RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES
Social Science Journal

Proficient Prognostication through Hybrid Approach for Heart Disease

By

Avantika Mahadik
Research Scholar, Pacific (PAHER) University, Udaipur, India
Email- avantika_mahadik@rediffmail.com

Dr. Prashant Sharma
Associate Professor, Department of Computer Science and Engineering Pacific (PAHER) University, Udaipur, India
Email- prashant.sharma@pacific-it.ac.in

Dr. Vaibhav Narawade
Professor, Dr. D Y Patil's Ramrao Adik Institute of Technology,Maharashtra, India
Email- vaibhav.narawade@rait.ac.in

Abstract

Machine learning uses variations of methods for disease prediction. The present article is aiming to give a thorough explanation of how random forest, decision tree, liner regression are used in our research, especially when combined and applied for the heart disease prognosis. The outcomes of an experiment comparing the implementation of forecasting techniques on the same dataset. In our research, we independently experimented the dataset with random forest, decision tree, and linear regression. The DTKNN[1] is our proposed model where we hybrid two machine learning algorithms for achieving the highest accuracy for heart disease prediction. 303 records and 1025 records from different regions combined together in the DTKNN to get 100% accuracy in the prediction of heart disease. In our article we compared DTKNN with other machine learning algorithms. Based on the outcomes of our experiments and analysis, we can accomplish that planned model generated the uppermost accurateness (100%) when using decision tree and KNN together.

Keywords: Heart Disease, Machine learning algorithms, Boosting, Ensemble classifier

Introduction

An enormous amount of data is generated in today's public and business sectors world. Data is important and essential to new development. To achieve the greatest and most optimised results in any sector, we need trustworthy information. Structured data is data that has an adequate specified model, a proper structure, and organisation. Working with unstructured or raw data frequently results in a number of issues, particularly if that data is used for analysis [19].

Today, healthcare sector is known as the largest sector which has a gigantic amount of medical data. This industry collects the data from various sources like hospitals, insurance companies, pharmaceutical industries, Epidemiological Surveillance, census, other health records, sample registration system, patient's disease registries, electronic health record and

[1] DTKNN- Decision Tree and K-nearest Neighbour

clinical surveys .The collected data is stored and used for analysis for better improvement in the medical field, research in drugs and predication or diagnosis of disease [19].

Various machine learning algorithms play a vital role in the healthcare industry for premature identification of diseases. If the medical practitioner as well as patient know the disease at the preliminary stage, then it is easy to prevent it through medication or further treatment. Every machine learning algorithm is good when it is used independently according to the tasks which are distributed. We can combine two or more algorithms together for getting better predictions from various points of view, especially in accuracy. In our research study, we found that if we integrated decision tree along with KNN for two different datasets, we achieved 100% accuracy in the prediction. We further discussed confusion matrix and classification report of random forest, support vector machine, gradient boosting classifier, KNN and linear regression. We used 303 records from one region and 1025 records from another region for experiment. We showed in this article how the hybrid model (after combining decision tree and KNN together) worked effectively and gave the highest accuracy in the prediction.

Related work

According to Polat K. (2019), in [18] SMOTE[2] is used to balance the dataset; the SMOTE technique artificially raised the amount of samples for the minority class in the Parkinson's dataset. According to the author, if random forest classification used for the prediction of Parkinson will give 87.037% of accuracy. And it increased to 94.89% when SMOTE and random forest combined together.

Javeed A. et al. (2019), in [5] developed the model by hybridizing RSA algorithm with random forest and named RSA-RF. RSA-RF model achieved 93.33% accuracy in heart failure detection and also gained the success to reduce the time complexity. According to the authors, reducing time complexity could be achieved by reducing number features. Authors selected 7 prominent features in RSA-RF [19].

In 2019, Senthilkumar et al. [3] [19] combined random forest and linear model together and came with new model. The new model is identified as HRFLM. Authors in [19][3], integrated random forest and linear model for preprocessing of data. According to authors, through HRFLM 88.70% accuracy of predicting heart disease could be achieved without any restriction in features selection [3][19].

Ensemble is powerful technique which can be upturning the accuracy in predicting [4] [19]. According C.Beulah Chirsta et al.[4] [19] if weak learner can integrated with strong learner then effectiveness of weak learner can be increased. In [4] [19], authors, combined multiple classifiers into proposed system to achieved best performance in prediction [19].

In [7] author K. Budholiya highlighted that optimization of hyper parameters are very essential. Hyper parameters are those parameters, on which learning model will finally determines the values for model and stop determining further learning. Kartik Budholiya et al. in [7], came with their proposed model for predication of heart disease which is based on tuning of hyper parameter and encoding of categorical variables in dataset. Researchers combined

[2] SMOTE- Synthetic Minority Over-sampling Technique

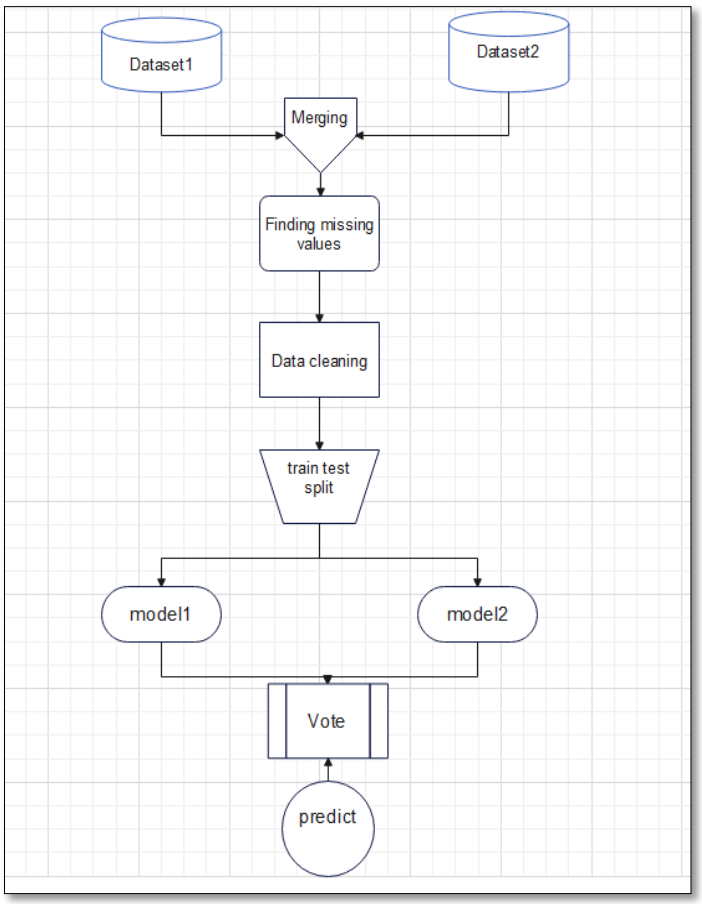
Bayesian classification and One-Hot (OH) encoding algorithm. Finding error or missing values in dataset at initial stage will give efficiency in prediction [7] [19].

Random forest, KNN, DT, and SVM algorithms have all been studied and explored by us. In several earlier studies, researchers combined their suggested model with core supervised machine learning approaches to build it. With high precision, at the early stage of heart attack, prediction is made possible by this combination. The summary of machine learning algorithms is shown in Table 7 when they are coupled with other algorithms to form the suggested model [19].

Portrayal of proposed system

The main objective of this study is to cultivate functioning system for the prediction of heart disease. For proposed system first we combined two datasets for larger and more reliable dataset, particularly from different regions. Later we integrated two machine learning algorithms for highest accuracy.

Fig 1. *Prediction of Heart Disease with DTKNN*



Description of Heart Disease Dataset

For this research work the Cleveland heart disease dataset acquired from the University of California, Irvine (UCI). According to the previous study, 13 features of heart disease are genuinely playing major role in the prediction [7]. Categorical and numerical are two forms of dataset which are explained into the following table no.1.

Table No 1. *Detailed features which are used for heart disease prediction [7]*

Feature No.	Feature Description	Type
1	Age	Numerical
2	Sex	Categorical
3	Chest pain type	Categorical
4	Resting blood pressure	Numerical
5	Serum Cholesterol	Numerical
6	Fasting blood sugar	Categorical
7	Resting electrocardiograph results	Categorical
8	Maximum heart rate achieved	Numerical
9	Exercise-induced angina	Categorical
10	St –depression	Numerical
11	St- slope	Categorical
12	Number of major vessels	Categorical
13	Thalassemia	Categorical
14	num (target value)	Categorical

Logistic Regression

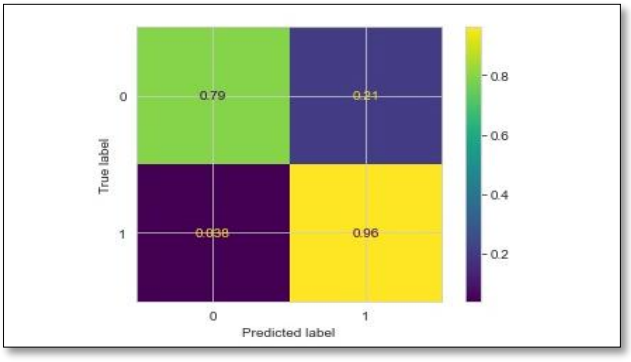
The primary application of the supervised machine learning technique, logistic regression, is classification tasks where the objective is to estimate the likelihood instance from the specific class [19][20]. Its term is logistic regression, and it is utilized for classification methods. Regression is used because it uses a sigmoid function to determine likelihood for the given class with the output of a linear regression function as input. The logistic regression predicts the likelihood that an instance will belong to a specific class or not, whereas the output of linear regression is a continuous value that can be anything. [19][20]. By merging both the datasets with best hyper parameters we have got an accuracy score of 0.8855421686674688

Fig 2. *Classification report for LR*

LR accuracy score: 0.8855421686746988

	precision	recall	f1-score	support
0	0.94	0.79	0.86	150
1	0.85	0.96	0.90	182
accuracy			0.89	332
macro avg	0.90	0.88	0.88	332
weighted avg	0.89	0.89	0.88	332

Fig 3. *Confusion matrix for LR*



Random Forest (RF)

A forest is made up of many different types of trees, and the more trees there are, the more robust the forest will be. Similar to this, the accurateness and problem-solving competence of a random forest improves through the algorithm's quantity of trees develops. In this technique various classifiers are integrated for the resolving the stimulating issues and improve model performance. Ensemble is another way of referring to this blending of various models and it can possible through bagging and boosting. Bagging is the process of generating a different training subset via replacement from a sample training dataset. The outcome is decided by a majority vote. Turning weak learners into strong ones by building successive models to achieve maximum possible accuracy is called boosting [19].

Steps involved in the random forest algorithm are [19],

Step 1: Pick random instances from the known data collection or training set.
Step 2: Build a decision tree for every training set of data.
Step 3: Voting will be conducted using an average of the decision tree.
Step 4: At the end, select the guess result that received the most votes.

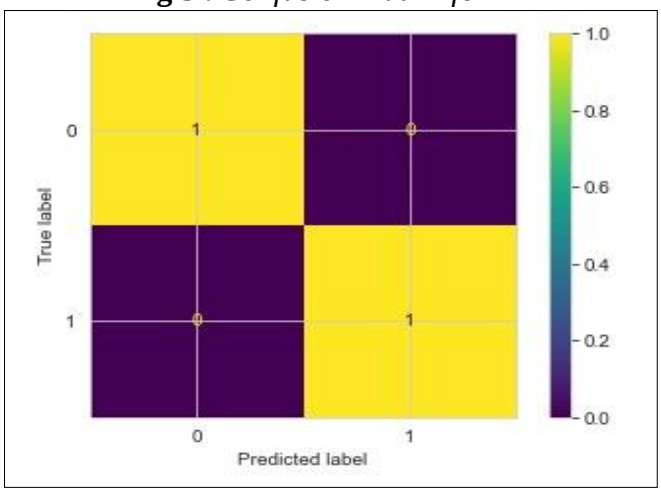
In this way random forest works and gives the better accuracy in the prediction of diseases.

Fig 4. *Classification report for LR*

RFC accuracy score: 1.0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	150
1	1.00	1.00	1.00	182
accuracy			1.00	332
macro avg	1.00	1.00	1.00	332
weighted avg	1.00	1.00	1.00	332

Fig 5 . *Confusion matrix for LR*



Decision Tree (DT)

This is consisting with leaf nodes, root nodes, decision nodes and branches. The probable outcome shows by leaf node. Leaf node does not split into further sub nodes. Whereas decision node having several branches and all are used to generate conclusions. This algorithm is more use for classification. Regression also possible with this algorithm [19].

Decision tree algorithm perform following steps [19],

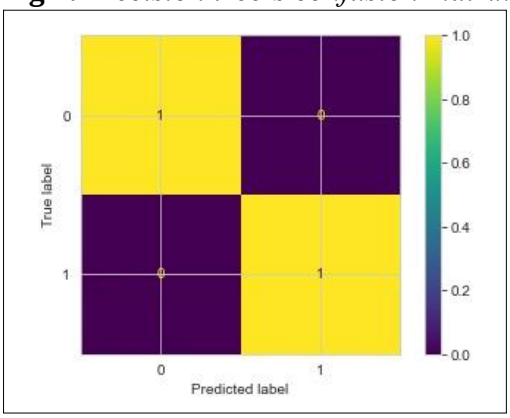
Step 1: Start the tree from the root node.
Step 2: Identify dataset's top attribute.
Step 3: Subset the root node to include probable outcomes as the finest traits.
Step 4: Form the decision tree node which has the finest feature in the step.
Step 5: Use the selections of the dataset generated in step 3 to iteratively develop new decision trees. Continue all the above iterations till you come to that point where you cannot further classify the nodes, and you can denote the last node as a leaf node.

Fig 6. *Classification report for DT*

DT accuracy score: 1.0

	precision	recall	f1-score	support
0	0.94	0.79	0.86	150
1	0.85	0.96	0.90	182
accuracy			0.89	332
macro avg	0.90	0.88	0.88	332
weighted avg	0.89	0.89	0.88	332

Fig 7. *Decision tree's confusion matrix*



According to the matrix, decision tree gives the best result for the true negative values which is better than any other models of machine learning for the prediction of heart disease.

Gradient Boosting Classifier

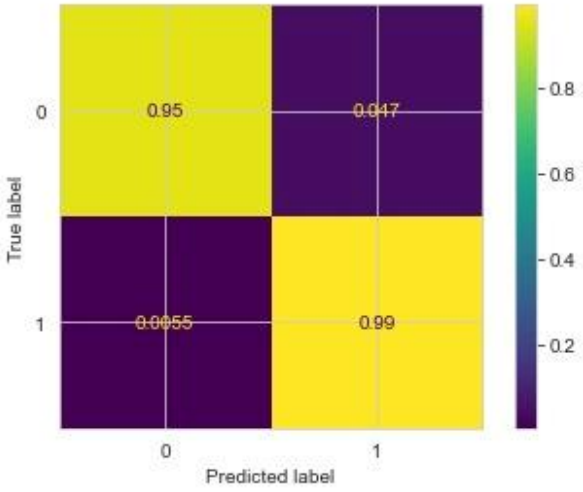
Each stage involves fitting n_classes_ regression trees on the loss function's negative gradient, which could be a binary or multiclass log loss. In the specific scenario of binary classification, a single regression tree is generated [21].

Fig 8. *Classification report of Gradient Boosting*

GradientBoostingClassifier accuracy score: 0.9759

	precision	recall	f1-score	support
0	0.94	0.79	0.86	150
1	0.85	0.96	0.90	182
accuracy			0.89	332
macro avg	0.90	0.88	0.88	332
weighted avg	0.89	0.89	0.88	332

Fig 9. *Confusion matrix for Gradient Boosting*



KNN (K-nearest Neighbour)

Agriculture, economics, text mining, and healthcare all make use of KNN algorithm for prediction. The basic goal of the KNN method is to anticipate a new sample point's categorization using data points that have been divided into various groups. It is distance-based and categorises things according to the classes of their close neighbours. KNN is most frequently utilised for classification tasks. The number of labelled points (neighbours) taken into account for classification is indicated by the parameter k in KNN. The number of these points needed to calculate the outcome is indicated by the value of k. calculating the distance and determining which categories are closest to our unidentified thing are our tasks [19].

Following are KNN Algorithm steps [19]:

Step 1: Consider labelled data

Labelled data is also called training set which are used to train the model. You can either use the labelled databases found in open sources like this one or manually label data to generate the training set.

Step 2: Searching for K-nearest neighbours

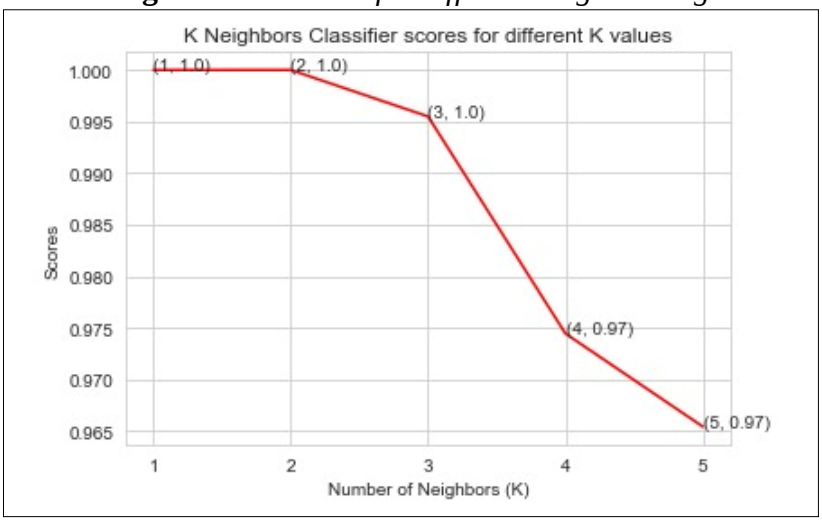
Finding a record's k-nearest neighbours entails locating the records that share the most features with it. This process is often referred to as the distance calculation or similarity search.

Step 3: Group the points

In classification problems, the algorithm chooses a class or group label by a majority vote, which means that the label that appears more frequently in neighbours is used. The K-distance is the separation between a specified query point and a set of data points. We must choose a distance measure in order to calculate it. Distance measurement techniques are several. Hamming distance, Minkowski distance, Manhattan distance and Euclidean distance are the popular metrics used to measure k-distance [19].

For a single dataset, the mean score for KNN is 0.9941176470588236 at neighbour 1 which gives the best result than other neighbours, but when datasets were merged, the modal was able to achieve better mean scores for your KNN model, which means it is now better at making predictions based on the data. Now the mean score is 1.0 in neighbouring 1, 2, and 3. Following figure 10 is showing KNN scores for different neighbouring.

Fig 10. *KNN scores for different neighbouring*



Support Vector Machine (SVM)

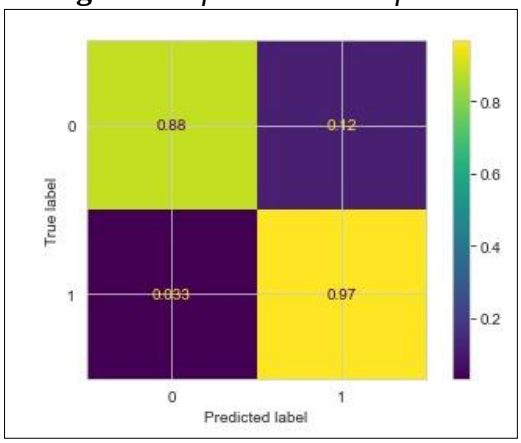
Support vector machine algorithm developed by Vladimir Vapnik. The main objective of this algorithm is to find out "hyperplane". Hyperplane is the accurate plane or line or decision boundary which can classify the number of features (is also referred by n-dimensional space) into sets. In order to build the hyperplane, support vector machine picks the extreme points and vectors. Support vectors are the term used for these extreme examples, and the support vector machine technique is named after them [19].

Fig 11. *Classification report for SVM*

SVM accuracy score: 0.9277

	precision	recall	f1-score	support
0	0.94	0.79	0.86	150
1	0.85	0.96	0.90	182
accuracy			0.89	332
macro avg	0.90	0.88	0.88	332
weighted avg	0.89	0.89	0.88	332

Fig 12. *Confusion matrix of SVM*



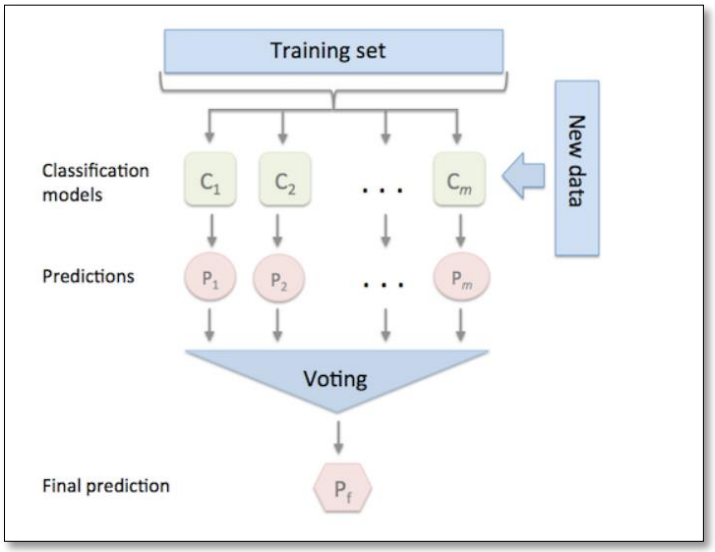
Ensemble Vote Classifier:

A popular machine learning technique termed ensemble learning combines numerous models to produce predictions that are more accurate than those produced by any one model

alone. Voting classifiers are a popular ensemble learning technique that combines the predictions of various independent models to provide a final prediction. In the voting classifier, it first trains several different models on the same dataset, using different algorithms or hyper parameters. These models can be of different types with different strengths and weaknesses. Hard and soft voting are two categories. In hard voting, the prediction of each model is counted as a vote and majority votes are used as the final prediction. On the other hand, in soft voting, each model's prediction is weighted by its confidence, and the weighted average is used as the final prediction. It means in soft voting majority of votes as well as level of confidence of each individual model would be considered for the prediction.

When predicting the target variable from complicated datasets with numerous features and numerous observations, the ensemble voting classifier is an effective tool. The voting classifier can improve comprehensive performance and minimize the risk of overfitting. Through ensemble voting, we can achieve highest performance by combining the advantages of various models and minimizing the drawbacks of individual models.

Fig. 13 *Ensemble Vote Classifier [17].*



By leveraging their supportive strengths and lowering their flaws, incorporating decision tree and KNN models can be an effective strategy to boost the performance of predictive models. Proposed model uses total 2 models

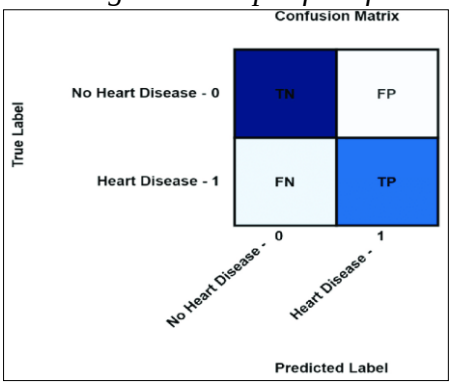
1. Decision tree
2. K – Nearest Neighbour

We have used a soft voting technique for the proposed model with equal weights.

Evaluation metrics

We examined the performance of proposed system with different evaluation variables. The percentage of the entire subjects correctly classified is referred to as accuracy. The percentage of those who test positive who actually have the condition is known as sensitivity. The percentage of those who test negative who do not have the condition is known as specificity. Recall and sensitivity are the same. The proportion of participants accurately classified as positive out of all those classified as positive is known as precision. A harmonic mean of recall and precision is the F1-Score [7].

Figure 14. *Showing the concept of confusion matrix [7]*



Following are the formulas for based on confusion matrix. Where, TP=true positive, FP = false positive, TN=true negative, FN= false negative

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$F1 = 2 \times \text{precision} \times \text{recall} / \text{precision} + \text{recall}$$

$$\text{Accuracy} = (TP + TN)/(TP + FN + TN + FP)$$

$$\text{Specificity} = TN + (TN + FP)$$

Experimental Results

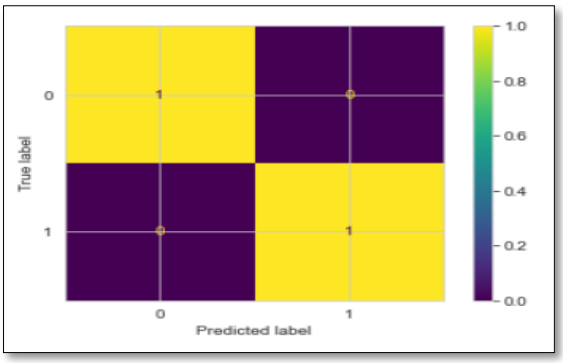
In this paper, the ensemble voting classifier technique has been employed as a classifier for training and testing. As shown in the tables below, the proposed models have the best accuracy of all of them, with an accuracy of 100%. We used 2 models for combined prediction, and we used different weights for each model. The highest accurate predictions were given more weight than the other models.

Fig 13. *Accuracy prediction score with hybrid model(DTKNN)*

Hybrid accuracy score: 1.0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	123
1	1.00	1.00	1.00	134
accuracy			1.00	257
macro avg	1.00	1.00	1.00	257
weighted avg	1.00	1.00	1.00	257

Fig 14 . *Confusin matrix of proposed model(DTKNN)*



Prediction accuracy based on 1025 rows of dataset :

Table 2. *Showing accuracy of proposed model for 1025 records*

Models	Accuracy
Logistic regression	0.8482
Random Forest Classifier	0.8716
MLPClassifier	0.9079
Gradient Boosting	0.9728
KNN	0.9942
Decision tree	1.0000
SVM	0.8560
Proposed Model(DT + KNN)	1.0000

Table 3. *Showing comparative result of DTKNN with other models for 1025 records*

Model	Accuracy	Sensitivity	Precision	Specificity
Logistic regression	0.8482	0.85	0.8292	0.8467
Random Forest	0.8287	0.8319	0.8260	0.8048
Gradient Boosting	0.9728	0.9754	0.9674	0.9703
MLPClassifier	0.8482	0.85	0.8292	0.8467
Decision tree	1.0000	1.0000	1.0000	1.0000
KNN	0.9610	0.9248	1.0	1.0
SVM	0.8560	0.8706	0.8439	0.8211
Proposed Model(DT + KNN)	1.0	1.0	1.0	1.0

Table 4. *Showing accuracy of proposed model for 303 records*

Models	Accuracy
Logistic regression	0.9077
Random Forest	0.86
MLPClassifier	0.9079
Decision tree	0.88
SVM	0.88
Proposed Model	0.9211

Table 5. *Showing comparative result of DTKNN with other models for 303 records*

Model	Accuracy	Precision	Recall	F1	Specificity
Logistic regression	0.9077	0.8823	0.909	0.901	0.9069
Random Forest	0.8684	0.8235	0.875	0.85	0.8636
MLP Classifier	0.9079	0.9422	0.9011	0.89	0.8823
Decision tree	0.8815	0.8235	0.9032	0.86	0.8666
SVM	0.8815	0.8235	0.90322	0.86	0.8666
Gradient Boosting	0.7895	0.7352	0.7812	0.79	0.7954
Proposed model	0.9211	0.8823	0.9375	0.93	0.909

After merging both the datasets for better prediction from two different regions, we received the following results. In table 6, we are showcasing accuracy prediction after combined the two datasets.

Table 6 . *Showing results of proposed system after integrating two datasets*

Models	Dataset accuracy		
	303 Records	1025 Records	1328 Records
Logistic regression	0.885	0.8482	0.885
Random Forest Classifier	0.86	0.8716	1
MLP Classifier	0.9079	0.9079	0.9849
Gradient Boosting	0.79	0.9728	0.9759
KNN	0.84	0.9888	1
Decision tree	0.92	0.9144	1
SVM	0.88	0.856	0.9277
Proposed Model DTKNN	0.9211	1	1

When two datasets come together, a bigger and diversified dataset is generated, which could improve the precision and dependability of predictions when they are used for training and testing. By combining two datasets, we can significantly expand the amount of dataset available to train the model and testing it. Such additional knowledge could lower the problem of overfitting. And that is why model could learns to predict training data too accurately and fights to adapt to new, unexpected results.

To address any gaps or biases in individual datasets, databases might be combined. Combining a dataset that covers a different demographic or geographical region with one that is biased towards a certain demographic or geographic area, for instance, can help to balance out the bias. It is necessary to ensure that the data is compatible and consistent before combining it. Additionally, we must think about any possible problems with data quality or the data duplication brought on by merging databases.

Overall, blending datasets can be an excellent way for generating a wider, more diversicity into the dataset which could be used in training and testing of the model and may produce superior predictions.

Conclusion

The accuracy and resilience of prediction models can be significantly enhanced by combining the two algorithms in the proposed model. With the ability to separate repeatedly the input data into subsets based on the most essential characteristics, the decision tree model, a common classification and regression technique, gets at a prediction or decision. It is a comprehensive and understandable model that can handle interactions and non-linear relationships among features. KNN model is a modest and strong classification as well as regression technique which operates by locating the k-nearest neighbours of a given input data point and applying their labels or values to predict the output. A non-parametric model, the underlying link between the features and the output does not have to take any particular functional form.

Decision tree and KNN , these two models can be joined to enhance around their flaws while exploiting their strengths. KNN is able to identify local patterns and anomalies in the data, while decision trees can handle complex and non-linear interactions among the features. We may produce a more precise and reliable prediction by employing a decision tree to divide the data into subsets, followed by the application of KNN to each subset.

Table 7. *Table showing summarization of background work done with integrated approach [19]*

References	List of algorithms which are used	Combined with other	Proposed model strategy and name	Dataset used for experiments	Performance result of proposed model (% of Accuracy)
[3]	RF,LM	-	Combined RF and LM in HRFLM combining	UCI repository	88.70
[4]	PART classifier, RF, Bayes Net , Naïve Bayes and MLP	-	Naïve Bayes+ Bayes Net+ RF+ Multilayer Perception	Cleveland data set from UCI repository	85.48
[5]	RF	RSA ,Grid Search	RSA-RF	Cleveland data set from UCI repository	93.33
[6]	Vote, SVM, LR, Neural Network, DT , Naïve Bayes and KNN	-	LR, Naïve Bayes combined with Vote	Cleveland dataset, Statlog dataset from UCI repository	87.41
[7]	Bayesian Classification	One-Hot (OH)encoding algorithm	Combined Bayesian classification with OH encoding	Cleveland data set from UCI repository	91.80
[8]	RF, SVM	CNN based deep learning model Inception Net, Residual Net	CNN based xception net model	Samples of Heart sound collected from https://github.com/yaseen21khan/Classification-of-Heart-Sound-Signal-Using-Multiple-Features	99.43
[18]	SMOTE , RF	Combined SMOTE and RF	For classification and balancing values of dataset	Cleveland data set from UCI repository	94.89

References

[1] Chang, V., Bhavani V. R., Xu A. Q., & Hossain M. A. (2022). "An artificial intelligence model for heart disease detection using machine learning algorithms", Healthcare Analytics, 2, 100016.

[2] Ali M. M., Paul B. K., Ahmed K., Bui, F. M. Quinn J. M., & Moni M. A. (2021). "Heartdisease prediction using supervised machine learning algorithms: Performance analysis and comparison". Computers in Biology and Medicine, 136, p.104672.
[3] Mohan S., Thirumalai C., & Srivastava G. (2019). "Effective heart disease prediction using hybrid machine learning techniques". IEEE access, vol.7, pp.81542-81554.
[4] Latha, C. B. C., & Jeeva S. C. (2019). "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques". Informatics in Medicine Unlocked, vol.16, p.100203.
[5] Javeed A., Zhou S., Yongjian L., Qasim I., Noor A., & Nour R. (2019). "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection". IEEE Access, vol.7, pp.180235-180243.
[6] Amin M. S., Chiam Y. K., & Varathan K. D. (2019). "Identification of significant features and data mining techniques in predicting heart disease". Telematics and Informatics, vol.36, pp.82-93.
[7] Budholiya K., Shrivastava S. K., & Sharma V. (2020). "An optimized XGBoost based diagnostic system for effective prediction of heart disease". Journal of King Saud University-Computer and Information Sciences.
[8] Roy T. S., Roy J. K., & Mandal N. (2022). "Classifier identification using Deep Learning and Machine Learning Algorithms for the detection of Valvular Heart diseases". Biomedical Engineering Advances, pp.100035.
[9] Uddin S., Khan A., Hossain M. E., & Moni M. A. (2019). "Comparing different supervised machine learning algorithms for disease prediction". BMC medical informatics and decision making, vol.19 (1), pp.1-16.
[10] Mienye I. D., Sun Y., & Wang Z. (2020). "An improved ensemble learning approach for the prediction of heart disease risk". Informatics in Medicine Unlocked, vol.20, pp.100402.
[11] Saxena K., & Sharma R. (2016). "Efficient heart disease prediction system". Procedia Computer Science, vol.85, pp.962-969.
[12] Tripoliti E. E., Papadopoulos T. G., Karanasiou G. S., Naka K. K., & Fotiadis D. I. (2017). "Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques". Computational and structural biotechnology journal, vol.15, pp.26-47.
[13] Uyar K., & İlhan A. (2017). "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks". Procedia computer science, vol.120, pp.588-593.
[14] Pathan M. S., Nag A., Pathan M. M., & Dev S. (2022). "Analysing the impact of feature selection on the accuracy of heart disease prediction". Healthcare Analytics, vol.2, pp.100060.
[15] Shailaja K., Seetharamulu B., & Jabbar M. A. (2018, March). "Machine learning in healthcare: A review". In 2018 Second international conference on electronics, communication and aerospace technology (ICECA), pp. 910-914.
[16] Khan M. A. (2020). "An IoT framework for heart disease prediction based on MDCNN classifier". IEEE Access, vol.8, pp.34717-34727.
[17] https://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/
[18] Polat K. (2019, April). "A hybrid approach to Parkinson disease classification using speech signal: the combination of smote and random forests". In 2019 scientific meeting on electrical-electronics and biomedical engineering and computer science (EBBT), pp. 1-3,IEEE.

[19] Dr. Prashant Sharma, Dr. Vaibhav Narawade and Avantika Mahadik (2023): "A comparison and analysis of supervised machine learning algorithms towards accurate predicting of heart diseases" DOI: http://biogecko.co.nz/.2023.v12.si01,pp.1-8
[20] www.geeksforgeeks.org
[21] https://github.com/topics/gradient-boosting-classifier