Predictive Modeling in Healthcare: A Survey of Data Mining Applications

Saja Ataallah Muhammed[1], Laith R. Flaih[2]

[1]Department of Computer Science & Information Technology, Salahaddin University-Erbil, Kurdistan Region, Iraq
[2]Department of Computer Science, Cihan University-Erbil, Kurdistan Region, Iraq

***Abstract*— Healthcare domain is the one that has taken responsibility of human well-being which is aligned with the third sustainable goal of the United Nations (UN). Lately, the use and utilizing of Artificial Intelligence (AI) in healthcare services has been increasing. Data mining and machine learning became very significant in the contemporary era. They are also used in the healthcare domain towards solving a variety of issues from disease diagnosis and classifying to risk assessment and survival prediction. Since healthcare services could be improved with predictive modelling, this paper investigates on predictive modelling using data mining in healthcare domain. It is a systematic review on predictive modelling in healthcare using data mining. This paper provides insights pertaining to the literature on the usage of predictive modeling using data mining to address problems in healthcare domain. It also provides important research gaps triggering possible research in the future. The review article also highlights the key advantages and disadvantages of the methods being followed and adopted by different studies in the field of study, focusing on how data mining techniques combined with other methods and algorithms contribute getting amazing results.**

***Index Terms*— Predictive Modelling, Data Mining, Healthcare, Machine Learning**

I. INTRODUCTION

"We are living in the Information Age", is a popular saying, since the mid-20th century when the traditional industry has been moved toward a revolutionary industry that is centered around technology (Hoover, 2006). but in the 21st century with the explosion of social media, mobile, and Internet of Things(IOT), we realized the fact that we are living in a data age, data generated by different sources is increasing constantly and rapidly, and tera or petabytes of data are generated due to the society computerizing, according to IDC, by 2025, it's predicted there will be 163 zettabytes of data(Reinsel,2017). News, medicine, commercial or business social media, etc. have caused explosive growth in the data amount available nowadays. Besides computerizing, the development of powerful computers, vast storage mechanisms and devices, sensing and data collection, and publication tools has played a crucial role in overwhelming the world with data but unfortunately lack to information necessary for decision-making. The transition from the information age to the data age has led to understanding complex relationships between data, and fast decision-making (Han J. Pei., 2023) Real-world data with high dimensionality or a large number of attributes or features, become no more processable and analyzable by standard statistical techniques that could only analyze data with a small number of dimensions. Data mining techniques have been used to analyze these data and try to find the "nuggets", meaningful insights or extract important knowledge from vast amounts of data with high dimensionality. (Han J. Pei., 2023) Data mining is the process that aims to find knowledge, interesting patterns, and nuggets from large amounts of data and it started in 1990. Opposite of statistical analysis that deals with small to moderate amounts of data and tries to analyze, describe, find relations and associations among data attributes. Data mining helps healthcare specialists, enterprise decision-makers, and experts in the process of decision-making to improve profit. Other techniques exist for instance machine learning which aims to build machines or models that learn from data to solve real-world problems. These techniques have been used for predictive modeling. (Han J. , 2023) Predictive

DOI:http://doi.org/10.24086/cocos2024/paper.1083

modeling is the process of predicting future behavior or values or estimations based on historical data using some statistical and mining techniques that are not possible to be performed manually. In healthcare field the use of data mining or even machine learning for building predictive modeling helped in making well-informed decisions, and fruitful estimates (Dep et. al., 2022). In 2019 during the COVID-19 pandemic, many estimation and predictive models have been built since then to predict the likelihood of positive, negative, and severe cases. Mining, Statistics, and data analysis played crucial roles back to that time not only to categorize but also to monitor the disease outbreak; therefore, making decisions to limit the spread and infection rate.(Rane et. al., 2023) Data collection and preprocessing are two essential steps in mining healthcare data, since healthcare data are available in tremendous amounts around the world, and different sources exist nowadays, like National Cancer Institute, OpenFDA, HMD, Data.gov, Kaggle…etc., but due to inconsistency, noise, and discrepancies, besides containing an attribute of specialists diagnosing information that represents the class or category of disease which is the most significant attribute in the classification process, that's why these data must be prepared and preprocessed carefully (e.g. handling missing data, applying transformation like normalization and discretization processes, noise smoothing,..etc.) are essential before starting mining process at hand, and getting accurate mining results. After the brief introduction about data mining, predictive modeling, and the importance of these techniques in building predictive models in the domain of healthcare. The remainder of this paper will focus on the different techniques proposed by researchers in this field that will lead finally to fruitful insight and research gaps to enable other research in the healthcare predictive domain.The key point that the paper will highlight and discuss is to review the existing predictive models, techniques used, their advantages and weaknesses, and limitations. Various kinds of diseases will be highlighted starting from diabetes, kidneys, cardiovascular, COVID-19, …etc.. The structure will be as follows: in the second section the methodology, while in the third section will include a brief literature review on the selected predictive modelling have been chosen for reviewing, the last two sections will include results summary and discussions and conclusions for section four and five respectively.

1. METHODOLOGY

In this section, the author will present the methodology used for the review of the present state of the art on predictive modeling in healthcare using data mining. Different digital libraries were used for collecting relevant articles. These libraries are IEEE, ACM, Elsevier, Springer, Scopus database and Google Scholar are used in the article search process. As many as 45 articles were initially collected from these sources. When different filters are applied such as inclusion exclusion criteria, finally 15 articles were found highly relevant and suitable for the review. The search process includes the usage of different search phrases such as predicting modeling using data mining and machine learning for predictive modeling in healthcare. Inclusion criteria includes a collection of only articles in English language, only peer-reviewed articles, their closeness to relevance and quality of the content. Table 1 shows the mapping of finally chosen articles with publishers.

Publisher		**References**
IEEE	4	1, 4, 6, 13
Elsevier	8	2, 5, 9, 10, 11, 12, 14, 15
Springer	3	3, 7, 8

Table 1: Shows final list of article references and their publisher

DOI:http://doi.org/10.24086/cocos2024/paper.1083

As presented in Table 1, the count of articles collected from each source and the corresponding article reference numbers are provided.

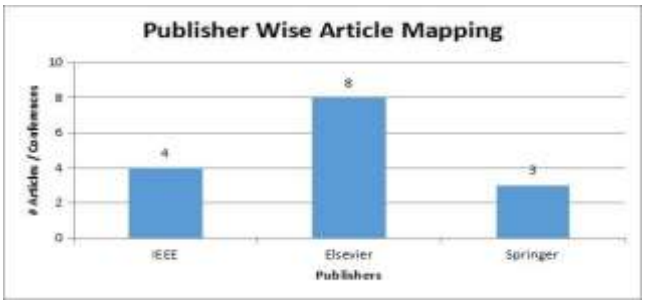

Figure 1: Number of articles collected from different publishers

As presented in Figure 1, as many as 8 articles are collected from Elsevier, three articles from Springer and four articles from IEEE. Out of 45 articles originally collected through search process, finally, 15 articles are used for review as given in Section 3.

3. LITERATURE REVIEW

In this section number of literatures will be reviewed on different prior works about predictive modeling using different mining techniques focusing more on the healthcare domain.

3.1. PREDICTIVE ANALYTICS USING DATA MINING

Predictive analytics is the process of analyzing data to perform the predictions needed. predictive analytics is used to predict the trends and analyze students success future depending on huge amount of historical data. The importance of such analytics is to address the problem of students dropping out their degrees and studying which in turn indicates the failure of institution to achieve their goal, and finally will lead to a huge lowering in human capital as skilled and professional graduates ratio degrades. Shafiq D. A. et al.(2022) started the paper with the critical importance of student retention in education and the methods employed to address this challenge, particularly through predictive models using Educational Data Mining (EDM) "which consists of discovering knowledge from educational data sources, to enhance the teaching practice, and classify and determine students that are at-risk (fails to graduate)" and Learning Analytics(LA), The focus is on factors influencing student performance and the application of various techniques, such as Machine Learning (ML) and Deep Learning(DL). (Shafiq D. A. et al., 2022) The authors in this survey highlighted the need for learning-based techniques to address student retention problems that are often addressed using ML and DL techniques. Gaps and limitations in these surveyed researches have been identified and addressed in coverage, quality, and focus on significant factors, ignoring less significant ones. It provides insights into both factors and features that are significant and less significant affecting the prediction of student's performance. recommendations for future research were concluded, including exploring clustering and ensemble methods, to understand the behavior of students are not supported. separating students according to their observable and similar patterns. Since students are not homogeneous(same characteristics) group, clustering techniques are unsupervised and will help to identify new characteristics and patterns to make new clusters and classify students that are for example academically weak, improving the overall predictive model's efficiency and accurateness. (Shafiq D. A. et al., 2022) In Shafiq D. A. et al.(2022) study limitations can be summarized to include the scope of the review and the need for further theoretical contributions, on the other hand, future directions involve more deep analysis and comparisons of different learning environments' predictive models.

DOI:http://doi.org/10.24086/cocos2024/paper.1083

1.1. Drugs Shortage Prediction Using Data Mining

In the domain of health care, medicine availability has an essential role in providing services to the public. Any shortage in drugs could have a serious consequence most importantly on patient's health, mortality, and the economy of the nation since the healthcare centers will lose their trust, some countries represent a therapeutic interface and destination for some incurable diseases. In Europe, a shortage of 39% between 2018 and 2019 has been addressed in a survey, with a ratio of 95% of hospitals globally suffering from this issue (EAHP. 2019). In this prediction of drug shortage area, probability plays a crucial role. (Nguyen A.et al. 2023) discussed the global issue of drug shortages in healthcare systems and explore the potential of data analytics. Collected data have been cleansed and checked, significant features have been extracted and engineered from data and finally, they analyzed the data using statistical methods including mean- shipping-time(average time elapsed between the shipment date and the ordering date.), correlation analysis to analyze the correlation between number of orders and mean-shipping-time, interquartile range (IQR) method to identify anomalies. (Nguyen A.et al. 2023) Research findings propose a promising opportunity to support advanced analytics techniques for predicting and managing shortages in drugs at local, national, and European levels. They have found that the shipping time provided by electronic. (Nguyen A.et al. 2023) data of drug orders and shipments contained early information about anticipated shortages. Limitations of the research can be summarized as focusing the analysis process on a single measure depended on only shipping and ordering dates in specific regions and time periods using the most simple analytical method. The researchers suggest including a geographical dimension and examining the spread of supply disruption at both local and European levels using more advanced mining techniques. (Nguyen A.et al. 2023)

1.1. Modelling Heart Disease Prediction Problem using ML and Data Mining

Good health is a blessing for all humankind, especially heart health. The leading cause of death and disability in the world. Two of the most significant factors that could be the main reason behind heart disease are hyperattention and anxiety and diabetes. Many techniques and methods were proposed for the early detection, classification, and analysis of different heart issues and disorders. Different data mining techniques have been combined and used to analyze data related to this topic. Merging these techniques with machine and deep learning approaches will lead to promising results.(Sharma R., 2022) The data being collected, include 9 features and 209 instances from the retrospective method from SRHC Govt Hospital, these data were preprocessed using discretization and binning techniques then different classification techniques and algorithms were tested to compare their performance in the classification process. these techniques are generalized linear model, logistic regression, Naïve Bayes, deep learning, decision trees, multilayer perceptron, and gradient-boosted trees. (Sharma R., 2022) They found that the deep learning method has a performance accuracy reached 71.4% using precision and recall measures to identify and calculate the ratio of misclassified cases, with a ratio of 28.6% and 72.2% for recall and precision respectively. (Sharma R., 2022) Researchers have concluded that due to the sensitivity of disease prediction and estimation one technique is not sufficient for deciding the best method for classification and prediction and because the health prediction models are related closely to human life which makes disease prediction a very sensitive topic. Selecting the most appropriate method or technique for this purpose could be dependent on the dataset used. For

DOI:http://doi.org/10.24086/cocos2024/paper.1083

instance, predicting breast cancer is different from predicting diabetes since the first one depends on genetic and lifestyle factors and regular datasets could be used while in the latter one x-rays and MRIs are used for prediction. So, the selection and the performance of a given algorithm depend mostly on the mining process at hand and dataset type. (Sharma R., 2022) Different accuracy measures and techniques have been proposed to evaluate the performance of the different techniques at hand. The findings showed that Deep learning has the best performance than the naïve classifier with an accuracy of 71.4% to 69.0% followed by gradient-boosted trees with a 66.7% accuracy and MLP(Multi-Layer Perceptron) 65.07%, while linear regression and decision trees has the same performance with accuracy ratio 64.3% and the voted perceptron has gained the least accuracy ratio reaches 56.93%.(Sharma R., 2022) We can conclude that integrating data mining techniques with the Deep Learning approach could be advocated to get higher performance and more accurate results.(Sharma R., 2022) Tarawneh M. and Embarak O. (2019) have stated that diseases of heart could be the leading cause of disability and death, the lack of experts and ignoring early symptoms could lead to fatal results and introduces challenges that are addressed by Tarawneh M. and Embarak by proposing an expert system to mine hidden patterns in heart disease datasets. Results showed that the combined models are efficient and diagnosing were correct and accurate. Heart diseases diagnosing are not simple and requires experience, diagnosing improperly could have not only bad but severe fatal consequences. using datamining techniques combined with building predictive models helps in correct prediction. Variety of prediction and classification techniques were used in the study, a hybrid approach was proposed by the researchers for better performance and more accurate heart diseases diagnoses. Different mining techniques analyzed Cleveland dataset, focusing on creating an expert system that quickly and accurately diagnoses heart diseases in both hospitals and educational environments. The recommendations for future studies included designing an accurate, efficient, and remote heart issues and disorders predictor system that support real time monitoring. In another study, Bharti, R. et. al., (2021) have showed the essential role of correct and early diagnosing in healthcare, especially heart diseases, because it's a life threaten, it's considered a challenging issue as mentioned earlier. Machine learning (ML) and datamining techniques is a good combination that showed fascinating reliable results in the instant diagnosing of diseases in the field of healthcare, contributing in saving human lives. To achieve high performance and assist healthcare experts to detect and identify heart issues, image fusion and several different ML methods been utilized. ML, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), were explored and their effectiveness for detecting the most threaten disease that cause 17.5M deaths per year around the world according to WHO have been highlighted throughout the reviewed studies. Most often data used for these studies need to be preprocessed, because the cleanness and data quality affects greatly on the prediction accuracy and efficiency. The conclusion stresses the need for reliable datasets, pre-processing, and appropriate algorithms, noting the promising results of artificial neural networks in heart disease prediction. The use of ML and image fusion is deemed essential for both healthcare authorities and patients, with a call for more high-quality datasets to improve predictive models and aid early detection of heart disease.

DOI:http://doi.org/10.24086/cocos2024/paper.1083

3.4. DIAGNOSE EYEDISEASES USINGVARIOUS FEATURES EXTRACTION APPROACHES AND ML ALGORITHMS

Similar to heart diseases, eye diseases are also one of the crucial issues and health problems that must get an extensive focus due to the importance of this organ, and disorders in the health of the eye could cause blindness. Abed and Al Bakry (2023) have utilized the Ocular Disease Intelligent Recognition (ODIR) dataset from Kaggle including 5000 instances each with 19 attributes (age, left and right eye fundus images, Doctors' diagnostic keywords, … etc.)to predict which patients might have eye illnesses. The prediction model was built to predict the eye illness based on the fundus image. These images first were preprocessed by converting the input image from color images to grayscale to improve the computation time and speed due to this conversion from colored (which requires 3 channels and more data to represent the image) into grayscale (only one channel is used to represent the image). Histogram Equalization (HE) was used to lower contrast, noise, and blurriness. (5x5) Gaussian filter is used to blur the image and fine the image by eliminating the noise from small image details. Bicubic interpolation was used to resize the images to 200x200. On the other hand feature extraction techniques were used to reduce the dimensionality. The input data will be reduced to feature maps, for this purpose SIFT (Scale-invariant feature transform) Features Extraction and GLCM)Gray Level Co-occurrence Matrix(Features Extraction techniques were used. Abed and Albakry(2023) have used different algorithms for classification including Naïve Bayes (NB), K-nearest Neighbor (KNN), Random Forest (RF), and Decision Trees (DT) Algorithms. To evaluate the performance of each of these algorithms, different measurements have been used by the researchers like, precision, recall and F-measure. Findings of this research can be summarized as follow: the RF has performed better than the other algorithms with precision 88%, followed by KNN, NB and DT with precisions 61%, 42% and 39% respectively. The model proposed successfully classified 5000 different fundus images with 8 different fundus types, in which various ocular diseases were represented with higher precision than other reviewed and proposed models. (Abed and Al Bakry, 2023)

3.2. DATA MINING AND PREDICTIVE MODELLING FOR COVID-19 DIAGNOSIS AND RECOVERY PREDICTIONS

"Health is the real wealth" is the lesson that Covid-19 has taught humanity during the pandemic. With Covid-19 disease predictive modeling and predicting its severity has become possible. Rane et. al. (2023) have focused their study on forecasting Covid-19 patients using different techniques, analyzing their merits and demerits, and categorizing models, focusing on the importance of accurate short-term predictions to effectively assign resources and holding back Covid-19's spread throughout the world. Various models were explored by the researchers such as Multilayer Perceptron (MLP), Model- based Least Square Support Vector Machine (LSSVM), Susceptible, Infection and Recover (SIR), and Susceptible, Exposed, Infection and Recover (SEIR). The key challenge in this pandemic was some contagious patients were symptomless that's why Symptomatic and asymptomatic factors were considered. Linear and non-linear models have been reviewed, and their characteristics were explained in detail, focusing attention on parameter evaluation needs calling for reducing the time complexity, and an overview of prediction techniques was provided beside future studies insight in the same field. (Ranem et al., 2023) Muhammad, L. J. et. al. (2020) have studied the recent pandemic which lacks agreed on proven treatment or drugs or proven vaccines. They have implemented different data mining algorithms including

DOI:http://doi.org/10.24086/cocos2024/paper.1083

(DT, SVM, NB, LR, RF, and K-NN algorithms) on the South Korean epidemiological dataset of COVID-19 patients to build a predictive model that predicts infected patients' recovery. The decision tree algorithm model has exhibited an outperforming accuracy that represents the highest among the other models with 99.85% accuracy. The model also predicted the minimum and maximum number of days required for recovery from the virus, plus the age group who are at higher risk not to recover from the virus. Such mining techniques are essential since there are non clinically proven treatments, to extract valuable and meaningful patterns and associations from the datasets to produce effective and accurate predictions about SARS-CoV, MERS-CoV, and Covid-19. The researchers also focused on the importance of mining and analyzing the vast collected COVID-19 datasets to improve the process of decision-making in healthcare, using mining applications to predict patient outcomes and evaluate effective treatments. (Muhammad, L. J. et. al., 2020)

1.1. Predictive Modelling and Data Analytics for Leveraging Clinical Diagnosis

With the use and the help of AI and Predictive analytics, decision making, and diagnosing, become very interesting for their ability of saving lives and diseases early detection. Divyashree N And Divyashree and Nandini (2022) have used "Local Weight Global Mean K-Nearest Neighbor (LWGMK-NN)" on ten datasets that could be easily found at "Kaggle", "UC Irvine Machine Learning Repository", and "data.World". with statistical resampling, 5-fold cross validation for 5 iterations implemented on Google's Colab with TPUs (Tensor Processing Unit) and GPUs (General Processing Unit). The algorithm is compared against nine different benchmark algorithms, including Logistic Regression and Random Forest Decision trees, Support Vector Machine, Gaussian Naive Bayes, Artificial Neural Networks, Stochastic Gradient Descent, and XGBoost. The prediction driven by these advanced data analytics and machine learning proves to be beneficial in various fields, particularly healthcare. The proposed algorithm's key features include data optimization, computation of Local Weights and Global Mean, which leads to a high and efficient predictions. Results and findings highlight the high accuracies across all datasets. The research suggests to improve and address many aspects, including further analysis of complexities and exploration in different domains. For future works they suggest involving the optimization of the LWGMK-NN algorithm, exploring additional metrics, and combining predictive analytics into a decision support system for prescriptive analytics. (Divyashree N And Nandini Prasad K S. , 2022)

3.3. Data Mining for Kidney Disease Diagnosis and Severity Prediction

Kidney diseases could also be life threaten disease that might lead to kidney failure El-Houssainy A. R., and Anwar, A. S. (2019) have showed that controlling on chronic kidney diseases(CKD) effectively depends highly on the early detection and right treatment. The researchers have proposed a system that help specialists accurately identify and detect the severity of the illnessaccurately with the use of data mining algorithms to extract useful insights and hidden patterns from laboratory and clinical data of 361 chronic kidney disease patients. Many techniques have been explored and studied including Radial Basis Function (RBF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Probabilistic Neural Networks (PNN. The PNN has achieved

DOI:http://doi.org/10.24086/cocos2024/paper.1083

better performance in prediction and classification of CKD severity with 96.7% accuracy which is better than the other techniques. It also showed that using machine learning techniques are very helpful in eliminating treatment and diagnosing errors.

3.4. Demand Supply Problem Solving Using Predictive Analytics

The predictive analysis is not only beneficial in the healthcare field many real world problems could be solved using mining techniques and predictive analysis Sathishkumar, V. E. et al. (2020) have discussed the problem of bikes rental in urban cities, how the predicting hourly demands for renting bikes has helped to enhance the mobility and assure a stable supply. This paper employs data mining techniques to address this challenge, utilizing weather information, bike rental counts per hour, and date information. Five statistical regression models were trained and evaluated. The Gradient Boosting Machine model, utilizing all predictors, demonstrates the highest R2 values of 0.96 in the training set and 0.92 in the test set. The paper also explores the significance of various predictors, revealing Temp and Hour as crucial variables. Weather data consistently improves prediction accuracy, highlighting its impact on rental bike demand. Future work is suggested for district-level predictions to enhance the availability of public rental bikes systematically.

3.5. Modelling Using Data Mining for Diabetes Classification and Risk Prediction

Diabetes causes many risks that can be predicted using data mining and predictive modelling. Yang, H.; et al. (2021) addressed the global diabetes epidemic, emphasizing the importance of early diagnosis to prevent chronic damage caused by hyperglycaemia. The research introduces a computational novel system for predicting and evaluating diabetes risk by integrating diverse physical examination data. A dataset includes a large amount of physical examination data was collected from an electronic medical record (EMR), from 2011 to 2017 from Luzhou Municipal Health Commission in China comprising 1,507,563 records from healthy individuals and diabetes patients, along with 387,076 follow-up records. A diabetes risk scorecard based on XGBoost and logistic regression was established for clinical convenience. The study also identified key factors influencing patient control through follow-up records. An online diabetes risk assessment system was developed for broader accessibility. The proposed system contributes to large-scale diabetes screening and personal lifestyle management, offering valuable guidance for human health management. Challenges in diabetes classification have been addressed by Azad, Ch. et al. (2021), in their study they have proposed a model comprising four layers: the first layer is data pre-processing, in which missing data, outlier detection and oversamples using synthic minority oversample technique (SMOTE) been treated, in the second layer selection and extracting of relative features were addressed using genetic algorithms and correlation analysis, in the third layer the model was trained using training data, while in the fourth layer different metrics and measurements were used to evaluate the results generated and the performance of model. An accuracy of 82.1256% was achieved using Synthetic Minority Oversampling Technique, Genetic Algorithm and Decision Tree (PMSGD) algorithm. The study suggested the use of machine learning algorithms combined with quality healthcare data for better and more accurate results in diagnosing and the prediction of diseases, suggesting future direction to include automatic diabetes exploring and analysis and the use of nature inspired algorithms. (Azad, Ch. et al., 2021)

DOI:http://doi.org/10.24086/cocos2024/paper.1083

1.1. Predicting Survival Probability of Heart Failure Patients

Heart failure patients can have chances of recovery. Predicting their survival probability is made possible with data analytics. Ishaq, A; et al. (2021) addressed the significant global issue of cardiovascular disease (CVD), a leading cause of mortality. Predicting the survival of heart disease patients poses a challenge in clinical data analytics. The research employs data mining techniques to analyze a dataset of 299 Heart-failure-clinical-records-dataset were taken from the UCI machine learning repository patients admitted to the hospital, aiming to identify significant features and effective models for predicting cardiovascular patient survival. Nine classification models, including Decision Tree, AdaBoost, Logistic Regression, and others, are utilized. The study addresses the imbalance class problem using Synthetic Minority Oversampling Technique (SMOTE). The Extra Tree Classifier (ETC) emerges as the most effective model, achieving an accuracy of 0.9262 with SMOTE. The study underscores the importance of early diagnosis in preventing heart failure, which is a major contributor to mortality. The proposed machine learning-based approach has the potential to enhance healthcare by predicting heart failure survival and aiding healthcare providers in focusing on major risk factors for improved patient outcomes. Future research could explore combinations of machine learning models and advanced feature selection techniques to further enhance performance.

1.1. Big Data Analytics for Pandemic Prediction

Health pandemics when predicted early can lead to better management of healthcare services. Ahmed, I. (2021) discussed the increasing prevalence of IoT devices and sensors across various sectors, producing substantial data with potential value, particularly in healthcare. Big data analytics emerges as a critical tool for analyzing this data, with a focus on its applications during the COVID-19 pandemic. The integration of IoT and big data technologies plays a pivotal role in addressing the challenges posed by the pandemic, facilitating tasks such as data collection, visualization, risk breakdown, case tracking, and prevention level monitoring. The paper presents a health monitoring framework utilizing big data analytics and IoT for the analysis and prediction of COVID-19. Descriptive, diagnostic, predictive, and prescriptive analyses are conducted, emphasizing the contribution of neural network-based models in diagnosing and predicting the pandemic. The results showcase the neural network's superior accuracy of 99%, highlighting the potential of these technologies in healthcare, including remote monitoring, early disease prediction, and efficient data management. The literature review further explores various studies, frameworks, and surveys related to big data analytics and IoT applications in healthcare, underscoring their significance in enhancing services, improving patient outcomes, and addressing global health challenges.

4. DISCUSSION AND COMPARISON

This section provides the summary of findings based on the literature review provided in Section

References	Modeling Technique (s)	Disease/Problem	Advantages	Disadvantages
Shafiq D. A. et al.(2022)	ML and predictive models	Student retention	Enhancing student learning and retention.	It has limited coverage and scope.
Nguyen A. et al. (2023)	Data analytics and data mining	Drug shortage prediction	Drug shortage detection with acceptable accuracy	Needs to be improve with multiple indicators.
Sharma R.,et al. (2022)	Naïve Bayes, Regression and Classification models	Heart Disease	Achieved high level of accuracy	Real and large dataset is not used.
Tarawneh M. and Embarak O. (2019)	Expert System	Heart Disease Symptoms	Detection of heart diseases automatically depending on early symptoms	Remote patient monitoring is yet to be explored.
Bharti, R. et. al., (2021)	ML, SVM, DT	Heart Disease	Early detection and classification	more high-quality datasets required
Abed and Al Bakry, 2023	ANN, NB, KNN, FR , DT	Eye Diseases	high accuracy	To be evaluated with more patient datasets

DOI:http://doi.org/10.24086/cocos2024/paper.1083

Ranem et al., (2023)	Forecasting using ML models	COVID-19	Covid-19 detection automatically	Accuracy is still an issue.
Muhammad, L. J. et. al. (2020)	Decision tree and data mining	COVID-19	Efficient prediction of Covid-19	Accuracy is still to be improved.
Divyashree N And Nandini Prasad K S. , 2022	LWGMK-NN	Clinical Diagnosis of Diseases	data optimization, computation effecient predictions	Does not include prescriptive analytics, analysis of complexities and exploration is required
El-Houssainy A. R., and Anwar, A. S. (2019)	ML, PNN, MLP, SVM, RBF	Kidney Disease Stages	PNN & SVM High level of performance in classification and prediction	Room for further improvement
Sathishkumar, V. E. et al. (2020)	Regression and classification models	Rental bikes sharing	Stable model for predictions in bike sharing	Needs more fine-grained intelligence generation
Yang, H.; et al. (2021)	Big data mining and Machine Learning	Diabetes	Risk assessment is found to be very useful for patients	Needs to be integrated with healthcare applications
Azad, Ch. et al. (2021)	PMSGD	Diabetes mellitus	Prediction of diabetes mellitus	To be improved for leveraging accuracy
Ishaq, A; et al. (2021)	Data mining and machine learning, supervised ML	Survival of Heart Failure Patients	Feature selection could improve training quality	Needs further improvement in feature selection.
Ahmed, I. (2021)	NN, IoT and Big data	COVID-19	Prediction of Covi-19	Needs further improvement in accuracy.

Table 2: Summary of findings

As presented in Table 2, the predictive modelling is found in different research endeavors in healthcare domain using data mining. The research includes disease prediction, risk assessment, classification of ailments and drug shortage estimation. It also throws light on detection of pandemics and various diseases.

5. CONCLUSIONS

In this paper, the author made a review on predictive modelling using data mining in healthcare domain. The review article threw light on various insights in healthcare domain in relation with predictive modelling using data mining and machine learning. The predictive modelling techniques found in the literature are used for different purposes such as drugs shortage prediction, heart disease diagnosis, prediction of Covid-19 and its severity, leveraging clinical diagnosis, discovering probability of recovery of patients from Covid-19, kidney disease stages prediction, prediction of risk in diabetic patients and so on. This research also covers predictive modelling the survival rate of heart failure patients, classification of diabetes mellitus and prediction of pandemic diseases. This research provides useful knowledge pertaining to the importance of predictive modelling and usage of techniques pertaining to data mining and machine learning in healthcare domain. Different ML techniques like Support Vector Machine (SVM), Random Forest (RF), Neural Networks (NN) and Multilayer Perceptron (MLP), to mention few, are widely used in predictive modelling.

1. REFERENCES

Hoover, Stewart M. (2006). Religion in the Media Age. Media, Religion and Culture (1st ed.). New York: Routledge. ISBN 978-0-415-31423-7.

Reinsel, D.; Gantz, J; Rydning, J. (2017). "Data Age 2025: The Evolution of Data to Life-Critical" (PDF). seagate.com. Framingham, MA, US: International Data Corporation. [online] Available at: https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf (Accessed 28 Dec. 2023)

Han J. Pei J. and Tong H (2023). Data Mining Concepts and Techniques Fourth Edition. USA: Elsevier.

SHAFIQ, D., A., et al., 2022, Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. IEEE. 10, pp.72480 - 72503.

DOI:http://doi.org/10.24086/cocos2024/paper.1083

[online] Available at: http://DOI:10.1109/ACCESS.2022.3188767 (Accessed 28 Dec. 2023)

EAHP. 2019 EAHP Medicines shortages report. 2019. [online] Available at: https://www.eahp.eu/sites/default/files/eahp_2019_medicines_shortages_report.pdf. (Accessed 26th Dec. 2023)

Nguyen, A. et al. (2023) 'On The Use Of Logistics Data To Anticipate Drugs Shortages Through Data Mining', Procedia Computer Science, 219, pp. 949–956. [online] Available at: http://doi:10.1016/j.procs.2023.01.371(Accessed 12nd 2023)

Sharma, R.; Gupta, Sh. and Garg P. (2022). Model for Predicting Cardiac Health using Deep Learning Classifier. IEEE., pp.25-30.[online] Available at: http://DOI:10.1109/CCiCT56684.2022.00017 (Accessed 11st 2023)Deb, A., M.; Koli, S. A.;
Akter, Sh.; and Chowdhury A. A., (2022). An Outcome Based Analysis on Heart Disease Prediction using Machine Learning Algorithms and Data Mining Approaches.IEEE, pp.1-7. [online]Available at: http://DOI:10.1109/AIIoT54504.2022.9817194(Accessed at 21st Nov. 2023)

Ranem, R.; Dubey, A.; Rasool, A.; and Wadhvani, R. (2023). Data Mining Based Techniques for Covid-19 Predictions. Elsevier. 218, pp.210-219.[online] Available at: https://doi.org/10.1016/j.procs.2023.01.003(Accessed 13th 2023)

Abed, Z. N. and Al-Bakry, A. M.(2023). DIAGNOSE EYES DISEASES USING VARIOUS FEATURES EXTRACTION APPROACHES AND MACHINE LEARNING ALGORITHMS.Iraqi Journal for Computers and Informatics Vol. [49], Issue [2][online] Available at:https://doi.org/10.25195/ijci.v49i2.437(Accessed 22nd 2023)

Divyashree, N And Nandini, Prasad K S. (2022). Improved Clinical Diagnosis Using Predictive Analytics. IEEE.10, pp.75158 - 75175. [online] Available at:http://DOI:10.1109/ACCESS.2022.3190416.(Accessed 11th Dec. 2023)

Muhammad, L. J.; Islam, Md. Milon; Usman, Sani Sharif and Ayon, Safial Islam (2020). Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patientsâ€™ Recovery. SN Computer Science, 1(4), 206–.[online]Available at: http://doi:10.1007/s42979-020- 00216-w (Accessed 12th Dec. 2023)

Bharti, R. (2021). Prediction of heart disease using a combination of machine learning and deep learning. DOI:http://doi.org/10.24086/cocos2024/paper.1083

Computational intelligence and neuroscience.[online] Available at: https://doi.org/10.1155/2021/8387680(Accessed 18th Dec. 2023)

Tarawneh, M. and Embarak, O. (2019). Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques. Acta Scientific Nutritional Health 3.7: pp(147-151) © Springer Nature Switzerland AG [online] Available at:DOI:10.1007/978-3-030-12839-5_41 (Accessed 18th Dec. 2023)

El-Houssainy A. R., and Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked, 15, 100178–[online] Available at: http://doi:10.1016/j.imu.2019.100178 (Accessed 20th Dec. 2023)

Sathishkumar, V. E. et al. (2020) Using data mining techniques for bike sharing demand prediction in Metropolitan City, Computer Communications 153, 353–366. [online]Available at: https://dl.acm.org/doi/abs/10.1016/j.comcom.2020.02.007 (Accessed: 29th Dec. 2023).

Oulad, S. Y.; Hajar, M. and Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. Fire Safety Journal, S0379711218303941–[online]Available at: http://doi:10.1016/j.firesaf.2019.01.006 (Accessed 20th Vov. 2023)

Yang, H.; et al. (2021). Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators.Information Fusion.[online]Available at: http://doi:10.1016/j.inffus.2021.02.015 (25th Nov. 2023)

Azad, Ch. et al. (2021). Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus . Multimedia Systems. [online]Available at:http://doi:10.1007/s00530-021-00817-2 (Accessed 28th Dec. 2023)

Ishaq, A; et al. (2021). Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques . IEEE Access. [online] Available at:http://doi:10.1109/access.2021.3064084 (Accessed 30th 2023)

Ahmed, I.; Ahmad, M.; Jeon, G. and Piccialli, F; (2021). A Framework for Pandemic Prediction Using Big Data Analytics. Big Data Research. [online] Available at: http://doi:10.1016/j.bdr.2021.100190 (Accessed 21st 2023)