirty and complex data, and based on data mining methods applied in the healthcare industry like classification, clustering, association, and outlier. Moreover, compared to user experience capabilities like supporting operating systems, interfaces, collaboration, and ease of use. Also, the study discussed the feature and limitations of each tool. Finally, they compared those types of tools in one table, concluding that the most suitable tools from these views for the healthcare field are RapidMiner and KNIME (Santos-Pereira et al., 2022).

6. Data Mining Challenges in Healthcare Field

With the continuous flow of the massive volume of medical data from all the heterogeneous sources and different data types, the complexity of the data increases. Despite the remarkable benefits derived from data mining applications in the healthcare domain, there are some limitations. For example, Electronic Health Records contain various clinical elements, so handling these datasets of millions of individuals becomes a major challenge and hurdles the decision-making. The key challenges include the following:

- Generally, the collected data is unorganized/inaccurate, which makes it difficult to gain insights into it.
- It is always challenging to maintain the correct balance between preserving patient privacy and ensuring adequate data quality and quantity for analysis (Sadiku et al., 2018).
- Management of this data and its standardization, privacy and security protection, storage efficiency, and transmission needs an abundant workforce (Elezabeth et al., 2018).
- It is hard to integrate genomic data into medical studies without the standards to handle bioinformatics, generate sequencing data, and support medical decision-making. Thus, it may cause longer time for the data analysis than usual.
- The language barrier arises when dealing with such voluminous data (Subrahmanya et al., 2022).
- Maintaining the balance between data security and the desire to share them (Krumholz, 2014).
- Data accessibility is an issue since raw data mining inputs mainly exist in various settings and systems, including microarray, administration, laboratories, clinics, and more. Hence, building a data warehouse before performing data mining is suggested, which can be expensive and time-consuming.
- Some data problems affecting data mining algorithms' performance include missing, inconsistent, corrupted, or unstandardized data (Tomar & Agarwal, 2013).
- The task domain knowledge is required for successful data mining applications, methodology, and tools.
- Investment of resources, precisely time, effort, and money, is substantial for healthcare organizations.
- Lack of collaboration among all parties involved in the data mining task (Fichman et al., 2011).
- Legal, ethical, and societal restraints of data to be used in data mining analysis. These include selling prescription information or other patients' data for research purposes, which puts patient privacy at stake (Sen & Khandelwal, 2018).
- Relying on Predictive Models is an issue in healthcare data mining. Although the models are used primarily to reduce uncertainties related to decision-making, they suffer from uncertainties related to themselves. However, these uncertainties can be minimized to acceptable levels through model management and regulatory adherence (Tekieh & Raahemi, 2015).
- Data sharing is a vital requirement for planning to deliver better treatment and services for a large population. However, this could be an obstacle as neither patients nor healthcare institutes are motivated to share their private data. Thus, it will be difficult to offer better treatment to the population or detect fraud and abuse in medical insurance companies (Ahmad et al., 2015).

7. Discussion

The reviewed papers cover various healthcare-related data mining and deep learning topics. In addition, the papers also touch on related topics such as information systems, knowledge management, and big data. Moreover, some focus on specific applications of data mining techniques, while others provide broader field overviews. Table (2) presents a comparison of some of the critical aspects of each paper:

Table 2: General summary of the literature

The paper	Methodology	Techniques	Findings	Limitations	Implications
Feblowitz et al. (2011)	Proposed a conceptual model using data mining techniques for summarizing clinical information in EHR.	Summarization, user interfaces, DSSs	The model improves the efficiency and accuracy of clinical decision-making.	No specific technical or implementation guidelines. Transforming conceptual framework into practical applications requires research and development. Did not address the challenges of data integrating from disparate systems or unstructured data.	Significant for healthcare information systems and decision support tools. Model provides a foundation for developing clinical summarization techniques and tools that can aid healthcare providers in quickly accessing and comprehending relevant patient information. These tools can improve patient care, enhance clinical decision-making, and increase efficiency by reducing information overload.
Fichman et al. (2011)	An overview of the role of information systems in healthcare, including benefits and challenges. Discuss the current research and future trends, including EHR, telemedicine, and health information exchange.	-	IS enhances healthcare delivery, quality, and efficiency. Evaluate current IS research on clinical decision-making, patient involvement, and healthcare analytics. Suggest cautious planning and effective implementation to overcome challenges.	Did not include empirical studies or data to support the findings. Did not capture the most recent developments in IS and healthcare as the field has evolved since 2011. Did not address the context-specific factors that influence the role and effectiveness of IS in healthcare.	Emphasizes the importance of investing in information systems and technology infrastructure to enhance healthcare delivery and outcomes. Highlights the need for interoperability standards, data governance frameworks, and privacy safeguards to address the identified limitations. Underscores the importance of continuous research and innovation in IS to keep pace with emerging healthcare needs and technologies.
Tomar and Agarwal's (2013)	A survey on healthcare data mining methods discusses the challenges, including data quality, privacy, and complexity.	Clustering, classification, association rule mining, outlier detection	Highlights the benefits, such as improving clinical decision-making, reducing medical errors, and identifying cost savings opportunities.	Not a comprehensive survey because new data mining, machine learning, and artificial intelligence techniques have emerged since the paper's publication in 2013.	Highlights the potential of data mining approaches in improving healthcare decision-making. Provides insights into the various techniques utilized for different healthcare applications. Emphasizes the need for further research and development to address the data mining challenges
Krumholz, H. M. (2014)	Discuss the opportunities and challenges associated with big data in healthcare. Discusses big data and new medical knowledge and proposes the need for a learning health system.	-	By analyzing large datasets, researchers and healthcare professionals can identify patterns, trends, and correlations to enhance diagnosis, treatment, and healthcare delivery.	Focuses on the benefits and potential of big data in healthcare and does not discuss the risks or ethical considerations associated with large-scale data analytics.	significant for healthcare stakeholders. Emphasizes the transformative potential of big data in healthcare and advocates for adopting a learning health system approach. It sheds light on the thinking, training, and tools needed to effectively harness big data to generate new knowledge and improve patient care.
Roski et al. (2014)	Provide a comprehensive overview of the potential opportunities and challenges associated with big data in healthcare and offers a clear and concise summary of the current state of big data in healthcare, including the limitations and ethical considerations that need to be addressed.	-	Presents value creation prospects, implementation issues, policy consequences, population health management, clinical decision support, customized medicine, and healthcare quality, outcomes, and prices enhancements.	Authors could have provided more concrete examples of how big data has been used to drive positive health outcomes, rather than just focusing on potential use cases. Legislative reforms are needed. Big data in healthcare presents security and privacy issues, as the authors note.	Highlights the opportunities presented by big data in healthcare, including improved outcomes, cost savings, and enhanced population health management. Emphasizes the need for policymakers to consider the policy implications and create an enabling environment for responsible data use. Collaboration among stakeholders is crucial to realizing the full potential of big data in healthcare and driving positive industry change.

Sen & Khandelwal (2014)	A detailed overview of healthcare data mining techniques and applications, such as disease prediction, patient monitoring, and drug discovery, focusing on EHR. Discusses the challenges and limitations of data mining.	DT, classification, clustering, association rule mining	Data mining in healthcare can introduce data quality issues and requires domain knowledge.	The authors did not offer suitable examples of data mining usage for better outcomes. Also, the paper is very technical and may be difficult to understand for readers without a background in data science.	Easy to follow making it accessible to a wide audience, including healthcare professionals and researchers.
Ahmad et al. (2015)	Review healthcare data mining techniques and discuss the challenges associated with using these techniques in healthcare, such as privacy concerns and the difficulty of obtaining high-quality data.	Partitional, hierarchical, and density-based) clustering, K-NN, SVM, DT, NN, BM, regression, association rule	Presence of some factors like noisy data, can greatly affect accuracy and performance of a data mining algorithm. The patient's confidential information privacy and data security are very important.	The paper lacks a critical evaluation and evidence quality for the studies reviewed. Also, no practical recommendations for healthcare practitioners or policymakers on effective data mining techniques.	This paper is recommended for researchers and healthcare professionals who want to learn more about data mining techniques and their potential applications in healthcare.
Jothi et al. (2015)	Provide a general review of the data mining in healthcare, algorithms, applications, and associated challenges, such as the need for high-quality data and interpretability. Also, provide examples of successful data mining applications.	Clustering, classification, association rule mining, and outlier detection	The article concludes with recommendations for future research in data mining for healthcare, including the need for interdisciplinary collaboration and the development of standard data mining tools.	While the authors highlight some potential benefits and challenges of data mining in healthcare, they do not provide specific guidance on addressing these challenges or how to integrate data mining into healthcare practice.	Provides a good resource for researchers and practitioners interested in data mining in healthcare.
Shafique et al. (2015)	Compare data mining process models (KDD, CRISP-DM, SEMMA), discuss each process model's advantages and limitations, and provide examples of their healthcare applications; heart disease diagnosis/prediction.	KDD, CRISP-DM, SEMMA, processes	The importance of domain knowledge and data preprocessing in data mining models. Data mining techniques improve the diagnosis and treatment of heart diseases.	Did not discuss potential biases or limitations of the data mining techniques used in the studies.	Highlights the potential benefits of data mining in healthcare for heart diseases, such as improved accuracy in diagnosing heart diseases and predicting patient outcomes.
Tekieh and Raahemi's (2015)	Present a survey on various data mining techniques used in healthcare, their importance, and the associated challenges like data integration, quality, and privacy.	Clustering, classification, association rule mining, anomaly detection	Data mining in healthcare, improves diagnosis and treatment, reduces readmissions, and identifies high-risk patients.	The study's sample size is not clearly stated, making it difficult to assess the representativeness of the results.	Data mining techniques are highly valuable in healthcare, particularly in reducing costs and improving patient outcomes; identify patterns and trends in healthcare data that can be used to inform clinical decision-making and improve patient care.
Amin and Ali (2017)	Overview on using multilayer perceptron (MLP) for data mining in healthcare operations and explore its application in predicting the risk of heart disease.	MLP	The effectiveness of MLP algorithms in healthcare operations is because they are relatively simple to implement and can handle nonlinear relationships between variables.	Did not discuss the ethical and privacy concerns associated with the use of patient data in data mining applications in healthcare.	Recommended to researchers and healthcare professionals interested in learning more about the application of MLP in healthcare operations, be applied in healthcare operations, specifically in predicting patient outcomes and optimizing healthcare resource allocation.
Sondakh (2017)	Compares various neural network algorithms for data mining in	MLP, RBF, LVQ, GRNN	MLP and RBF performed better than LVQ and GRNN in terms	The use of a single data set, which may not represent all healthcare	The study provides insights into the performance of neural network algorithms in predicting healthcare

	healthcare data and their applications, such as disease diagnosis and patient monitoring. Discusses the advantages and limitations of data mining models using neural networks.		of accuracy, sensitivity, specificity, and F-measure, and the size of the dataset had a substantial impact on the performance of the algorithms.	settings, and the limited number of algorithms considered.	outcomes, which can guide the selection of appropriate algorithms for healthcare data mining projects.
Dwivedi et al. (2018)	Provide an overview of various data mining algorithms used in healthcare and discuss some associated challenges, such as the need for interpretability and accurate data.	ANN, DT, RF, NB, SVM, K-NN	decision trees, rule-based classifiers, and clustering algorithms are widely used in healthcare applications, but selecting a suitable algorithm depends on the type of data and the problem.	Using data mining can result in data privacy and security concerns, data quality issues, and the interpretability of the results.	Healthcare providers need to develop a comprehensive data governance framework to address data privacy and security issues and ensure data quality, and researchers need to build more interpretable data mining algorithms to enable healthcare providers to better understand and use the results.
Elezabeth et al. (2018)	Review the role of big data mining in healthcare applications, such as EHR, telemedicine, and clinical DSSs, as well as potential benefits and challenges. Provide examples of big data mining usage for predicting disease risk and improving patient outcomes.	Data preprocessing, analysis, visualization	Big data mining provides insights into the traditional methods and helps healthcare professionals make better-informed decisions. Data privacy and security are important for healthcare since data mining access sensitive information.	No case studies or examples of successful implementation of big data mining in healthcare. Did not discuss the associated challenges, such as issues related to data quality and standardization, integration of data from multiple sources, and lack of skilled personnel.	Big data mining has the potential to revolutionize healthcare by reducing costs and improving patient outcomes.
Sadiku et al. (2018)	Briefly overviews data mining, its various techniques, and applications in healthcare, such as disease diagnosis, treatment prediction, and drug discovery. Discuss the limitations of data mining, such as the need for extensive and high-quality data sets and the potential for bias in data mining models.	Classification, clustering, association rule mining	Data mining can improve diagnosis, treatment, healthcare management, reducing costs, and increasing patient satisfaction.	The paper is relatively short and lacks in-depth analysis or empirical evidence to support its claims.	Easy to follow by a wide range of audiences introducing basic knowledge.
Helén (2019)	A qualitative analysis of the narratives surrounding data mining in healthcare.	The narrative of efficiency, the narrative of individualization, and the narrative of collaboration.	Narratives are crucial in transforming healthcare by providing insights into disease patterns and treatment outcomes.	Did not provide a technical analysis of data mining or empirical evidence of its effectiveness in healthcare but relies on qualitative analysis and theoretical frameworks.	Encourages a more reflective and critical approach to create a better understanding of the potential benefits and limitations of data mining in healthcare. The narratives complement each other, not mutually exclusive.
Ma et al. (2019)	Proposes a medical examination data mining system based on the decision tree model to predict diseases with a comprehensive dataset analysis.	DT, preprocessing, feature selection, and model evaluation	The model extracts hidden patterns and relationships effectively, and the decision tree algorithm achieves high	Used a small sample size for testing the system and the lack of comparison with other data mining algorithms. Decision tree algorithms may not	Well-structured and can be used as a reference for similar studies. The system can help identify patients at risk of developing certain health conditions and provide early interventions, leading to better health outcomes if validate system effectiveness with larger samples and

			accuracy in disease prediction.	perform good on large datasets and can be prone to overfitting.	compare its performance with other data mining algorithms.
Shirazi et al. (2019)	Provide an application-based review of recent healthcare data mining advances, such as disease diagnosis, treatment prediction, and drug discovery, and examples of recent research in each area. Discuss the challenges of healthcare data mining; the need for large and high-quality datasets, and the potential model's bias.	-	Various data mining techniques are used to develop predictive models for various medical applications. Integrating various data sources (EHR, medical images, genetic data) enhanced models' accuracy. Data mining has also been used for identifying patterns and trends in healthcare data (identifying disease outbreaks, and detecting adverse drug reactions).	Did not provide a comprehensive review of all recent advances in data mining applications in healthcare. Did not include a detailed discussion of specific applications.	Data mining can potentially to improve healthcare outcomes and reduce costs, but careful consideration is needed to ensure that data is used ethically and with due regard for privacy and confidentiality concerns. Further research is needed to develop effective data mining techniques and to address the challenges associated with using these techniques in healthcare.
Fernando et al. (2020)	Discuss medical anomalies and the challenges of using traditional machine learning methods to detect them. Survey deep learning-based anomaly detection methods in medical images, signals, and EHR. Discuss the potential benefits and challenges of using deep learning for medical anomaly detection, and the limitations of current approaches and future research directions.	CNN, RNN, GAN	The key challenges in medical anomaly detection are limited annotated data and class imbalance. Various deep learning architectures and algorithms are proposed to address these challenges. Applying transfer learning, domain adaptation, and unsupervised learning techniques in medical anomaly detection.	Did not conduct any experiments to validate the effectiveness of the reviewed methods. Focuses on deep learning-based anomaly detection methods, so it may not cover other relevant areas of medical data analysis.	A useful overview of the state-of-the-art deep learning-based methods for medical anomaly detection for interested researchers and practitioners in the healthcare industry. Highlights the potential of deep learning techniques to improve medical diagnosis and treatment. If the limitations of current approaches are considered carefully when designing and implementing new methods.
Joshi et al. (2020)	Discuss data mining for predicting obesity using demographic and clinical data. Present a data mining predictive model to predict obesity by analyzing the factors affecting obesity.	SVM, DT, CRISP-DM process	Compared the performance of various classification methods, and the decision tree algorithm gave the best accuracy of 93.4% for obesity prediction.	Used data from a single hospital, limiting the generalizability of the findings. Did not consider the effect of key factors, such as physical activity and diet, on obesity, which may affect the prediction model accuracy.	Demonstrate the potential of data mining techniques in predicting obesity and can be used as a basis for further research in this area. Information sources on healthcare policies and programs to prevent and treat obesity by identifying factors contributing to obesity.
Ali and Buti (2021)	Provide an overview of data mining's importance, different techniques, and their applications in the healthcare sector; disease diagnosis, prediction, and	Predictive analytics, clustering, association rule mining	Data mining can improve the accuracy and efficiency of diagnosis, treatment, and prevention of diseases, identify and manage risk factors, predict	Provide a broad overview of data mining healthcare applications without going into specific examples or case studies. Did not discuss the challenges and	Data mining plays a significant role in transforming healthcare by improving diagnosis, treatment, prevention of diseases, and optimizing resource allocation, if addressing the challenges and ethical considerations of using data mining in healthcare to

	prevention. Discuss potential benefits and associated challenges; high-quality data necessity and the data complexity issue.		disease outbreaks, and optimize healthcare resource allocation.	ethical considerations that arise when using data mining in healthcare, such as patient privacy and informed consent.	ensure patients' privacy and rights are protected.
Jamjoom (2021)	Using a real dataset, investigate the potential use of data mining techniques to predict the likelihood of a cesarean delivery operation. Uses a combination of techniques to make predictions and compares the performance of these techniques.	DT and ANN	The proposed model achieved high accuracy in predicting cesarean delivery. The importance of selecting suitable data mining methods and features to improve the accuracy of the predictive model.	Relied on a single dataset may limit the external validity of the findings. The small sample size affects the generalizability of the results.	A comprehensive dataset analysis, including preprocessing, feature selection, and model evaluation. Data mining techniques can be effectively used in healthcare to predict cesarean delivery operations, and further research using larger sample sizes and multiple datasets validate the results and improve the accuracy of the predictive model.
Abdullah et al. (2022)	Review of Bayesian deep learning in healthcare, its applications, including disease diagnosis, drug discovery, personalized medicine, and medical image analysis, and its challenges, such as the lack of large-scale labeled datasets and computational resources.	Bayesian deep learning	Bayesian deep learning has advantages over traditional deep learning approaches, such as better uncertainty quantification, model interpretability, and complex and high-dimensional data handling.	Did not provide a detailed analysis of specific studies or applications of Bayesian deep learning in healthcare. Did not discuss potential ethical and legal implications, such as issues related to data privacy and informed consent.	provides insights into the challenges of implementing Bayesian deep learning in healthcare and the need to address them for successful implementation. Bayesian deep learning could be used to improve the accuracy and reliability of healthcare systems, leading to better patient outcomes.
Suzan et al. (2022)	Review the current state of knowledge management (KM) in healthcare, focusing on big data mining and expert systems. Discuss various knowledge management techniques, their applications, the challenges, and the use of knowledge management to improve healthcare outcomes.	-	Provide healthcare organizations a wide range of KM benefits, including improved efficiency, better decision-making, and enhanced patient outcomes. Highlight several associated challenges, such as the need for specialized knowledge and expertise, cultural barriers, and the complexity of healthcare systems.	Lack of empirical research to support the claims made regarding the benefits of KM in healthcare. Did not provide a detailed analysis of the specific tools and techniques that can be used to implement KM in healthcare settings.	Provides a useful overview of KM's current state in healthcare and highlights the potential benefits and challenges benefiting healthcare organizations looking to implement KM strategies. The importance of a clear understanding of the goals and objectives of KM initiatives and the need for effective leadership, communication, stakeholders' collaboration, and selecting appropriate technologies to support KM activities, such as expert systems and data analytics tools.
Kaul et al. (2022)	This paper discusses the application of deep learning to improve healthcare outcomes and the challenges associated with applying deep learning in healthcare. Provide examples of successful applications of deep learning in	CNN, RNN, and deep belief networks.	Deep learning detects abnormalities in medical imaging, assists radiologists in making accurate diagnoses, predict patient outcomes, identify high-risk patients, improves people's health management using EHRs, and	Did not present any original research or empirical data. Instead, it synthesizes and summarizes existing research on deep learning in healthcare. Focus mainly on the technical aspects of deep learning without discussing the social, ethical, and	Deep learning has great potential to improve the accuracy and efficiency of medical diagnosis and healthcare outcomes, with further research and development to overcome the limitations and challenges associated with its implementation.

	medical imaging, electronic health records (EHRs), and clinical decision support systems (CDSSs).		provides tailored advice based on patient data, which supports treatment decisions making.	legal implications of using this technology in healthcare settings.	
Santos-Pereira et al. (2022)	A systematic literature review identifies and analyzes the most relevant data mining tools used in the healthcare industry; 22 relevant tools. Then analyze tools based on their features, strengths, weaknesses, and applications to recommend their use and selection.	R, Python, Weka, KNIME, RapidMiner, SPSS	R, Python, Weka, KNIME, RapidMiner, and SPSS are common data mining tools in healthcare. The criteria for evaluating the tools include the ability to preprocess data, algorithms, visualization tools, scalability, and ease of use. The choice of tool depends on the needs and goals of the project.	Limited to articles published before the end of 2020 may not include more recent developments in the field. Findings cannot be generalizable to all healthcare settings or regions, as the tools used may vary depending on factors such as resource availability and cultural context.	The study provides insights into the most relevant data mining tools and their strengths and weaknesses to assess healthcare professionals making informed decisions when choosing a tool for their specific needs, which will improve patient outcomes, reduce costs, and optimize healthcare processes.
Subrahmanya et al. (2022)	A systematic literature review with case studies analysis. Discuss the role of data science in healthcare advancements, the benefits, limitations of its usage, such as improved patient outcomes and the need for ethical and legal frameworks. Present its applications, such as disease diagnosis, treatment prediction, and drug discovery.		Data science enables accurate and personalized diagnoses, improves treatment outcomes, and optimizes healthcare operations leading to healthcare advancement. Identify key applications, including DSSs, predictive analytics, natural language processing, and images analysis.	Relies heavily on case studies and expert opinions and does not include empirical evidence from randomized controlled trials or other rigorous research studies.	Healthcare professionals and organizations get insights into the potential benefits and challenges of implementing data science in healthcare. Healthcare organizations are recommended to invest in data science technologies and infrastructure and develop strategies to address the challenges associated with using these technologies.

The papers listed here are all related to data mining and its applications in healthcare. They cover various aspects of data mining, including techniques, tools, benefits, and policy implications. Overall, the papers provide a comprehensive overview of the current state of data mining in healthcare and highlight its potential to transform the healthcare industry.

These papers demonstrate that data mining has many potential applications in healthcare, including improving diagnosis, treatment, and overall healthcare management. Using data mining techniques requires a designated guideline for each one. For example, use a statistical method to detect the redundant and inappropriate attributes before employing classification techniques to reduce noise and outlier adverse effects on the model performance. In contrast, use feature selection methods to recognize the relevant and valuable attributes to enhance the model performance and accuracy. Furthermore, the performance of a model depends on the testing dataset. Thus, using a cross-validation method is necessary, i.e., using each dataset record for training and testing (Tomar & Agarwal, 2013). It is vital to minimize the semantic gap to ensure fetching meaningful patterns during the data sharing across distributed healthcare repositories. That will improve treatment effectiveness services and customer relationship management worldwide (Ahmad et al., 2015). A new beneficial application is personalized medicine, which tailors appropriate medical care to an individual using the unique physiological makeup and medical history of that individual. As a result, it will provide precise diagnoses, effective treatments with lower cost, and minimized side effects, for instance, genetics-driven personalization (Fichman et al., 2011).

However, implementing data mining techniques in healthcare is not without challenges, including privacy concerns, data quality issues, and the need for appropriate policies and regulations. Generally, data mining algorithms rely on patient data, and privacy concerns can arise when sharing and analyzing such sensitive information. Also, data mining algorithms are only as good as the data they analyze, and data quality issues can lead to inaccurate results. Moreover, some data mining algorithms are complex, and healthcare professionals may need more training to interpret the results accurately. In addition, data mining algorithms in healthcare may be subject to regulatory challenges, and healthcare professionals must ensure that they comply with relevant laws and regulations. Furthermore, data mining algorithms should be used to supplement clinical judgment, not replace it. Overreliance on algorithms can lead to misdiagnosis and other issues.

Additionally, the choice of a data mining algorithm depends on the specific application and the characteristics of the data being analyzed. Therefore, it is essential to carefully consider the advantages and disadvantages of different algorithms and the potential biases and limitations of the data before selecting an appropriate algorithm for analysis.

8. Conclusion & Future Work

Healthcare is one of the sectors that will benefit significantly if data mining analysis is appropriately used to solve key search questions in the domain. Using data mining techniques in healthcare has various administrative, services, and medical care benefits. However, these algorithms' performances vary, and each has benefits and drawbacks. Additionally, many DM algorithms are used in healthcare for different applications and analysis purposes. Similarly, abundant applications like heart disease detection, obesity prediction, and cesarean delivery operations forecasting are available. Moreover, several common data mining tools are used to support medical data analysis. Despite all the benefits data mining brings to the healthcare sectors, DM still faces many challenges to achieve the best results. This study aims to review comprehensively data mining techniques implementation and applications in the healthcare domain, including its importance, widespread techniques, their advantages and disadvantages, applications, public tools, challenges of data mining analysis, and possible future directions to the benefit of data mining analytics in this industry optimally.

Future directions:

There is a continuous search for better options to improve the quality of medical services and treatment. One way is to capitalize on technological innovations for better future development. Also, the transformation of patients, physicians, and healthcare providers' vision of care delivery is started. On top of that, the rise of data science and the introduction of many data mining applications allow health providers to offer better data mining analysis for customized customer services (Subrahmanya et al., 2022). Additionally, data mining applications have tremendous potential and effectiveness. Therefore, many elements must be considered to direct the future and enhance the benefits of data mining outcomes. These considerations will be summarized in the following lines. First, use the best method to capture, store, prepare, and extract data. Second, it is vital to standardize the data-sharing method across the institute and the clinical vocabulary. Consequently, data integration and text mining must be considered since healthcare data contains only quantitative data, such as physicians' notes or clinical records. A possible direction is to examine the potential of data mining on digital diagnostic images (Fichman et al., 2011). In the absence of international legislation, it is also crucial to have legal regulations in place to protect data privacy, prevent misuse, and preserve ethical aspects of data mining. Similarly, educational institutions must consider the data mining major due to the multidisciplinary complexity surrounding this field (Sen & Khandelwal, 2018). Another element to work on is changing insurance reimbursement models to find suitable approaches to reasonably repay doctors and patient satisfaction metrics to characterize better care. In addition, researchers need to decide on the appropriate data mining tools and methods that improve the overall performance of medical care services. Also, the identification of risks and solutions for mitigation is an essential requirement to reduce settlement rates (Elezabeth et al., 2018).

A recommendation for further research would be to investigate the potential benefits and challenges of implementing data mining techniques in specific healthcare settings, such as hospitals or clinics, and to develop appropriate policies and regulations to support these efforts. Additionally, the research could focus on developing new data mining techniques and tools specifically designed for healthcare applications.

Acknowledgement

The authors would like to thank Deanship of scientific research in King Saud University for supporting this research.

Compliance with Ethical Standards

Competing Interests: The authors declare that they have no conflicts of interest.

References

Abdullah, A., Hassan, M., & Mustafa, Y. (2022). A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges. *IEEE Access*, *10*, 1-1. https://doi.org/10.1109/ACCESS.2022.3163384

Ahmad, P., Qamar, S., & Rizvi, S. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*, *120*, 38-50. https://doi.org/10.5120/21307-4126

Amin, M., & Ali, A. (2017). *Application of Multilayer Perceptron (MLP) for Data Mining in Healthcare Operations*.

Dwivedi, A., Rehman, K., Ghosh, M., & Chandan, R. (2018). Data Mining Algorithms in Healthcare. *International Journal of Computer Applications*, *180*, 26-31. https://doi.org/10.5120/ijca2018916901

Elezabeth, L., Mishra, V. P., & Dsouza, J. (2018). *The Role of Big Data Mining in Healthcare Applications*. https://doi.org/10.1109/ICRITO.2018.8748434

Feblowitz, J., Wright, A., Singh, H., Samal, L., & Sittig, D. (2011). Summarization of clinical information: A conceptual model. *Journal of biomedical informatics*, *44*, 688-699. https://doi.org/10.1016/j.jbi.2011.03.008

Fichman, R., Kohli, R., & Krishnan, R. (2011). Editorial Overview: The Role of Information Systems in Healthcare: Current Research and Future Trends. *Information Systems Research*, *22*, 419-428. https://doi.org/10.2307/23015587

Helén, I. (2022). *Tales for opening a techno-future: Narrating the promise of data mining in healthcare*.

Jamjoom, M. (2020). *Data Mining in Healthcare to Predict Cesarean Delivery Operations using a Real Dataset*. https://doi.org/10.5220/0010366700200026

Jothi, N., Abdul Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare – A Review. *Procedia Computer Science*, *72*, 306-313. https://doi.org/10.1016/j.procs.2015.12.145

Kaul, D., Raju, H., & Tripathy, B. K. (2022). Deep Learning in Healthcare. In (pp. 97-115). https://doi.org/10.1007/978-3-030-75855-4_6

Krumholz, H. (2014). Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health affairs (Project Hope)*, *33*, 1163-1170. https://doi.org/10.1377/hlthaff.2014.0053

Ma, G., Zhang, L., Cui, G., & Cheng, Y. (2019). Design of Medical Examination Data Mining System Based on Decision Tree Model. *Journal of Physics: Conference Series*, *1237*, 022022. https://doi.org/10.1088/1742-6596/1237/2/022022

Raju, S., Joshi, A., Choudhury, T., & Sabitha, S. (2020). Data Mining in Healthcare and Predicting Obesity. *Advances in Intelligent Systems and Computing*, *1090*, 877-888. https://doi.org/10.1007/978-981-15-1480-7_82

Roski, J., Bo-Linn, G., & Andrews, T. (2014). Creating Value In Health Care Through Big Data: Opportunities And Policy Implications. *Health affairs (Project Hope)*, *33*, 1115-1122. https://doi.org/10.1377/hlthaff.2014.0147

Sadiku, M., Eze, K., & Musa, S. (2018). Data Mining in Healthcare. *International Journal of Advances in Scientific Research and Engineering*, *4*, 90-92. https://doi.org/10.31695/IJASRE.2018.32881

Santos-Pereira, J., Gruenwald, L., & Bernardino, J. (2022). Top data mining tools for the healthcare industry. *Journal of King Saud University - Computer and Information Sciences*, *34*(8, Part A), 4968-4982. https://doi.org/https://doi.org/10.1016/j.jksuci.2021.06.002

Sen, I., & Khandelwal, K. (2018). *DATA MINING IN HEALTHCARE*. https://doi.org/10.13140/RG.2.2.22189.38887

Shafique, U., Majeed, F., & Qaiser, H. (2015). Data Mining in Healthcare for Heart Diseases. *International Journal of Innovation and Applied Studies*, *10*, 2028-9324.

Shirazi, S., Baziyad, H., & Karimi, H. (2019). An Application-Based Review of Recent Advances of Data Mining in Healthcare. *Journal of Biostatistics and Epidemiology*, *5*, 235-245. https://doi.org/10.18502/jbe.v5i4.3864

Sondakh, D., Klabat, U., Mononutu, J., & Utara, M. (2017). Data Mining for Healthcare… Data Mining for Healthcare Data: A Comparison of Neural Networks Algorithms. *CogITo Smart Journal*, *3*, 10-19. https://doi.org/10.31154/cogito.v3i1.40.10-19

Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. M. Z., Paul, R., Smriti, K., Naik, N., & Somani, B. K. (2022). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science (1971 -)*, *191*(4), 1473-1483. https://doi.org/10.1007/s11845-021-02730-z

Sura I. Mohammed ALI, R. H. B. (2021). Data Mining In Healthcare Sector. *MINAR International Journal of Applied Sciences and Technology*, *3*(2), 87-91. https://doi.org/10.47832/2717-8234.2-3.11

Suzan Katamoura, M. S. A. (2022). A Review On Implementing Knowledge Management In

Healthcare. *Turkish Journal of Computer and Mathematics Education*, *13*(2), 958-969. https://doi.org/10-20220510

Tekieh, M., & Raahemi, B. (2015). *Importance of Data Mining in Healthcare: A Survey*. https://doi.org/10.1145/2808797.2809367

Tharindu Fernando, H. G., Simon Denman, Sridha Sridharan, Clinton Fookes. (2020). *Deep Learning for Medical Anomaly Detection -- A Survey*. arXiv. https://doi.org/10.48550/arXiv.2012.02364

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. Bio-Science and Bio-Technology,