Contents available at the publisher website: G A F T I M . C O M

International Journal of Computations, Information and Manufacturing (IJCIM)

Journal homepage: https://journals.gaftim.com/index.php/ijcim/index

Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms

Rehan Ahmed [1], Maria Bibi [2], Sibtain Syed[3]

[1, 3] *Pak-Austria Fachhochschule Institute of Applied Sciences and Technology, Pakistan*

[2] *University of Engineering and Technology Taxila, Pakistan*

A R T I C L E I N F O

Keywords:

Heart Diesease Prediction, KNN, Machine Learning Hybrid Model, SVM.

Received: May, 11, 2023

Accepted: June, 22, 2023

Published: June, 23, 2023

A B S T R A C T

The largest cause of mortality worldwide is heart disease, and early identification is critical in limiting disease development. Early approaches for detecting cardiovascular illnesses assisted in determining the progressions that should have happened in high-risk persons, reducing their risks. The major goal is to save lives by recognising anomalies in cardiac circumstances, which will be performed by identifying and analysing raw data produced from cardiac information. Machine learning can provide an efficient method for making decisions and creating accurate forecasts. Machine learning techniques are being used extensively in the medical business. A unique machine learning technique is provided in the proposed study to predict cardiac disease. The planned study made advantage of open source heart disease dataset from kaggle. Hybrid algorithms for machine learning prediction are the logical mixture of many previous methodologies designed to improve efficiency and produce improved outcomes. The presented work introduces a hybrid method that employs the notion of categorization for prediction analysis. We used real patient data to build a hybrid technique to predicting cardiac disease. KNN and SVM classification techniques were utilized in this paper. Jupyter Notebook is used to implement this hybrid method. A hybrid technique outperforms other algorithms in the prediction analysis of heart disease.

1. INTRODUCTION

The practise of collecting useable information and patterns from a range of raw datasets is usually referred to as data mining. It comprises analyzing massive amounts of data and discovering trends or patterns using one or more techniques. It is useful in a variety of contexts, including analysis, research, and healthcare. Because data mining is a method of investigation and Numerous excellent early prediction systems for healthcare have evolved from medical datasets, which can detect trends in large volumes of data (J. H. Joloudari, et al., 2019). Improving the level of healthcare in the medical sector is best described by the guiding variables that have an influence on it, among which is healthcare data, which may be viewed as the system's foundation for improvement for each patient. The application of data mining techniques to extract knowledge from medical records or datasets will help in the discovery of sickness occurrence, evolution, recognition, and significant facts to establish the sources of diagnostic procedures based on the components existing inside healthcare. An investigation for the information cycle for the classification of illnesses

might potentially involve data mining techniques. Therefore, it will reveal hidden relationships and detect patterns in the data, leading to better and enhanced diagnostic recognition. This paper gives us insight into the subject of heart disease prognosis. The heart, which weighs around 3 pounds, is a vital body organ. The ribcage shields it, as it is located on the left side of the chest. All the body's organs are supplied with blood by the heart through a system of blood arteries. The blood helps the body stay healthy by supplying it with the minerals and oxygen it needs. Heart illness or heart dysfunction can cause serious health issues like heart attacks, strokes, and even death. To guarantee appropriate treatment and care, it is crucial to identify any heart disease symptoms at an early stage. The aim of this study is to design a system that aids in determining whether a patient has cardiac disease by suggesting a hybrid approach utilizing data mining techniques. For this, predictions are made using a predictive analysis model and a variety of algorithms. This model's procedure is divided into four steps. Pre-process the raw data at this step. At the second stage, transform the data that has been processed into a useable form for model. Model training in the third stage. Fourth step uses a learning model to create predictions and then reviews them as necessary.

2. LITERATURE REVIEW

(Jabbar, et al., 2016) conducted research on heart disease by utilising the random forest method and Cleveland dataset. In order to carry out the investigation, the author uses the Chi Square attribute selection of features as well as the GA-based selection of features model. Despite the fact that the evaluation was confined to existing machine learning models, the experimental findings reveal that the suggested model with GA feature selection beat the present models. (Al-Milli, and Nabeel, 2013) explores the use of a back propagation neural network in predicting cardiac disease. The author used a deep learning model known for its accuracy in disease prediction and implemented it using a deep learning. The Cleveland dataset was used in the study, and a simulation was done in Matlab. While the research has yielded promising results, there is room for improvement by employing deep learning models and applying the findings to real-world applications. (Hashi, E.K. and Zaman, M.S.U, 2020) A cognitive method is used in this article to predict heart disease. The study assesses five machine learning techniques for prediction based on their accuracy. The logistic model tree is used to increase prediction accuracy by utilising an ADA boost and bagging model. The experimental results show that the random forest model predicted cardiac disease with great accuracy. (Soni, Jyoti, et al., 2011) The author of this article investigates the prediction of cardiac disease using methods based on data mining. Decision tree algorithm, KNN method,Bayesian classification, neural network classifications, and techniques are all evaluated in the study. Furthermore, the author looks into the usage of genetic algorithms for feature selection in identifying critical traits for heart disease. The decision tree model achieves good accuracy in the experiments. (Alkeshuosh, Azhar Hussein, et al, 2017) this publication describes a method for detecting cardiac disease that employs the Particle Swarm Optimization method. The author created explicit rules based on the Particle swarm optimization algorithm and tested them to find a more precise rule for detecting heart disease. Following the evaluation of the rules, the author employed the C 5.0 algorithm for disease classification based on binary classification. The author verified great accuracy was achieved using Particle sworm optimization and a Decision Tree algorithm. (UCI Official site) offers a study on the prediction of heart disease using data mining techniques. The research looks at techniques like the KNN algorithm, decision tree algorithms, neural network classifications, and Bayesian classification techniques. Furthermore, the author investigates the use of genetic algorithms for feature selection of critical heart disease features. The study tests different strategies and assesses their performance, concluding that the decision tree model achieves excellent accuracy. (P. M. Barnaghi, et al., 2012) In this article a author uses random forest and decision tree first for prediction and to measure their accuracy. After that he uses a mixed hybrid approach consisting decision tree and random forest for the prediction of heart disease and measure the accuracy and compare from the previous. And it gives an excellent performance as compared to other models. (Khalid et al., 2023)provide insights into machine learning-based techniques that can contribute to enhancing

the accuracy of heart disease prediction, while (Aslam et al., 2022) emphasize the importance of data security and privacy compliance in the context of improving healthcare analytics. (Saeed et al., 2019) explore the practices and challenges of nomadic knowledge sharing, providing insights that can inform the development of collaborative approaches in improving heart disease prediction accuracy using hybrid machine learning algorithms.

3. ALGORITHMS USED

3.1. KNN

A supervised learning algorithm called the K-Nearest Neighbor (KNN) classifier divides a given dataset into various clusters based on the user's citations (P. M. Barnaghi, et al., 2012). This algorithm is flexible and can be applied to classification and regression issues. KNN's fundamental premise is that items that are similar often cluster together, and the algorithm finds these clusters by measuring the separation between different data points. Since KNN holds the data and performs operations on it during the classification phase rather than learning directly from the training dataset, it is sometimes referred to as a lazy learner algorithm. A new data point's classification is determined by the majority vote of its closest neighbours. Modifying the value of k can have an impact on the accuracy of algorithm.

3.2. Support Vector Machine

SVM is a specialised supervised machine learning classifier that may be utilised in statistical learning (Alkeshuosh, Azhar Hussein, et al., 2017) for linear and non-linear dataset categorization. It works by utilising a non-linear mapping function to change the original dataset into a more understandable representation. SVM seeks for a linear hyperplane in this newly transformed space that can partition the data points into different classes. The hyperplane is an ideal decision-making boundary, and SVM creates them with support vectors. The hyperplane can split data into multiple classes by utilizing an appropriate function for nonlinear function. Despite being a precise classification approach, SVM is computationally expensive since it involves addressing quadratic issues using mathematical functions that require sophisticated calculations that can take time (Soni, Jyoti, et al., 2011)

3.3 Hybrid Approach

A hybrid is typically defined as a combination of two or more elements with traits that are either similar or dissimilar. Different elements have various characteristics, however once they're joined, the final element may have both characteristics. A hybrid approach combines two or more algorithms, each init hold thier set of benefits. Integration of algorithms produce new results that may be more accurate and precise than utilising the techniques alone. The hybrid model is created by merging the KNN and SVM methods. Svm probabilities are used in the hybrid model. The knn probabilities are combined with the train data and sent into the svm algorithm. Likewise, SVM probabilities are determined and supplied into the test data. Lastly, values are forecast. Machine learning is applied to a preprocessed dataset, and the predicted cardiovascular disease for the provided test dataset.

5. METHODOLOGY

The suggested work is written in Python and implemented in Jupyter Notebook. All of the methodology's implementation phases are used here. The dataset used to train the system is obtained from the UCI machine learning library (UCI Official site).

5.1. Data Description.

This study utilized a heart disease Dataset obtained from the UCI machine learning repository to construct a model. The Dataset included various attributes such as sex, age, resting blood pressure, chest pain, fasting blood sugar, cholesterol, maximum heart rate achieved, resting electrocardiogram, ST depression induced by exercise, exercise-induced angina, number of major vessels, slope of peak exercise ST, pred attribute, and thalassemia.

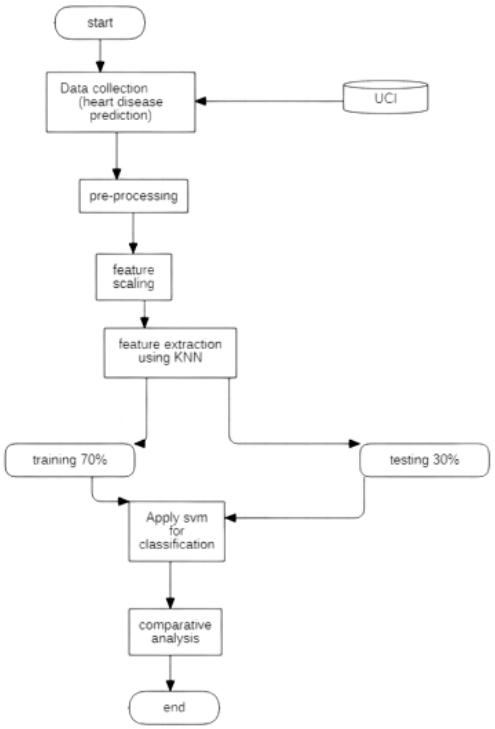

Figure 1

5.2. Working

Proposed workflow utilized two machine learning algorithms and a hybrid model to achieve accurate predictions of heart disease. The advantages of this approach include implementing an optimized model through the use of the hybrid model. The methodology involved collecting the dataset from uci.edu, performing data visualization, and splitting the dataset into test and train data. We are applying the KNN and SVM models for training and analysis. The model is trained using 70% of the dataset as training input, and the remaining 30% is used as testing data for heart disease prediction. The KNN, SVM, and Hybrid of both are used to predict heart disease. The predicted values are then plotted and compared for accuracy.

5.3. Camparative analysis

In this scientific work, the proposed strategy was compared against current approaches to establish their usefulness. The findings demonstrated that the suggested method is more precise, efficient, and appropriate for predictive analysis. The assessment is carried out using confusion matrix parameters, that are typically used to evaluate a model's performance on a dataset containing known actual values. A confusion matrix is a table that summarises the outcomes of a classification issue prediction, offering four values: True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP). Several factors are obtained from the confusion matrix to compare the strategies. The study computed three factors for each technique: accuracy, precision, and recall.

Accuracy: This parameter calculates the proportion of values obtained that agree with the true values.

Acuracy Formula = TP + TN/TP + FP + TN + FN

Precision: This parameter computes the proportion of results that are relevant.

Precision Formula = TP /TP + FP

Recall: This parameter computes the proportion of total relevant values categorised properly by the algorithm.

Recall Formula = TP/ TP + FN

6. DISCUSSION ON THE RESULTS

Python was used to implement our suggested study, together with relevant libraries such as sklearn, pandas, and matplotlib. This study's dataset was taken from uci.edu and consisted of heart disease cases. To predict cardiac disease, algorithms based on machine learning such as KNN and SVM were used. A hybrid model integrating KNN and SVM was also created to increase the uniqueness of this study. According to the findings of this study shown in table 1, the hybrid model were successful in diagnosing cardiac disease. KNN had an accuracy of roughly 75%, SVM had an accuracy of 76% , and the hybrid model had an accuracy of 81% .

Table 1. Experimental results

Algorithm	Accuracy	Precision	Recall	F1
SVM	76	80	80	80
KNN	75	80	78	79
Hybrid	81	80	89	84

Table 1 shows the performance metrics achieved by the proposed hybrid technique and based algorithms following their implementation, include recall, precision, accuracy, and f1-score.

Table 2. Accuracy results

Algorithm	Accuracy
SVM	76
KNN	75
Hybrid	81

Table 2 compares the proposed hybrid model to existing algorithms on the basis of the accuracy performance indicator. The proposed hybrid technique outperforms all existing algorithms in terms of accuracy. The suggested hybrid approach's parameter values outperform the results of other algorithms, as indicated by the graph shown in

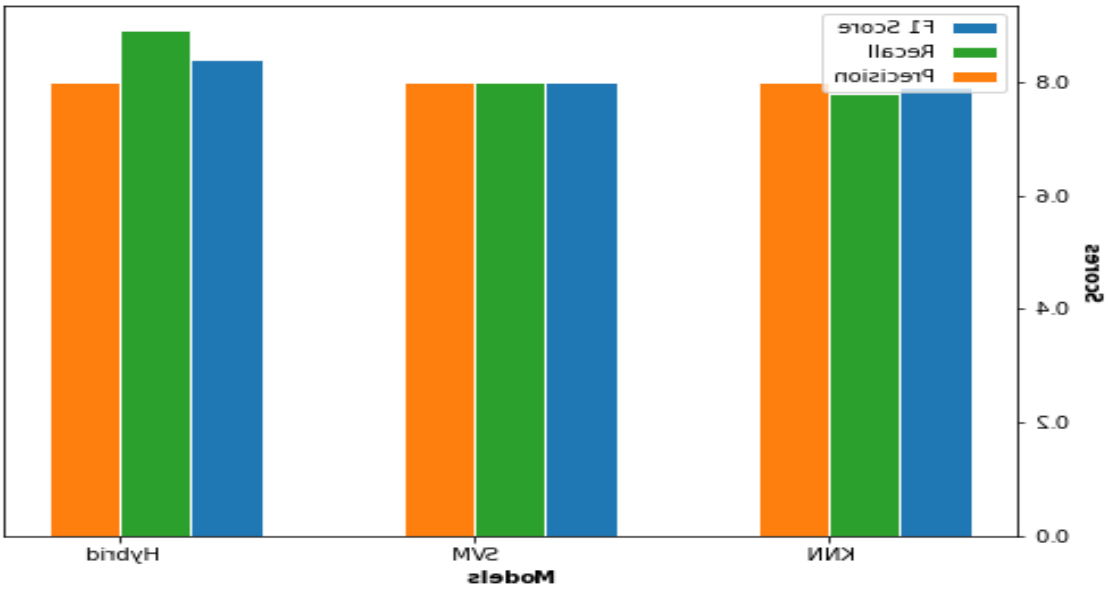

Figure 2

Figure 2. This performance analysis reveals that the suggested approach outperforms existing algorithms in predicting liver disorders.

7. CONCLUSION

The method of information mining includes analyzing crude information to reveal critical designs that can advise future applications. This strategy utilizes different classification strategies to anticipate heart disorders. Our research approach utilizes knn to extricate properties from a endless dataset and applies svm classification to create a model for predictive analysis. Compared to built up calculations such as KNN and SVM, our proposed approach yields prevalent comes about. Our investigation shows that the cross breed approach accomplishes a 81% exactness rate in anticipating heart disease prediction, outperforming the execution of other calculations on the same dataset.

REFERENCE

J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection," Informatics Med. Unlocked, vol. 17, no. October, p. 100255, 2019, doi: 10.1016/j.imu.2019.100255.

Jabbar, M. A., B. L. Deekshatulu, and Priti Chandra. "Intelligent heart disease prediction system using random forest and evolutionary approach." Journal of Network and Innovative Computing 4.2016 (2016):175- 184

Al-Milli, Nabeel. "Backpropagation neural network for prediction of heart disease." Journal of theoretical and applied information Technology 56.1 (2013): 131-135.

Hashi, E.K. and Zaman, M.S.U., 2020. "Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. Journal of Applied Science Process Engineering", 7(2), pp.631-647.

Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-48.

Alkeshuosh, Azhar Hussein, et al. "Using PSO algorithm for producing best rules in diagnosis of heart disease." 2017 international conference on computer and applications

(ICCA). IEEE, 2017.

Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-48.

"UCI Official site." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset).

P. M. Barnaghi, V. A. Sahzabi, and A. A. Bakar, "A Comparative Study for Various Methods of Classification," Int. Conf. Inf. Comput. Networks, vol. 27, no. Icicn, pp. 62–66, 2012.

Dr. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, R. Sai Sura et al. "Heart Disease Prediction using Hybrid machine Learning Model." Sixth International Conference on Inventive Computation Technologies [ICICT 2021]. IEEE