International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-13 Issue-2, May 2023

Innovations in Healthcare Analytics: A Review of Data Mining Techniques

Shikha Bhardwaj, Neeraj Bhargava, Ritu Bhargava

Abstract: ***This review article provides an overview of the current state of data mining applications in healthcare, including case studies, challenges, and future directions. The article begins with a discussion of the role of data mining in healthcare, highlighting its potential to transform healthcare delivery and research. It then provides a comprehensive review of the various data mining techniques and tools that are commonly used in healthcare, including predictive modelling, clustering, and association rule mining. The article also discusses several key challenges associated with data mining in healthcare, including data quality, privacy, and security, and proposes potential solutions to address these issues. Finally, the article concludes with a discussion of the future directions of data mining in healthcare, highlighting the need for continued research and development in this field. The article emphasises the importance of collaboration between healthcare providers, data scientists, and policymakers to ensure that data mining is used ethically and effectively to improve patient outcomes and support evidence-based decision-making in healthcare.***

Keywords: Data Mining; Health Care; SVM; ANN

I. INTRODUCTION

Data mining, also known as knowledge discovery in databases (KDD), is the process of extracting valuable and actionable insights from large datasets. With the increasing amount of data being generated across various domains, including healthcare, finance, and retail, data mining has become a crucial tool for informed decision-making and strategic development. Data mining techniques enable organisations to identify hidden patterns, relationships, and trends in their data that can be used to make informed decisions, improve efficiency, and gain a competitive advantage [1, 2]. The process of data mining involves several steps, including data preprocessing, feature selection, model building, and evaluation. Data pre-processing involves cleaning, transforming, and integrating data from multiple sources to prepare it for analysis and interpretation. Feature selection is identifying the most relevant features or variables in the data that are likely to impact the outcome.

Manuscript received on 15 April 2023 | Revised Manuscript received on 20 April 2023 | Manuscript Accepted on 15 May 2023 | Manuscript published on 30 May 2023.

*Correspondence Author(s)
Shikha Bhardwaj*, Department of Computer Science, Mahatma Jyoti Rao Phoole University, Jaipur (R.J), India E-mail: shikhabhardwaj90390@gmail.com, ORCID ID: https://orcid.org/0009-0009-4759-3919
Prof. Neeraj Bhargava, Department of Computer Science, M.D.S University, Ajmer (R.J), India E-mail: profneerajbhargava@gmail.com, ORCID ID: https://orcid.org/0000-0002-1824-499X
Dr. Ritu Bhargava, Sophia Girls' College, Ajmer (R.J), India Email: drritubhargava92@gmail.com, ORCID ID: https://orcid.org/0000-0001-6629-9402

Model building involves selecting an appropriate algorithm or method to extract patterns and relationships from the data. Finally, evaluation consists of measuring the accuracy and performance of the model and validating its usefulness in real-world applications. However, it has gained significant attention in the healthcare sector in recent years. The healthcare industry generates massive amounts of data, and data mining can be utilised to extract valuable insights that improve patient outcomes, reduce costs, and enhance efficiency [3].

This review article examines the diverse applications of data mining in the healthcare sector. We will explore the various techniques employed in data mining, the challenges associated with applying data mining in healthcare, and the potential benefits of implementing data mining in this context. Finally, we will discuss some case studies where data mining has been successfully used in the healthcare sector.

II. TECHNIQUES USED IN DATA MINING.

Data mining involves using various techniques to extract valuable insights from data. Figure 1 represents the classification of data mining techniques.

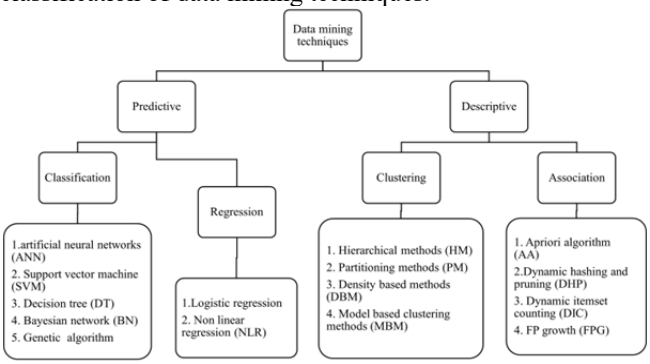

Figure 1. Classification of Data Mining Techniques

The following are some of the techniques commonly used in data mining.

2.1 Predictive Modelling

Predictive modelling is a technique used to predict outcomes based on historical data. In the healthcare sector, predictive modelling can be used to predict patient readmissions, disease progression, treatment response, and other outcomes. The technique involves building a model that can learn from historical data and make predictions on new data. Several classifications of algorithms are used for predictive modelling, each with its advantages and disadvantages. Predictive modelling can be further divided into algorithms based on classification and regression techniques, as shown in Figure 1. Some of healthcare's most used classification algorithms include decision trees, random forests, support vector machines, and artificial neural networks.

2.1.1 Regression

Regression analysis is used to identify the relationship between two or more variables. In healthcare, regression analysis determines the factors that influence patient outcomes.

2.1.1.1 Logistic Regression:

Logistic regression is a classification algorithm used to model the probability of an event occurring, such as readmission or mortality, based on one or more predictor variables. The technique is based on the logistic function, which maps any input to a value between 0 and 1, representing the probability of the event occurring. Logistic regression is a straightforward and efficient method, making it a popular choice for healthcare applications. In a research study, it was investigated that in most cases, only a small portion of the data was sufficient to generate a model with performance comparable to that of one generated using the entire dataset, but with significantly less model complexity [4].

2.1.2 Classification:

Classification involves the use of algorithms to categorise data into predefined groups. In healthcare, classification is used to diagnose diseases, predict patient outcomes, and identify risk factors associated with specific conditions.

2.1.2.1 Decision Trees

Decision trees are a classification algorithm used to model decision-making processes. In healthcare, decision trees can be used to predict patient outcomes based on clinical characteristics. The technique involves building a tree-like model that splits the data based on the most informative predictor variables. Decision trees are easy to understand and interpret, making them a popular choice for healthcare professionals.

Jegelevicius and Lukoševicius [5] applied a decision support system for the differential diagnosis of intraocular malignancies utilising information from ultrasound-captured pictures of the eyes. Using the See5.0/C5.0 data mining system, a predictive modelling method application for decision tree generation was demonstrated.

2.1.2.2 Random Forests

Random forests are an ensemble classification algorithm that combines multiple decision trees to improve performance. In healthcare, random forests can be used to predict patient outcomes based on many predictor variables. The technique involves building multiple decision trees on random subsets of the data and then combining the results. Random forests are robust and can handle noisy data, making them a popular choice for healthcare applications.

2.1.2.3 Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning algorithm that has been widely used in healthcare to predict clinical outcomes, such as disease diagnosis and patient survival. The algorithm works by identifying a hyperplane that separates the data into two or more classes, maximising the margin between the hyperplane and the nearest data points. SVM can handle both linear and non-linear data, making it suitable for a wide range of healthcare applications.

The implementation of SVMs in healthcare involves several steps. First, the data must be collected and pre-processed to ensure that it is ready for analysis. This may include removing missing data, standardising variables, and transforming variables as needed. Next, the data is divided into training and testing sets, and the SVM model is trained using the training set. The optimal hyperplane is found by optimising the SVM objective function, which involves minimising the classification error and maximising the margin between the classes.

Once the SVM model has been trained, it can be used to predict the class labels of new data points. The model is evaluated using the testing set, and its performance is assessed using metrics such as accuracy, sensitivity, and specificity. The model can also be visualised using techniques such as decision boundaries and feature importance plots, which help interpret the results and identify the key variables driving the classification.

In healthcare, SVM has been used to predict disease diagnosis, patient survival, and treatment response. For example, SVM has been used to indicate the risk of heart disease based on patient demographics, clinical characteristics, and biomarkers. SVM has also been used to indicate the response to cancer treatment based on tumour characteristics and patient demographics. These predictions can then be used to guide treatment decisions and improve patient outcomes. One advantage of SVM is its ability to handle high-dimensional data with a relatively small number of samples. This is important in healthcare, where datasets can be complex and sample sizes may be limited. SVM also has a strong theoretical foundation and can handle noisy data by using kernel functions to transform the data into a higher-dimensional space. However, SVM also has some limitations in healthcare. For example, it can be sensitive to the choice of kernel function, and selecting hyperparameters can be a challenging task. Additionally, the interpretability of SVM can be limited, which is a concern in healthcare where understanding the underlying factors driving predictions is crucial.

2.1.2.4 Artificial Neural Networks

Artificial neural networks (ANNs) are a classification algorithm inspired by the structure and function of the human brain. In healthcare, ANNs can be used to predict patient outcomes based on many predictor variables. The technique involves building a network of interconnected nodes that process information and make predictions. ANNs can handle complex relationships between variables and can learn from noisy data, making them a popular choice for healthcare applications. Hachesu et al. [6] implemented and compared the performance of three data mining techniques — SVM, ANN, and decision tree — to determine the length of stay for cardiac patients. The authors reported that all three algorithms can predict the length of stay to varying degrees of accuracy. The SVM had the best fit, according to the results. Patients with heart disease, high blood pressure, or lung or respiratory conditions had a clear trend for length of stay to be prolonged.

Retrieval Number: 100.1/ijsce.B36090513223
DOI: 10.35940/ijsce.B3609.0513223
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

2.1.2.5 Bayesian networks

Bayesian networks are a probabilistic graphical model that represents the conditional dependencies between variables in a system. In healthcare, Bayesian networks can be used to model complex relationships between clinical variables and predict patient outcomes. A Bayesian network consists of two components: nodes and edges. Nodes represent variables in the system, while edges represent the conditional dependencies between the variables. The direction of the edges represents the direction of causality in the system. For example, if variable A causes variable B, there will be an edge from A to B in the Bayesian network.

Bayesian networks are based on Bayes' theorem, which states that the probability of a hypothesis (H) given evidence (E) is proportional to the likelihood of the evidence given the hypothesis times the prior probability of the hypothesis. In other words, Bayesian networks use probabilities to calculate the possibility of an event occurring based on available evidence.

To build a Bayesian network, data is collected and used to estimate the probabilities of the variables and the conditional dependencies between them. This involves defining prior probabilities and updating them as new data becomes available. The network can then be used to make predictions about new data.

In healthcare, Bayesian networks can be used to predict patient outcomes based on clinical variables. For example, a Bayesian network can be used to indicate the likelihood of a patient developing a particular disease based on their age, sex, medical history, and other clinical variables. The network can also be used to identify the most critical variables in predicting the outcome. In a research study Mc McLachlana et al. [7] categorised and quantified the variety of medical diseases for which healthcare-related BN models have been suggested, as well as the methodological variations among the most prevalent medical conditions to which they have been applied.

One of the advantages of Bayesian networks is their ability to handle missing data and noisy data. They can also handle complex relationships between variables and can be updated as new data becomes available. However, Bayesian networks can be computationally intensive and require a large amount of data to build an accurate model.

2.2 Descriptive Modelling

Descriptive modelling is a set of data mining techniques that aims to summarise and describe data in a valuable and informative way. It is used to understand patterns and relationships within the data, identify trends, and explore the characteristics of the data.

2.2.1 Clustering

Clustering is a technique used to group similar data points together. In the healthcare sector, clustering can be used to group patients based on their clinical characteristics or to identify patterns in large datasets. The technique involves partitioning data points into groups (clusters) based on their similarity, such that data points within the same cluster are more similar to each other than data points in other clusters. Several clustering techniques are used in data mining. Here are some of the most common ones [8].

2.2.1.1 *K-means clustering*

K-means clustering is a popular unsupervised machine learning algorithm used in healthcare to identify groups of patients with similar clinical characteristics. The algorithm works by iteratively partitioning the data into K clusters, where K is a user-defined number of clusters. The goal is to minimise the within-cluster sum of squares, which is a measure of the sum of the squared distances between each data point and the centroid of its assigned cluster.

In healthcare, k-means clustering has been utilised to identify patient subgroups based on clinical characteristics, including age, gender, comorbidities, and biomarkers. For example, k-means clustering has been used to identify subgroups of patients with similar disease severity or risk factors for chronic diseases such as diabetes or cardiovascular disease. This information can then be used to tailor interventions and treatments to the specific needs of each patient, improving patient outcomes and overall healthcare efficiency.

Implementing k-means clustering in healthcare involves several steps, including data collection, data cleaning and preprocessing, and determining the optimal number of clusters, among others.

The first step is to collect the data that will be used to identify patient subgroups. This may include patient demographic information, clinical characteristics such as lab results and comorbidities, and health behaviour data. Once the data has been collected, it needs to be cleaned and pre-processed to ensure that it is ready for analysis. This may involve removing missing data, standardising variables, and transforming variables as needed. The next step is to determine the optimal number of clusters to use in the analysis. This can be achieved using various methods, including the elbow method, silhouette analysis, and hierarchical clustering. Once the number of clusters has been determined, k-means clustering can be implemented. This involves randomly selecting K initial centroids and assigning each data point to the cluster with the closest centroid. The centroids are then updated by calculating the mean of each cluster and reassigning the data points to their closest centroid. This process is repeated until convergence, which is typically defined as when the centroids no longer move significantly. The final step is to validate the results of the clustering analysis. This may involve visualising the clusters to ensure that they are clinically meaningful and actionable. It is also crucial to evaluate the stability of the clusters and their reproducibility across various samples.

When implementing k-means clustering in healthcare, it is essential to carefully consider the clinical context and ensure that the clustering results are clinically meaningful and actionable. This may involve incorporating domain knowledge and expert input into the analysis, as well as validating the results with clinical stakeholders to ensure accuracy and relevance. By carefully implementing and validating k-means clustering, healthcare providers can identify patient subgroups with similar clinical characteristics and tailor interventions and treatments to the specific needs of each patient, improving patient outcomes and overall healthcare efficiency.

Zheng et al. [9] implemented a hybrid of the support vector machine and K-means (K-SVM) algorithms to extract important information and identify the tumour. The hidden patterns of benign and malignant tumours were recognised separately using the K-means method. The training model calculated and treated each tumour's participation in these patterns as a new feature. The new classifier was then created using an SVM to distinguish between the incoming tumours. When tested on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the University of California, Irvine machine learning repository, the suggested methodology improved the accuracy to 97.38% based on 10-fold cross-validation.

2.2.1.2 Hierarchical clustering

This technique involves building a tree-like structure (dendrogram) of clusters by recursively merging or splitting clusters. There are two types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering starts with each data point as a separate cluster. It merges them iteratively based on similarity, while divisive clustering starts with all data points in a single cluster and recursively divides them into smaller clusters.

2.2.1.3 Density-based clustering

Density-based clustering methods are a popular approach in healthcare for identifying clusters of patients with similar clinical characteristics or health behaviours. These methods involve identifying areas of high density in the data and grouping together data points that are close to each other in terms of their distance or similarity. One of the key advantages of density-based clustering methods is that they can identify clusters of any shape or size and can handle noise and outliers in the data.

One example of density-based clustering methods in healthcare is the use of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to identify subgroups of patients with similar clinical characteristics. DBSCAN works by identifying areas of high density in the data and grouping together data points that are close to each other. The algorithm defines clusters as areas of high density, and outliers as data points that do not belong to any of these clusters. This approach has been used in healthcare to identify subgroups of patients with similar clinical characteristics, such as comorbidities or risk factors for chronic diseases.

Density-based clustering methods have several advantages over traditional clustering methods. Firstly, they can handle noise and outliers in the data, which is common in healthcare datasets. Secondly, they can identify clusters of any shape or size, which allows for more flexible clustering solutions. Finally, they do not require the number of clusters to be specified in advance, which can be challenging in healthcare datasets where the optimal number of clusters may not be known. Overall, density-based clustering methods can be used for identifying clusters of patients with similar clinical characteristics or health behaviours in healthcare. By identifying these clusters, healthcare providers can tailor their interventions and treatments to the specific needs of each patient, improving patient outcomes and overall healthcare efficiency.

2.2.1.4 Fuzzy clustering

This technique allows data points to belong to more than one cluster with varying degrees of membership. Fuzzy clustering is functional when data points are not separable and belong to multiple groups simultaneously [10]. Fuzzy C-means (FCM) is a popular algorithm for fuzzy clustering [11], [12].

The performance of all of the main fuzzy clustering algorithms, including PCM, KT2FCM, PFCM, FCM, FCM-σ, T2FCM, IFCM-σ, KIFCM-σ-σ, NC, CFCM, IFCM, KIFCM, and DOFCM, has been thoroughly examined and experimentally analysed in the review work by Gosain and Dahiya[13].

Albayrak et al. [14], clustered the thyroid gland data using the fuzzy c-means and hard k-means algorithms. According to the authors, the fuzzy c-means algorithm outperformed the hard k-means algorithm in medical diagnostic systems.

2.2.1.5 Model-based clustering methods

Model-based clustering methods are a popular approach in healthcare for identifying patient subgroups with similar clinical characteristics. These methods involve fitting a statistical model to the data to identify clusters or groups of patients based on shared features or characteristics. One of the key advantages of model-based clustering methods is that they enable the identification of complex patterns and relationships in the data that may not be readily apparent using traditional clustering methods. One example of model-based clustering methods in healthcare is the use of Gaussian mixture models (GMMs) to identify patient subgroups with similar clinical characteristics. GMMs assume that the data is generated by a mixture of Gaussian distributions, with each distribution representing a different patient subgroup. The parameters of the GMM, including the number of clusters and the mean and variance of each cluster, are estimated using maximum likelihood estimation. This enables the identification of patient subgroups based on clinical characteristics, such as age, gender, comorbidities, and biomarkers. Model-based clustering methods have several advantages over traditional clustering methods. Firstly, they allow for the identification of complex patterns and relationships in the data that may not be easily detected using conventional methods. Secondly, they provide a probabilistic framework for clustering, which allows for uncertainty in the clustering process to be accounted for. Finally, they can handle missing or incomplete data, which is common in healthcare datasets. Overall, model-based clustering methods are a powerful tool for identifying subgroups of patients with similar clinical characteristics or health behaviours in healthcare. By identifying these subgroups, healthcare providers can tailor their interventions and treatments to the specific needs of each patient, improving patient outcomes and overall healthcare efficiency.

2.2.2 Association Rule Mining

Association rule mining is a popular data mining technique that involves discovering interesting relationships, patterns, and associations among large datasets.

This technique is widely used in market basket analysis, web usage mining, and other domains where patterns in large datasets need to be identified and analysed.

Association rule mining involves finding correlations between different items or attributes in a dataset. The objective is to identify rules that show how the presence or absence of one item can be used to predict the presence or absence of another item. The rules are expressed in the form of "if-then" statements, where the antecedent is the condition, and the consequent is the predicted outcome.

Several algorithms are used in association rule mining, including Apriori, Eclat, and FP-growth. Apriori is the most popular algorithm and is widely used in practice.

2.2.2.1 Apriori algorithm

The Apriori algorithm is one of the most widely used algorithms for data mining and association rule learning. It employs a bottom-up approach, generating candidate item sets of increasing size and then pruning those that do not meet the minimum support threshold. The minimum support threshold is a user-defined parameter that specifies the minimum number of transactions in which an itemset must appear to be considered frequent. The algorithm consists of two main steps: candidate generation and candidate pruning. In the candidate generation step, the algorithm generates a set of candidate item sets of size k+1 by joining frequent item sets of size k. In the candidate pruning step, the algorithm prunes the candidate item sets that do not meet the minimum support threshold. This process is repeated until no more frequent item sets can be generated.

The Apriori algorithm is an efficient algorithm for mining large datasets because it avoids generating all possible item sets by using the principle of monotonicity. This principle states that any subset of a frequent itemset must also be frequent. Thus, the algorithm only needs to generate frequent item sets, and not all possible item sets.

The Apriori algorithm has numerous applications in data mining, including market basket analysis, recommendation systems, and web mining. Market basket analysis is one of the most common applications, where the algorithm is used to find associations between items that are frequently purchased together. Recommendation systems utilise algorithms to suggest items to customers based on their past purchase history. In web mining, the algorithm is used to find patterns in web logs, such as which pages are frequently visited together.

Overall, the Apriori algorithm is a powerful tool for discovering patterns and associations in large datasets. Its simplicity and efficiency make it a popular choice for data mining tasks, and its broad range of applications make it a versatile tool for both businesses and researchers.

2.2.2.2 FP growth algorithm

The FP-growth algorithm is a popular data mining algorithm used for identifying frequent patterns in datasets. It is a bottom-up approach that generates a compressed representation of the dataset in the form of a frequent pattern (FP) tree, which is then used to mine frequent itemsets. This algorithm is particularly useful for large datasets because it avoids the expensive generation of candidate itemsets that are necessary in other frequent pattern mining algorithms, such as the Apriori algorithm.

The FP growth algorithm works by first constructing an FP tree from the dataset. The tree is built by scanning the dataset once to identify all frequent items, sorting them in descending order of frequency, and then creating a tree structure that represents the frequency of each item set. This tree structure allows the algorithm to efficiently find all frequent itemsets in the dataset, without the need to generate and test candidate item sets.

Once the FP tree is constructed, the algorithm recursively finds all frequent item sets by traversing the tree and collecting all item sets that meet the minimum support threshold. This process is repeated for each item in the tree, and the resulting frequent item sets are stored in a list.

The FP growth algorithm has several advantages over other frequent pattern mining algorithms. Firstly, it is very efficient, especially for large datasets, as it avoids generating candidate itemsets. Secondly, the compressed representation of the dataset in the form of an FP tree makes it easier to mine frequent itemsets, as it reduces the number of database scans required. Finally, the algorithm is highly scalable, as it can easily handle datasets with millions of transactions and thousands of items.

The FP-growth algorithm is widely used in various applications, including market basket analysis, bioinformatics, and social network analysis. In market basket analysis, the algorithm is used to identify which items are frequently purchased together. In bioinformatics, it is used to identify frequent patterns in genetic data. In social network analysis, the algorithm can be used to identify frequent patterns in social interactions between users.

III. CASE STUDIES

Numerous case studies demonstrate the potential of data mining in healthcare. The following are some examples of how data mining has been used in the healthcare sector:

Predicting readmissions: Researchers developed a data mining algorithm that could identify patients at high risk of being readmitted to the hospital within 30 days of discharge. The algorithm analysed electronic health records to identify patients with certain risk factors such as previous hospitalisations, chronic diseases, and length of stay. This information was then used to develop a risk score for each patient. The algorithm was able to predict readmissions with a high degree of accuracy, allowing healthcare professionals to intervene early and prevent readmissions [15].

Diagnosing breast cancer: In this study, ten machine learning (ML) algorithms were evaluated using the Wisconsin Diagnostic Breast Cancer dataset. These algorithms included decision trees, linear discriminant analysis, passive-aggressive, forests of randomised trees, logistic regression, gradient boosting, naive Bayes, nearest centroid, SVM, and perceptrons. After compiling the data, the authors evaluated performance and contrasted the various classification methods. With an F1 score of 96.77%, the gradient boosting approach outperformed all other techniques [16].

Wang et al. [17] implemented the logistic regression and decision tree algorithm to investigate the improvements of the

survivability prognosis of breast cancer based on the experimental results.

Eye disorders in young adults: This research presents an association rule-based Apriori data mining approach for a web-based hospital information management system (HIMS) that extracts common patterns from patient data related to eye disorders [18].

IV. BENEFITS OF USING DATA MINING IN HEALTHCARE

The use of data mining in healthcare has the potential to provide numerous benefits. The following are some of the potential benefits of using data mining in healthcare:

Data mining can be utilised to develop predictive models that identify patients at risk of developing specific health conditions. This allows healthcare professionals to intervene early and provide preventive care, improving patient outcomes. Data mining can be used to identify inefficiencies in healthcare delivery and to optimise resource allocation. This can lead to cost savings and a more efficient healthcare system. Data mining can help to develop personalised treatment plans for patients based on their characteristics. This can lead to better treatment outcomes and improved patient satisfaction.

With the help of Data mining, it is easy to identify patterns and trends in healthcare data that inform medical research. This can lead to the development of new treatments and therapies. In addition to the above applications, text mining extracts information from unstructured data sources, including clinical notes, medical literature, and social media.

V. CHALLENGES OF USING DATA MINING IN HEALTHCARE

The use of data mining in healthcare presents its challenges. The following are some challenges that must be addressed when implementing data mining in healthcare. The quality of the data used in data mining is critical. Healthcare data is often complex, heterogeneous, and incomplete. Therefore, data cleaning and pre-processing are essential to ensure the data is accurate and reliable. Furthermore, healthcare data is highly sensitive, and patient privacy must be protected. Data mining algorithms must be designed to protect patient privacy and prevent data breaches. Healthcare data is often stored in different formats and systems. Therefore, data mining algorithms must be designed to integrate data from multiple sources, providing a comprehensive view of patient health. Data mining algorithms can generate complex models that are difficult to interpret. Therefore, data mining results must be presented clearly and understandably to healthcare professionals [19].

VI. CONCLUSION AND FUTURE SCOPE

Data mining can transform healthcare delivery by providing valuable insights into patient health, optimising resource allocation, and improving patient outcomes. However, data mining in healthcare has challenges, including data quality, privacy and security, integration, and interpretability. Despite these challenges, numerous successful case studies demonstrate the potential of data mining in healthcare. As the volume of healthcare data continues to grow, data mining will play an increasingly important role in healthcare delivery, research, and innovation.

Despite the challenges faced, the future of data mining in healthcare is promising. Continued research and development in this field will be necessary to address these challenges and to ensure that data mining is used ethically and effectively to improve patient outcomes and support evidence-based decision-making in healthcare. Future directions include using artificial intelligence and machine learning to automate data mining processes and develop more accurate and personalised predictive models. Additionally, the use of data mining in precision medicine and personalised healthcare is an exciting area of research, as it has the potential to revolutionise the way healthcare is delivered and to improve patient outcomes.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is not subject to interpretation.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. Shekhar S, Xiong H (2008) Active Data Mining. Encycl GIS 10–10. [CrossRef]
2. Liao SH, Chu PH, Hsiao PY (2012) Data mining techniques and applications - A decade review from 2000 to 2011. Expert Syst Appl 39:11303 11311. [CrossRef]
3. Jothi N, Rashid NA, Husain W (2015) Data Mining in Healthcare - A Review. Procedia Comput Sci 72:306–313. [CrossRef]
4. John LH, Kors JA, Reps JM, et al (2022) Logistic regression models for patient-level prediction based on massive observational data: Do we need all data? Int J Med Inform 163:104762. [CrossRef]
5. Jegelevičius D, Lukoševičius A, Paunksnis A, Barzdžiukas V (2002) Application of Data Mining Technique for Diagnosis of Posterior Uveal Melanoma. Informatica 13:455–464
6. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F (2013) Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc Inform Res 19:121 129. [CrossRef]
7. McLachlan S, Dube K, Hitman GA, et al (2020) Bayesian networks in healthcare: Distribution by medical condition. Artif Intell Med 107:101912. [CrossRef]
8. Berkhin P (2006) A survey of clustering data mining techniques BT - Grouping Multidimensional Data. Group Multidimensional Data 25–71 [CrossRef]
9. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 41:1476 1482. [CrossRef]
10. Fernandez-Basso C, Gutiérrez-Batista K, Morcillo-Jiménez R, et al (2022) A fuzzy-based medical system for pattern mining in a distributed environment: Application to diagnostic and co-morbidity. Appl Soft Comput 122:108870. [CrossRef]

11. Hong TP, Lee YC (2008) An overview of mining fuzzy association rules. Stud Fuzziness Soft Comput 220:397 410. [CrossRef]
12. Hong TP, Lin KY, Wang SL (2003) Fuzzy data mining for interesting generalised association rules. Fuzzy Sets Syst 138:255 269. [CrossRef]
13. Gosain A, Dahiya S (2016) Performance Analysis of Various Fuzzy Clustering Algorithms: A Review. Procedia Comput Sci 79:100–111. [CrossRef]
14. Simhachalam B, Ganesan G (2014). Possibilistic fuzzy C-means clustering on medical diagnostic systems. Proc 2014 Int Conf Contemp Comput Informatics, IC3I 2014 1125–1129. [CrossRef]
15. Mohanty SD, Lekan D, McCoy TP, et al (2022) Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare. Patterns 3:100395. [CrossRef]
16. Kadhim RR, Kamil MY (2023) Comparison of machine learning models for breast cancer diagnosis. IAES Int J Artif Intell 12:415–421. [CrossRef]
17. Wang KJ, Makond B, Wang KM (2013) An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. BMC Med Inform Decis Mak 13:. [CrossRef]
18. Gulzar K, Ayoob Memon M, Mohsin SM, et al (2023) An Efficient Healthcare Data Mining Approach Using Apriori Algorithm: A Case Study of Eye Disorders in Young Adults. Information 14:1 14. [CrossRef]
19. R. Pallavi Reddy (2020) A Review on Data Mining Techniques and Challenges in the Medical Field. Int J Eng Res V9:329–333. [CrossRef]

AUTHORS' PROFILES

Shikha Bharadwaj is a research scholar in the Department of Computer Science, Mahatma Jyoti Rao Phoole University, Jaipur (India). She is currently conducting her research in the field of data mining.

Prof. Neeraj Bhargava, working as Professor at M.D.S University, Ajmer. He is the Head of the Department of Computer Science and School of Engineering and Systems Science at MDS University, Ajmer. He has over 26 years of teaching experience and has guided numerous research projects throughout his career. He has been a prominent figure in teaching and research, and his papers have had a significant impact on young researchers in India and abroad.

Dr. Ritu Bhargava is working as a Lecturer at Sophia Girls' College, Ajmer. She has been a senior academic and a prominent faculty member of Computer Science. She has been teaching at many government and private firms as a visiting faculty member.