Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1 : 2024
ISSN : **1906-9685**

MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

#1**THANGALLAPALLI KALYANI,** *Assistant Professor, kalyanisriramula14@gmail.com,*

#2**JONUKUTI SHEKAR,** *Assistant Professor, jonukutishekar@gmail.com ,*

Department of Computer Science and Engineering,

SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TELANGANA.

ABSTRACT: The multidisciplinary discipline of healthcare data mining sprang directly from the integration of database statistics. It is an excellent resource for determining the efficacy of various healthcare therapies. The application of machine learning algorithms and techniques to data visualization can help with the management of diabetes-related cardiovascular disease, a subset of cardiovascular sickness that affects diabetics. Diabetes is a chronic medical condition characterized by either an insufficient amount of insulin produced by the pancreas or an inefficient utilization of the insulin produced within the body. Circulatory disease, often known as heart disease, is an umbrella term for a number of illnesses that can have a negative impact on a person's circulatory system. There are several data mining classification algorithms that can be used to forecast cardiovascular illness; however, there has been little investigation into how this might be done for diabetics. We improved the decision tree model's capacity to predict whether or not a diabetic will acquire heart disease. This was done because the decision tree model regularly outperformed the naive Bayes and support vector machine models.

Keywords: Prediction, Machine Learning, Classification, SVM.

1. INTRODUCTION

Data collection and analysis in healthcare are both difficult and time-consuming operations. The digital revolution and subsequent technological improvements have enabled the generation of massive amounts of patient data that integrate numerous dimensions. This broader topic includes several subfields, including as diagnostic data, patient files, medical histories, and healthcare facilities. It is critical to efficiently handle and analyze enormous datasets that are also compact and complicated in nature in order to gain significant insights that help direct and improve decision-making processes. The discovery of hidden patterns is a topic that offers a lot of potential for future research in the field of medical data mining.

Machine learning and data mining approaches have had a transformative impact on healthcare businesses. These technologies have enabled the detection of significant patterns as well as linkages and correlations among numerous variables contained inside massive databases. As a result, these strategies have enabled healthcare organizations to shift. The significance of data analysis in the field of healthcare is based on its ability to supply and evaluate previously obtained data in order to support the design of prospective course-of-action plans. The technology makes it much easier to explore large datasets because it integrates a number of advanced analytic methodologies and cutting-edge algorithms. The

healthcare industry follows a set of defined practices when it comes to obtaining, organizing, and reviewing patient information. Using this method, we may focus in on the most efficient ways to solve problems that develop during service delivery and uncover their root causes. This could lead to more accurate diagnostics, more efficient surgical procedures, and more effective drugs in the long term. Furthermore, it provides a solid foundation for understanding the inner workings of a number of components employed in the medical industry. Using information acquired from multiple medical databases allows for the early detection and treatment of ailments.

The examination of a patient's symptoms in order to determine the nature of the health issue that the patient is experiencing is what is meant by the term "diagnosis." There is no agreement among medical professionals on whether or not a diagnostic examination is an essential component of any particular diagnostic procedure or sequence of procedures. Accurate diagnosis of chronic diseases is critical in medicine, and this requires a thorough examination of a wide range of symptoms. It may be difficult to put this strategy into practice, and doing so may result in incorrect judgments. When making a clinical decision, it is critical to consider both the patient's self-reported symptoms as well as the physician's prior knowledge and expertise in the field of sickness diagnosis. It is becoming increasingly difficult for medical professionals, such as doctors and other healthcare workers, to keep up with the rapid rate of change in clinical practice caused by the creation of novel medical systems and treatments. To provide the best possible care, medical professionals and other healthcare workers must be well-versed in the various diagnostic criteria, patient medical data, and pharmacological therapies. When people opt not to pursue formal education and instead rely on their "gut feelings," which are the result of a lack of information and a limited amount of experience working with patients, errors are possible. A person's mental abilities can be hampered for a variety of reasons, including, but not limited to, inability to multitask, poor analytical skills, and poor short-term memory, to mention a few. The lack of clinical testing data and a thorough medical history for patients presents a substantial difficulty for medical practitioners in their efforts to consistently offer reliable diagnoses. Even the most experienced doctors can benefit from computer-aided diagnostic tools, which can help them make more accurate and objective diagnoses. As a result, there is a lot of interest in the topic of how to delegate the process of diagnosis by integrating machine learning approaches with the expertise of medical professionals. Data mining and machine learning technologies are being actively used by researchers to rapidly translate easily accessible data into information valuable to the diagnosis process. This is being done to enhance the diagnostic procedure. On the subject of how successfully machine learning systems can perform diagnostic tasks, extensive empirical research has been undertaken. According to research, the diagnosis accuracy rate of machine learning algorithms is 91.1%, which is higher than even the most experienced doctors' diagnostic accuracy rate (79.97%). When explicit machine learning techniques are applied to datasets relevant to the disorders under consideration, the best feasible diagnoses, prognoses, preventative plans, and therapies can be obtained.

2. RELATED WORK

Bayesian classifiers make predictions by combining a structural model with a set of conditional probabilities. At least in theory, each element is thought to contribute the same amount. The technique then sets the occurrence of each variable value in an unexpected scenario after evaluating the baseline probability for each category. The Bayesian network classifier is based on a graphical representation of the joint probability distribution of a set of qualities that all belong to the same category.

In order to forecast renal disease, both the SVM algorithm and the Naive Bayes technique were used. The researchers chose to utilize a piece of software called an Adaptive Neuro-Fuzzy

Inference System (ANFIS) to classify the various signs of kidney disease. The main goal of this work was to design a reliable classification method by carefully considering a variety of factors. The Naive Bayes approach was substantially faster than the SVM method, but the SVM method produced significantly more accurate classifications. The results show that the Support Vector Machine (SVM) method outperforms the Naive Bayes technique in predicting renal disease. The method of Naive Bayes was applied.

The primary goal of this study was to predict the development of heart disease, and the researchers chose to do so by applying a fuzzy method based on the concept of a membership function. The authors used the Fuzzy KNN Classifier to solve problems caused by the dataset's intrinsic ambiguity and unpredictability. The dataset contained 550 entries, which were then divided into 25 categories, each of which contained 22 different things. The dataset was partitioned along the center to make training and evaluation easier. After completing the preparatory steps, we used the fuzzy K-nearest neighbors (KNN) approach. A variety of evaluation criteria, including accuracy, precision, and recall, were used to determine the success of this method. The results show that the fuzzy K-nearest neighbors (KNN) classifier beats its more traditional equivalent, the classical KNN classifier.

Methods based on the Artificial Neural Network (ANN) algorithm are urgently required for cardiovascular disease prediction. The researchers created an interactive prediction approach based on classification using an artificial neural network algorithm and thirteen clinically relevant parameters. The proposed method has proven to be a valuable tool for healthcare practitioners in the diagnosis and prognosis of cardiovascular disease, with an amazing accuracy rate of 80% in these undertakings.

The researchers used an algorithmic methodology to find solutions to difficult questions about the prognosis of cardiovascular illnesses. The Naive Bayes method was used to improve the efficacy, precision, and overall allure of this intelligent system. This technology advancement has the potential to help medical workers diagnose and treat people who have had a myocardial infarction. The addition of Short Message Service (SMS) features, the development of mobile applications for the Android and iOS platforms, and the addition of a pacemaker capability are all potential system changes.

Support vector machines, often known as SVMs, have been used to aid in the diagnosis of diabetes and breast cancer by leveraging their versatility. The diagnostic approach was improved by including Adaptive Support Vector Machines (SVM), a type of machine learning that can be very effective and diverse in its application. The findings improved when the bias value used by typical Support Vector Machines (SVM) was changed. The classifier in question generates 'if-then' rules. Using the technologies now under investigation, a diagnosis of diabetes and breast cancer was made 100% of the time. Future research should concentrate on establishing more effective ways for managing the bias parameter in typical Support Vector Machines (SVM).

A novel strategy for predicting the onset of type 2 diabetes has been created using a hybrid model that incorporates clustering and classification strategies. The three components that make up the composite model that is used in the prediction process are k-fold cross-validation, the K-means clustering algorithm, and the C4.5 classification method. We were able to get a classification accuracy of 88.38% using a hybrid technique, which is a very promising result. Because of the study's findings, medical personnel will be able to make better informed clinical decisions, which has the potential to greatly improve diabetes treatment.

3. FRAMEWORK FOR MULTIPLE DISEASE PREDICTION

This framework employs a variety of machine learning techniques, including the decision tree method, naive Bayes, and support vector machine.

The Bayesian classification is a well-known probabilistic approach to data categorization that use Bayes' theorem and assumes that each

variable may be considered independently. Despite its utility, Bayesian categorization is still not well known or understood. There is no direct relationship between the presence or absence of one characteristic within a category and the presence or absence of any other feature within that same category. When certain circumstances are met, the mechanism will activate. Using Bayes' theorem, one may determine the likelihood of a third event occurring given the occurrence of the second event. For the sake of Bayes' theorem, we shall refer to event B as the dependent event, and event A as the prior event. By dividing the values of both samples by two, we can calculate Sample B's proportional relevance in relation to Sample A. The conditional probabilities of event B given event A are computed by first dividing the frequencies of both events by their totals, and then dividing the products of that division. When there aren't enough training examples, the Naive Bayes Classifier can't estimate some parameters accurately. The parameters in question include the variables' medians and variances. Because people are thought to be independent, class variances must be discovered. This concept is not confined to only two or three categories; rather, it is limited to those at most.

The use of Support Vector Machines (SVMs) to facilitate kernel learning is a widespread method in the field of machine learning, particularly when dealing with large-scale prediction difficulties. In terms of generalization and scalability, the SVM classifier beat other classifiers on both linear and nonlinear data. This was true for both sorts of information. When paired with other well-established methods based on statistical learning and optimization theory, the performance of the support vector machine (SVM) classifier in the field of pattern recognition is extremely amazing. The primary goal of this study was to give a comprehensive viewpoint that distinguished between high-quality data and low-quality data by identifying the most common causes of errors. The major motivation behind the SVM categorization scheme is the quest of financial gain.

When the data shows linear separability, it is simple to choose a hyperplane that will divide it into two groups because all that is required is to identify the one that provides the best match. Support Vector Machines (SVMs) use kernel functions to solve complicated problems and translate input into higher-dimensional spaces. Linear kernel functions (LKFs), polynomial kernel functions (PKFs), sigmoid kernel functions (SKFs), exponential radial basis kernel functions (ERBKFs), and generalized radial basis kernel functions (GRBKFs) are some examples of kernel functions. The Radial Basic Function (RBF), which is now the most advanced kernel function, is often regarded as the most effective alternative.

The use of decision trees enables the classification of big datasets. Categorization trees, often known as "decision trees," are used to organize information in a hierarchical fashion. A classification tree's "root" node represents a working hypothesis, while the "leaf" nodes are in charge of creating grouping results.

The mistake rates incurred by the classification approach were meticulously and comprehensively reported.

The tree that is generated can be used to generate rules. Simple concepts are used to build decision trees. There are other alternative decision tree algorithms that can be utilized, including ID3, C4.5, and CART. It is correct to say that the C4.5 data mining system employs a well-developed decision tree algorithm. The purpose of this inquiry is to compare and analyze the financial returns of various possibilities. The C4.5 algorithm's ability to effectively analyze datasets with both categorical and continuous features contributes significantly to the algorithm's extraordinary effectiveness. When compared to prior systems of its kind, this one has a lower impact on memory capacity and is more forgiving of occasional failures in calculation. It is frustrating that derivatives that appear to have no purpose are continually generated. The ID3 technique is significantly reliant on the collection of many types of data. The Classification and Regression Tree, or CART, method is widely used to construct a binary decision tree, which is a tool used by computer systems. The Gini index is used

on the decision tree to determine which nodes within the structure have the least reliable information. The ID3 algorithm does not consider any values that are not present when analyzing discrete features.

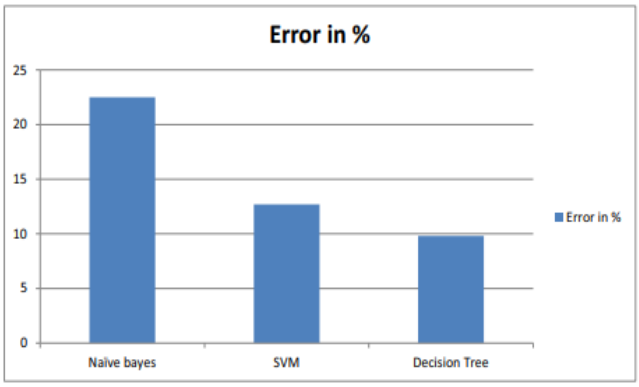

Fig. 2. The following results will go over the classification algorithm's accuracy..

The Cleveland dataset is the primary data source that will be used in this methodology. The Cleveland data compilation went underwent a preprocessing phase to make it more accurate and to remove irrelevant data. When the preparation time is completed, the data supplied to you will be consistent and well-organized. Several machine learning algorithms, like Support Vector Machines (SVM), Naive Bayes, and the Decision Tree C4.5, are currently being used to analyze data that is already available. The algorithms that have been mentioned are capable of appropriately categorizing the data that has been presented to them. Following that, the classification results are used to train a model, which will later be employed in the prediction process. When new patient data is presented, this architecture uses previously obtained training knowledge stored inside the classes to create predictions about the data's normalcy or abnormality. Furthermore, it provides possible disease symptoms. Figures 1 and 2 show the successful and poor results of machine learning, respectively.

4. CONCLUSION

Database statistics play an important part in the development and evolution of the multidisciplinary subject of healthcare data mining. This can be attributed in part to the usage of database statistics. This excellent tool makes comparing different treatment options and determining which ones are the most successful much easier. Including machine learning elements in the process of visualizing obtained data. Diabetes is a long-term medical disorder defined by either insufficient insulin production by the pancreas or insufficient insulin intake by the body. The term "circulatory illness," sometimes known as "heart disease," refers to a group of illnesses affecting the cardiovascular system. The terms "circulatory illness" and "heart disease" are sometimes used interchangeably; however, despite the existence of a number of data mining classification algorithms for predicting cardiovascular illness, there is currently insufficient information to make good estimates for those who have both diabetes and cardiovascular disease. When it comes to predicting the likelihood of diabetics acquiring heart disease, the decision tree model outperformed both the naive Bayes model and the support vector machine model.

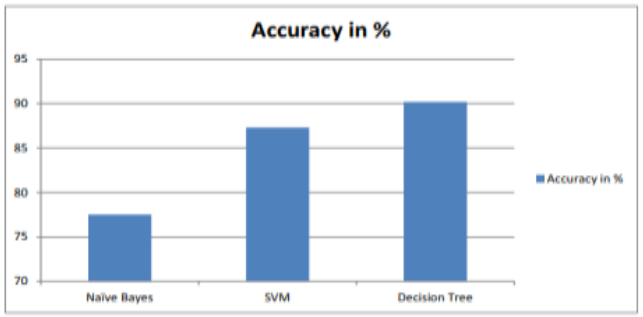

Fig. 1. The classification's results are consistent..

Declaration of Potential Conflicts of Interest

The authors swear an oath that they have no financial or other personal stake in the study's outcomes in any way, shape, or form.

REFERENCES

1. R. Manne, S.C. Kantheti, Application of artificial intelligence in healthcare: chances and challenges, Curr. J. Appl. Sci. Technol. 40 (6) (2021) 78–89, https://doi.org/10.9734/cjast/2021/v40i631320.
2. M. Sivakami, P. Prabhu. Classification of algorithms supported factual knowledge recovery from cardiac data set, Int. J. Curr. Res. Rev. 13(6) 161- 166. ISSN: 2231-2196 (Print) ISSN: 0975-5241 (Online).
3. M. Sivakami, P. Prabhu. A Comparative

Review of Recent Data Mining Techniques for Prediction of Cardiovascular Disease from Electronic Health Records. In: Hemanth D., Shakya S., Baig Z. (eds) Intelligent Data Communication Technologies and Internet of Things. ICICI 2019. Lecture Notes on Data Engineering and Communications

4. Technologies, vol 38. Springer, Cham 477-484. ISSN 2367-4512 ISSN 2367-4520 (electronic), ISBN 978-3-030-34079-7 ISBN 978-3-030-34080-3 (eBook) 2020.
5. P. Prabhu, S. Selvabharathi. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus. In 2019 3rd International Conference on Imaging, Signal Processing and Communication, ICISPC 2019 (pp. 138–142) Institute of Electrical and Electronics Engineers Inc. ISBN:9781728136639. 2019.
6. N. Jothi, N.A. Rashid, W. Husain, Data mining in healthcare – A review, Procedia Comput. Sci. 72 (2015) 306–313.
7. H. Polat, H. Danaei Mehr, A. Cetin. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, J. Med. Syst. 41(4) 2017 55.
8. K.B. Wagholikar, V. Sundararajan, A.W. Deshpande, Modeling paradigms for medical diagnostic decision support: a survey and future directions, J. Med. Syst. 36 (5) (2012) 3029–3049.
9. E. Gürbüz, E. Kılıç, A new adaptive support vector machine for diagnosis of diseases, Expert Syst. 31 (5) (2014) 389–397.
10. M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification, Expert Syst. Appl. 41 (5) (2014) 2239–2249.