Heartbeat Feature Extraction for Chronic Cardiovascular Disease Prediction via Computer Vision

Jyoti Tiwari
Research Scholar
Dept. of CSE
Sagar Institute of Research &
Technology, Bhopal

Dr Ritu Shrivastava
Head & Dean
Dept. of CSE
Sagar Institute of Research &
Technology, Bhopal

Prof Rupali Chaure
Research Scholar
Dept. of CSE
Sagar Institute of Research &
Technology, Bhopal

ABSTRACT: Due to the difficulties in diagnosing heart illness, it is now one of the world's most ultra-hazardous and severe diseases. Machine learning may assist the medical field by providing accurate and timely illness diagnoses. Within the healthcare industry, a popular approach used for analyzing huge amounts of data is Data mining. This paper examines many aspects of heart illness and develops a model using supervised learning techniques like Logistic Regression, Naïve Bayes, decision trees, K-nearest neighbor (KNN), random forest, Support Vector Machine (SVM), and the proposed method. The proposed approach is an organization and regression approach that used ensemble learning. The recommended method is based on maximizing an objective arbitrary loss function to develop and generalize the ensemble model stage by step. It utilizes a dataset of the Cleveland catalog of heart patients at UCI. There are around 303 occurrences plus 76 characteristics within the collection. Only 14 out 76 characteristics are tested, despite their importance in proving the effectiveness of alternative algorithms. We employed an ensemble of SVM, KNN, and ANN in the suggested approach to obtain an accuracy of 95.61%. In comparison to other authors' models, which included Naive Bayes, Decision Trees, and Support Vector Machine classifiers, KNN, Random Forest, the UCI heart disease dataset had an accuracy of 95.61 percent, sensitivity of 92.3 percent, Precision 83.7 percent, and specificity of 93.8. Percent.

INDEX TERMS Heartbeats classification, estimation · Data mining · Decision tree · Random forest · Electrocardiogram (ECG). Naïve Bayes · K-NN · Artificial Intelligence

I. INTRODUCTION

Heart disease instances are rising at an alarming rate, and it's critical and worrying to be able to predict such diseases in advance. This is a tough diagnostic to make, therefore it must be done correctly and quickly. In the modern world, heart disease prediction has become a key idea that is influencing society's health. The primary idea is to use the Random forest algorithm to determine the age group and heart rate. Our research explains how a system calculates the heart rate and condition depending on the user's inputs, such as blood pressure and other factors. (Jindal et al., 2021)When compared to other algorithms, RFA implementation delivers a better experience and more accurate results. This aids in the early detection of sickness and is utilized in a variety of ways, including providing input to determine the heart rate based on the health state. The use of invasive methods to diagnose heart disease is based on medical professionals analyzing the patient's medicinal history, physical checkup description, and examination of worrying symptoms. Due to human error, there is frequently a delay in diagnosis. Due to these limitations, scientists have resorted to new techniques for illness prediction, such as Data Mining and Machine Learning. Data mining is critical in developing an intelligent model for a medical system to identify heart disease (Akyildiz et al., 2006) using a patient dataset that includes risk factors linked with the condition. Medical professionals may be able to assist with the detection. Researchers have developed several software tools and methods for building an effective medical decision support system. Machine learning assists computers in learning and acting appropriately. It allows the computer to learn a complicated model and forecast data, as well as do complicated math calculations on large amounts of data. The heart disease prediction systems based on machine learning will be exact and decrease the risk.(Farran et al., 2013)

The goal of this paper is to see if the patient is at risk for cardiovascular heart disease based on their medical characteristics such as gender, age, chest discomfort, fasting sugar level, and so on. A dataset with the patient's medical history and distinctiveness is selected from the UCI repository. Thus, to be done, we classify the patients base on 14 therapeutic distinctiveness to see if he/she are at risk of heart disease. To prevent bias in training the method and testing its occurrences, the dataset is split into two halves (70 percent /30 percent, training/testing). 70% of the data is utilized to teach the machine learning method, while the residual 30% is utilized to assess the suggested activity categorization system's performance. For the data acquired from the UCI repository, we utilized python and pandas' operations to conduct heart disease categorization(Heydari et al., 2016)

The advantages of the method are: The effects of cardiovascular disease are difficult to predict. Data mining techniques do not assist in making smart decisions. Large datasets for patient records are too large to handle. Increased precision for successful cardiac disease detection is an advantage. Using the proposed method and feature selection, it handles the roughest (enormous) quantity of data. Doctors' time complexity should be reduced. For patients, this approach is also cost-effective.(Bansal et al., 2015)

II. Literature Review-

In the biomedical sector, machine learning has previously been used in diabetic prediction (C & G.M, 2015), the association of heart disease and diabetes (Bhramaramba et al., 2011), and the study of diabetes proteins (King, n.d.,1992). The practice of gathering helpful data and information from the bulky database is acknowledged as data mining. To categories numerous heart disease variables in guessing heart disease, different data mining procedures likewise

- regression
- association rule,
- clustering

Classification methods such as

- Naive Bayes, ,
- random forest,
- K-nearest neighbors
- Decision tree are employed.

The classification approaches are compared using a comparison analysis (Ho, 1995).

Random forests, as well recognized as random decision forests, are a band learning scheme for classification, regression, and additional responsibilities that works by training a huge number of decision trees of the modules (classification) or the mean estimate (regression) of the precise trees (Ho, 1998). The model was assembled on two random forest constraints, which are the utmost significant RF parameters.(Bahrammirzaee, 2010)

Forecasting, commercial modeling, economics, medicinal applications(Hoptroff, 1993), and other disciplines have used neural networks extensively (Azar, 2013)(Brown et al., 2000). These techniques have also proved effective(Furey et al., 2000) in bioinformatics (Ranawana & Palade, 2005)(Yasdi, 2000). Further ANN model solicitations are addressed in (Wu & Vapnik, 1999) and the references therein.

Support vector machine (SVM) was initially proposed by (Policy, 2019) in his work on statistical learning theory, SVM abbreviated as supervised machine learning approach that is used for categorization and regression.(Kanmani et al., 2007) The eight-quality metrics were used to assess the categorization models (Blake, 2007). These quality metrics for classification analysis were examined. Samples that did not have heart disease were classified as positive, while those that did have heart disease were classified as negative. Table 1 shows different methods and accuracy proposed by different authors Logistic regression is a real-valued input vector discriminative classification approach(Learning & Kidwell, n.d.,1997). Features or predictors are the measurements of the input vector to be classified.

The k-nearest neighbor (KNN) technique is an instance-dependent learning method that does not begin with a fully developed hypothetical model. As an alternative, it employs a simpler notion, with the nearest instance in the input space most likely belonging to the same class(Celesti et al., 2017) .

Multiple types of heartbeats were investigated in(Alarsan & Younes, 2019) , and the author achieved a 93.4 percent accuracy rate. For the classification of ECG beat types, the author developed a convolution neural network.

DNN was used to categorize heartbeats using deep learning (Deep Neural Network). According to the author, categorizing accuracy reached 99 percent; however, there were only two categories of classification and the dataset dimensions were around 84,900 entries(Deepika & Seema, 2017). The study of the hidden Markov model shown that that emerging a fast classifier for heartbeat classification is possible. (Deepika & Seema, 2017)

III. Methodology- Supervised machine learning classifiers come in a variety of shapes and sizes. Some examples of these methods include naïve Bayes, linear discriminate analysis (LDA) and quadratic discriminate analysis (QDA), generalized linear models, stochastic gradient descent, support vector machine (SVM), linear support vector classifier (Linear SVC), decision trees, neural network models, nearest neighbors, and other approaches. Collaborative methods bring together a group of weak learners to generate a group of robust users(Sharma et al., 2020)The mission of these calculations is to expand accuracy. This can be achieved using two strategies. One of the strategies is the use of feature engineering, and the other strategy is the use of boosting algorithms. Boosting algorithms concentrate on those training observations which end up having misclassifications. There are five vastly used boosting methods, which include AdaBoost, CatBoost, LightGBM, XGBoost and gradient boosting. To do so, two methods may be employed. Among the approaches used are feature engineering and boosting strategies. The training observations that cause misclassifications are the target of boosting algorithms.(Ho, 1998)

Table 1: Literature review in terms of Accuracy

Title	**Year**	**Method**	**Accuracy**
M.Kumari et.al. (Kumari & Godara, 2011)	2011	Support Vector Machine, Artificial neural networks (ANNs), Decision Tree, and RIPPER classifier	84.7
C.S. Dangre et.al. (S.Dangare & S. Apte, 2012)	2012	Decision Trees, Naive Bayes, and Neural Networks	90.74
V.Chaurasia et.al. (Vikas & Saurabh, 2013)	2013	Tree and Bagging algorithm	85.03
P.Purushottam et.al. (Purushottam et al., 2016)	2015	Decision tree classifier	87.3
Bahrami b. et.al, (Bahrami & Shirvani, 2015)	2015	J48, Naïve Bayes, KNN, SMO	83.73
K.Dwivedi et.al. (Dwivedi, 2018)	2018	Naïve Bayes KNN Logistic Regression Classification Tree	83% 80% 85% 77%
Ibrahim fair et.al. (Alarsan & Younes,	2019	Gradient-Boosted Trees model	93.4

2019)			
S.Devansh et.al. (Patel et al., 2021)	2020	Naïve Bayes K-NN Decision tree Random forest	88.157% 90.789% 80.263% 86.84%

Table 2: Heart Disease Characteristics Datasets(Patel et al., 2021)

Sr.no.	**Attribute**	**Representative Icon**	**Details**
1	Age	AGE	Patient age (In years)
2	Sex	SEX	Gender of patient (male-0 female-1)
3	Chest Pain	CP	Chest pain type
4	Rest blood pressure	TRESTBPS	Resting blood pressure (in mm Hg on admission to hospital, values from 94 to 200)
5	Serum cholesterol	CHOL	Serum cholesterol in mg/dl, values from 126 to 564)
6	Fasting blood sugar	FBS	Fasting blood sugar>120 mg/dl, true-1 false-0)
7	Rest electrocardiograph	RESTECG	Resting electrocardiographic result (0 to 1)
8	Max Heart rate	THALCH	Maximum heart rate achieved (71 to 202)
9	Exercise-induced angina	EXANG	Exercise included agina(1-yes 0-no)

10	ST depression	OLDPEAK	ST depression introduced by exercise relative to rest (0 to .2)
11	Slope	SLOPE	The slop of the peak exercise ST segment (0 to 1)
12	No. Of vessels	CA	Number of major vessels (0-3)
13	Thalassemia	THAL	Defect types; 3—normal; 2—fxed defect; 1—reversible defect
14	Target	TARGET	0 or 1

3.1 Gradient Boosting:

Gradient boosting (GB) is a technique that produces new concepts incrementally from a traditional model, to lower the loss function with each new model. To calculate this loss function, the gradient descent method is utilized. When the loss function is applied, the respectively innovative model fits the data with additional accuracy, resulting in increased overall accuracy as shown in figure 1. Boosting, on the other hand, must come to a halt at some point, or the model will over fit. As a termination criterion, you might use a prediction accuracy threshold or a maximum number of models created.(Alarsan & Younes, 2019)

3.2 DESCRIPTION AND COLLECTION OF DATA

The Irvine's machine learning repository data came from the University of California. This machine learning repository UCI data has a set called Heart Disease Data Set. The machine learning repository at UCI includes a huge and diversified collection of datasets from a variety of areas. The machine learning community, ranging from beginners to experts, regularly uses this data to experimentally interpret data. Several academic papers and research initiatives have made use of this resource. Only 14 of the dataset's 76 properties are used in this analysis; the 14 attributes are presented in Table 2.(King, n.d.)

3.3 TOOLS USED

This research makes use of a variety of tools. They're all free of charge and open source.

- Python 3.5
- NumPy 1.11.3
- Matplotlib 1.5.3
- Pandas 0.19.1
- Seaborn 0.7.1

- SciPy and Scikit-learn 0.18.1

Python is a high level general programming language and is very widely used in all types of disciplines such as general programming, web development, software development, data analysis, machine learning etc. Python is used for this project because it is very flexible and easy to use and also documentation(Patel et al., 2021) and community support is very large. NumPy is very powerful package which enables us for scientific computing. It comes with sophisticated functions and is able to perform N-dimensional array, algebra, Fourier transform etc. NumPy is used very where in data analysis, image processing and also different other libraries are built above NumPy and NumPy acts as a base stack for those libraries.(Sharma et al., 2020)

3.4 PROPOSED METHODOLOGY FLOWCHART AND ALGORITHM

XGBoost and Gradient Boosting Machines (GBMs) are also both composite tree approaches that use the fundamental GBM framework through system optimization and algorithmic improvements. The foundation learners, or models that make up the ensemble, might be from the same learning algorithm or from separate learning algorithms. Bagging and boosting are two types of ensemble learners that are commonly employed. Though these two approaches may be used to a variety of statistical models, decision trees have been the most popular.(Vikas & Saurabh, 2013)

Accuracy is a degree that replicates how effectively your algorithm predicts the true positives within your system's total positives. The number of genuine positives collected by our system is determined by categorizing them as genuine positives. When the data is skewed, F-measure is preferred over accuracy because it provides a good balance of precision and recall. As a consequence, the formula in F-measure was utilized as a performance statistic as shown in equation (1) for other parameters like sensitivity, specificity, and precision shown in Equations 2,3, and 4.(Khan et al., 2021)

$$\text{Accuracy}(ACC)=\frac{TP+TN}{TP+TN+FP+FN}\text{X } 100\% \qquad (1)$$

From total samples, the capability of the classifier model to expect correct positive and negative classes is calculated by accuracy.

$$Sensitivity\ (SEN)=\frac{TP}{TP+FN}\times 100\% \qquad (2)$$

It enumerates the capability of the classifier model in expecting positive class labels correctly.

$$Specificity\ (SPE)=\frac{TN}{TN+FP}\times 100\% \qquad (3)$$

Contrary to sensitivity, the classifier model can envisage undesirable instances correctly.

$$Precision=\frac{TP}{TP+FP}\times 100\% \qquad (4)$$

Precision calculates the amount of positive class estimates that truly fit the positive class.

Where TP stands for True Positive, FP for False Positive, and FN for False Negative. The proposed approach is a classification and regression approach that uses ensemble learning. It is proficient in producing an active model made up of weak learners, such as decision trees. The suggested technique is based on maximizing an objective arbitrary loss function to develop and simplify the collaborative model stage by step. In an iterative process, the suggested approach builds its model from the prior loss function of the negative gradient.

Along with accuracy for proposed model sensitivity and Minimizing the loss function is a critical issue in machine learning, and it must be optimized. The loss function, in other words, reflects the difference between the expected output and the target. A low loss function value indicates a good prediction or classification result. The model is enchanting in a positive path, which is the Gradient of the loss function, once the loss function drops consecutively and repetitively.

1.5 Algorithm Steps:

XGBoost is a machine learning method that has lately dominated Kaggle tournaments for structured or tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. Author(Mariot et al., 1964) proposed the XG Boost algorithm, which is based on the GBDT structure. XG Boost is a computational intelligence challenging issue that uses a configurable end-to-end tree strengthening system. Distributed and parallel data processing speeds up learning, permitting for much more rapid model investigation. It has received considerable attention as a result of its great success in Kaggle's ML contests (Mohan et al., 2019) Unlike the GBDT, the XG Boost optimization process includes a convolution operation to more per. The goal function can be described as continues to follow:

$$O=\sum_{i=1}^{n} L(y_i,F(x_i))+\sum_{k=1}^{\tau} R(f_k)+C \qquad (5)$$

where $R(f_k)$ signifies the regularization tenure at the k-time repetition and C is a continuous term, the regularization term $R(f_k)$is articulated as

$$R(f_k)=\alpha H+\frac{1}{2}\eta\sum_{j=1}^{H} w_j^2 \tag{6}$$

Where, w_j is the output result of each tree structure, H is the portion of the plant, denotes the punishment parameter, and symbolizes the complex nature of leaves. The blossoms, in particular, denote the performance of the predicted components of the classification rules, while the leaf node symbolizes the tree node that can then be split. Furthermore, rather than using the first-order derivative as in the Gradient Boost Decision Tree, XGBoost uses a second-order Similarity transformation of objective functions. If the Euclidean distance is the mean square error, then the intellectually function can be inscribed as

$$O=\sum_{i=1}^{n}[P_i w_{q(x_i)}+\frac{1}{2}(q_i w_{q(x_i)}^2)]+\alpha H+\frac{1}{2}\eta\sum_{j=1}^{H} w_j^2 \tag{7}$$

Where $(q(x_i))$ indicates a function that assigns data points Therefore, the objective function is also expressed as

$$O=\sum_{\mathrm{j}=1}^{\mathrm{T}}[P_j w_j+\frac{1}{2}(Q_j+h)w_j^2]+\alpha H \tag{8}$$

Where $P_j, Q_j=\sum i\varepsilon I_j$,and q_j and I_j indicates all samples in the leaf node$j$.

IV. RESULTS & DISCUSSIONS

The simulation tool is Jupiter notebook, which is suitable for python programming tasks. Jupiter notebook also includes rich text components and code, such as figures, equations, and links, among other things. These documents are ideal for bringing organized explanation and its discoveries, and executing data scrutiny in actuality, thanks to their blend of rich text components and code. Jupiter Notebook is a web-based collaborative visual, maps, charts, conceptions, and descriptive script tool that is open-source.

4.1 RESULT ANALYSIS-

The training sample is the set of order to train the program. 70% of the data was used for preparation within that study. In the computer vision community, 60 to 70% percent of data will be used for teaching on typical, but this varies based on the goal of the trial. In the k-fold classification model, the training set is separated into four Portions, and a training plus training set is formed from each 10-part. A theory is used, and the findings of all of the other training and test sets are combined. The histogram of attributes in Figure 2 displays the variety of dataset characteristics and the code that was used to build it.

The status of heart health is depicted in Figures 3 , which varies from excellent to terrible. The blue bar signifies the male population, while the red bar characterizes the female population. In this data collection, the male population appears to be more vulnerable to heart disease. in figure 3 flow chart of classification algorithm shows the different name of algorithms used in implementation and for comparison.

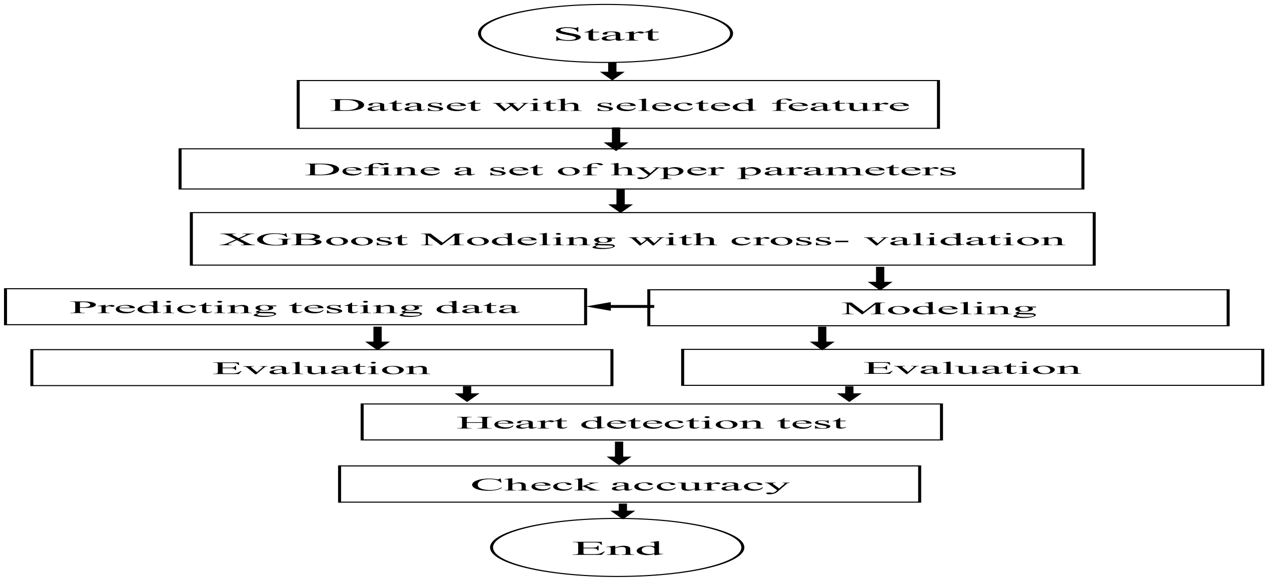

Start
Dataset with selected feature
Define a set of hyper parameters
XGBoost Modeling with cross- validation
Predicting testing data
Modeling
Evaluation
Evaluation
Heart detection test
Check accuracy
End

Figure 1: Flow Chart of Proposed Algorithm

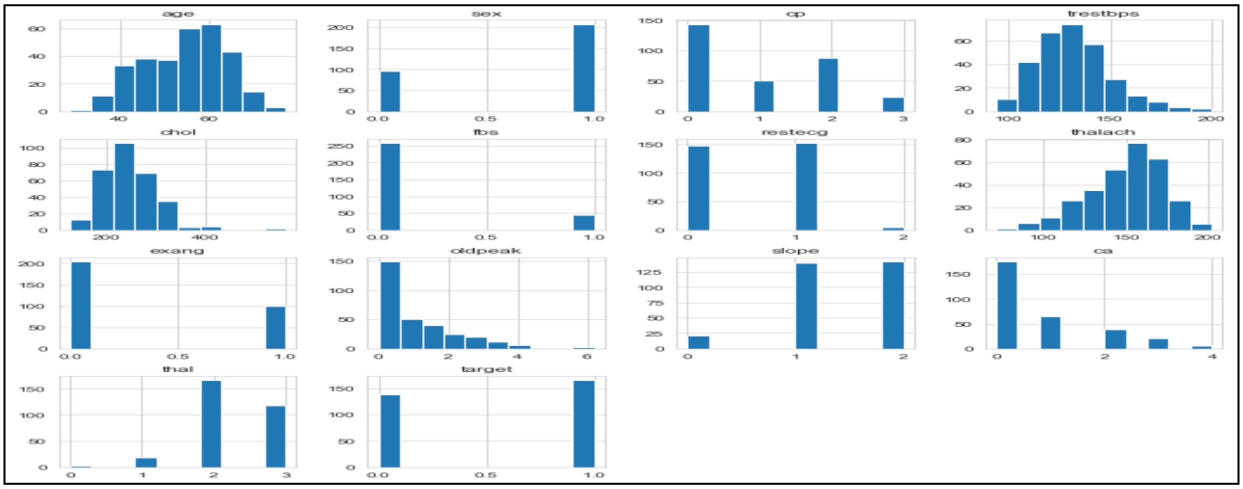


Figure 2: range of dataset attributes

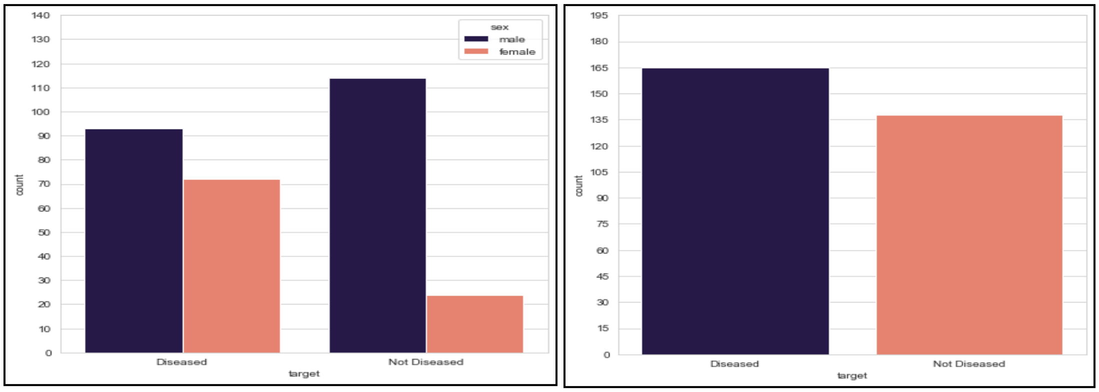


Figure 3 : Bar Plot of Number of Patients with Disease & According to Gender.

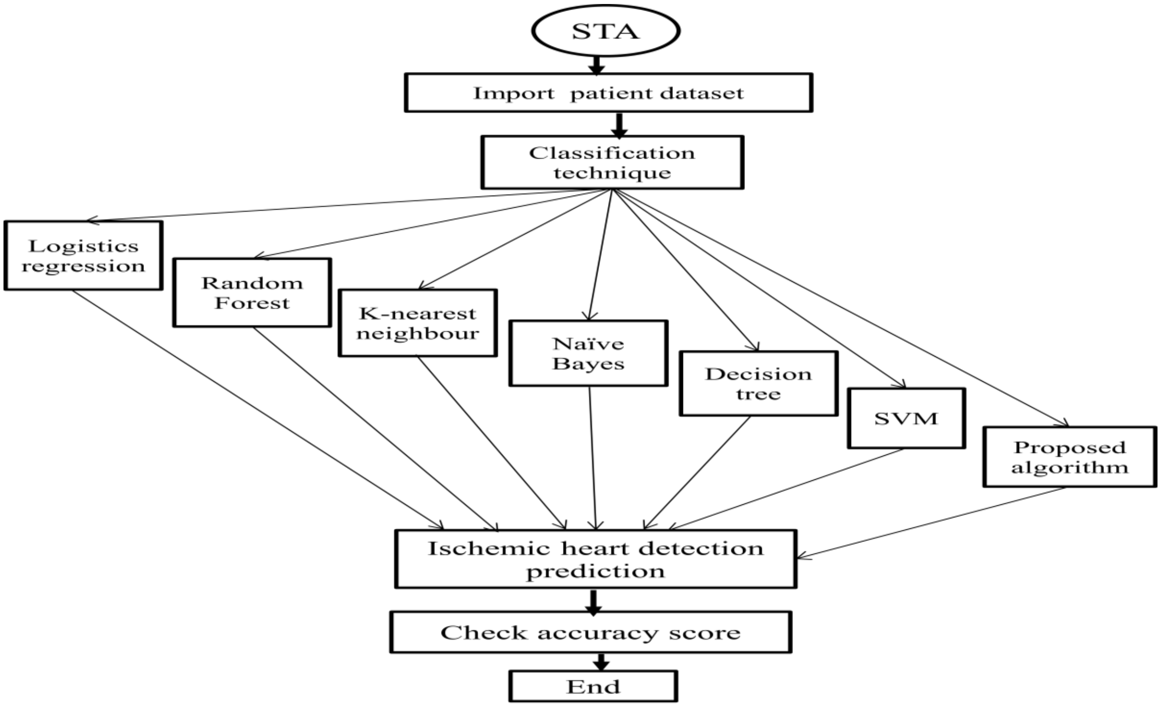
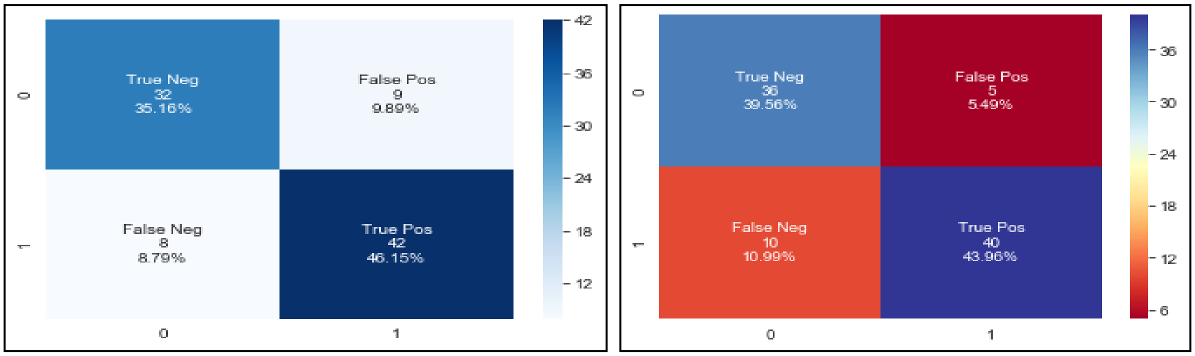


Figure 4: flow chart of all classification algorithms

4.1.1 LOGISTIC REGRESSION

- The Logistic Regression's confusion matrix is as follows:

To measure efficiency, the L2 premium, which may be the inverse of the factor size and is handled by Logistic Regression, was utilized. The train has an accuracy of 83.88 percent, while a test has an accuracy of 85.25-point margin



(a) (b)

Figure 5: Logistic Regression & Naïve Bayes Confusion Matrix

The values are determined using consistency, resolution, recalls, and the F1 score. It's time to revisit the descriptions of TP, FP, TN, and FN. A true positive (TP) is a false positive that its model projects properly; while a false positive (FP) is a positive result that it is model predictions mistakenly. A true negative (TN) is a valid conclusion that the model correctly predicts so that a false negative (FN) is a negative result that the theory wrongly predicts. While our dataset isn't broad enough, we didn't employ cross-validation.

There were two components to the dataset: test set. In the tables below, the outcomes of several methods are listed. and we'll go through each technique. Figure 5 shows a flow chart of all classification algorithms with all steps.

It functions very well so far, but it's not fully to comprehend. As a result, applying Logistic Regression is simple and effective. But on the other hand, logistic regression has the drawback of assuming linearity between the dataset's

Attributes as shown in figure 5 (a). In actual life, data is rarely separate, and my dataset is no exception. As a reason, we were unable to sustain an outstanding result.

4.1.2 NAÏVE BAYES

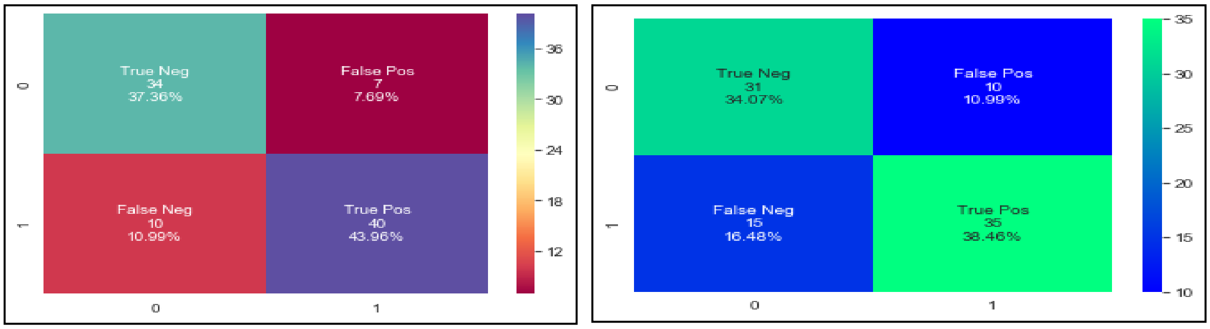
The Naive Bayes confusion matrix is as follows in figure 5 (b):

The train has an accuracy of 83.47 percent, while the test has an accuracy of 85.25 percent. Naive Bayes has the advantage of being able to make predictions with a small amount of training data. The disadvantage of Naïve Bayes is that it presupposes that all features are mutually independent.

However, we seldom meet a dataset with mutually independent characteristics, which may explain why we can't get high accuracy of 90%.

4.1.3 RANDOM FOREST

Random Forest's confusion matrix is as follows in figure 6(a):



(a) (b)

Figure 6: Random Forest & Decision Tree Confusion Matrix

The railroad is 100 percent accurate, and also the assessment is 91.80% accurate. We start also with the desktop environment (n estimators=100, suggesting that only the forest contains 100 trees, and max depth = Nothing here, insinuating also that nodes were often strained till all blossoms are clean or represent too little samples than in the level standard to divvy up an internal node). While we have 100% test accuracy, we only have 85.25 percent test accuracy. We believe it's due to an excessive amount of over fitting. We reordered the dataset and explored with either the random soviet from 1 to 2000 because one probable problem is that the training data never was specialized throughout the training phase. Again, when the random state equals 1826, the accuracy score is 91.80%. We then tried a few other values for n estimators (from 10 to 300) and maximum profundity (from 10 to 300), with the maximum accuracy score remaining at 91.80%. As random state=1825, this shows that other preset options are sufficient to attain the best test accuracy. Its woodland, for example, contains 100 healthy trees. Because the algorithm has not been updated for training data, let alone testing data, under fitting will occur with a small number of birds. When the number of trees in a model hits its maximum, the model has grown excessively intricate and sensitive to incoming data, resulting in over fitting. Random Forest has the benefit of being able to cope with datasets with a lot of features, balance the variance, and not be affected by data noise. Random Forest beats all other models among these five.(S.Dangare & S. Apte, 2012)

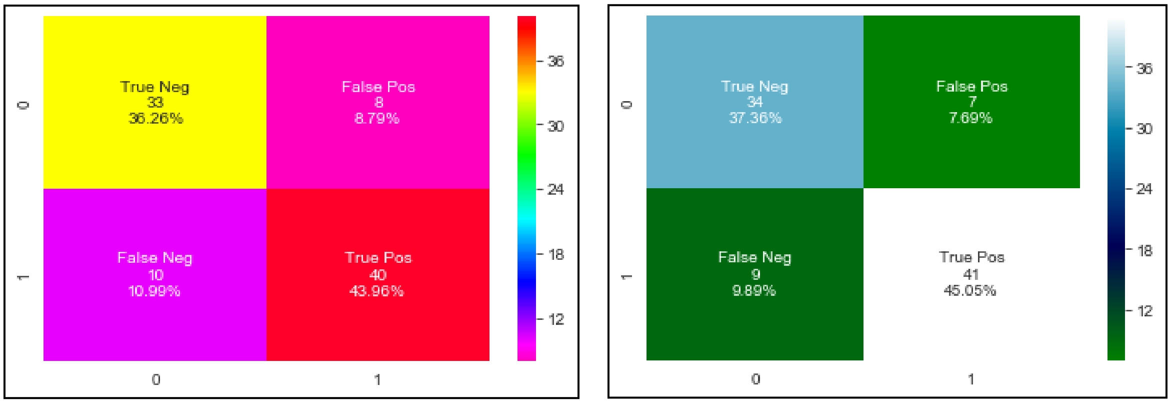
4.1.4 DECISION TREE

The Decision Tree's confusion matrix is as follows in figure 6(b):

The decision tree now has the least value of 77.55 percent, and that increases to 82.17 lakh hectares when combined with either the enhancing strategy. (Ho, 1995)shows that random forest performs badly, with 42.8954 % of instances correctly identified, but (Yasdi, 2000) utilizes so same database but builds Decision Trees using the J48 technique, resulting in an accuracy of 67.7%, which is lesser and yet greater than those of the previous. Author (Ranawana & Palade, 2005) had able to attain a 71.43 percent accuracy rate. Another author used overlapping decision trees in tandem with feature extraction to achieve 92.2 points per game (Vikas & Saurabh, 2013).

4.1.5 K-NEAREST NEIGHBOUR

The confusion matrix of KNN is shown in figure 7(a):



(a) (b)

Figure 7: Confusion Matrix for KNN (a) & SVM (b) .

KNN gives an accuracy of 83.16 percent when the value of k is set to 9 and the 10-cross validation technique is employed. KNN with Ant Colony Optimization beats other techniques with 70.26 percent accuracy and a 0.526 percent error rate. The author reported 87.5 percent efficiency .

4.1.6 SUPPORT VECTOR MACHINE

SVM's confusion matrix is as follows in figure 7(b):

For a tiny dataset, the learn tutorials say utilizing sklearn SVM. The accuracy obtained is 89.26 percent, while the test accuracy is 86.89 percent. In high-dimensional settings, SVM has the advantage of being very efficient. The main disadvantage is that the SVM has a huge number of parameters that must be carefully chosen to achieve the best results. We merely use SVM's default parameters to be safe. Furthermore, the test accuracy is greater than that of Logistic Regression at 86.89 percent.

4.2 PROPOSED ALGORITHM

In proposed method after execution, testing and training we discover that the accuracy of the proposed method is much competent as compared to another algorithm's.

We employed an array of SVM, KNN, and ANN in the prescribed group to attain a 95.61 percent share. Sensitivity of 92.3 percent, Precision 83.7 percent, and specificity of 93.8. Percent using a popular model that would include Naïve Bayes, Decision Tree, and Support Vector Machine classifiers. For the Proposed Algorithm, below is the confusion matrix is shown in figure 8:

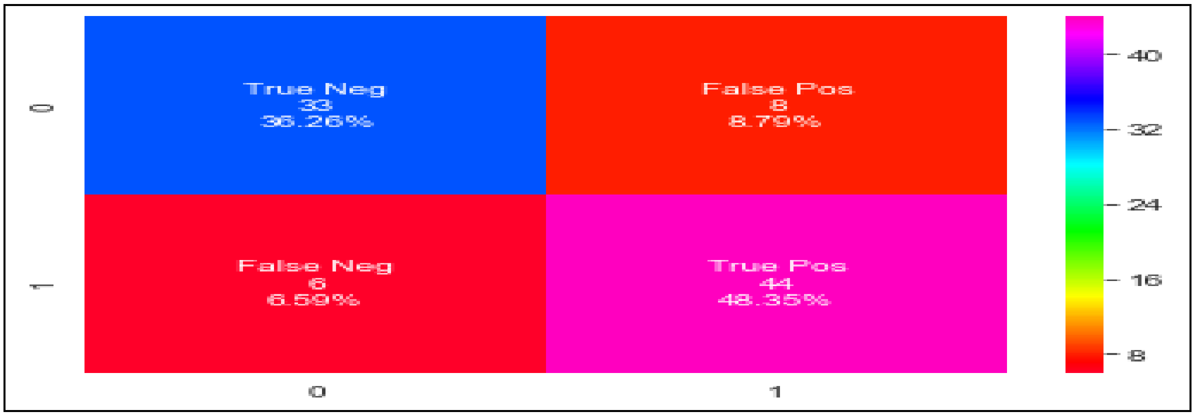


Figure 8: Matrix of Confusion for Proposed Algorithm.

Graph 1 shows the ROC curve for accuracy. We discovered that the accuracy of the random forest is significantly more efficient than other algorithms after using a machine learning system for train and test. The accurateness of each algorithm should be intended using the confusion matrix as shown in Figures, the number of counts of TP, TN, FP, and FN are supplied, and the value is computed using the accuracy equation (1). As shown in table 3 authors (Dwivedi, 2018) achieve maximum accuracy of 85% from the logical regression method. (Bhramaramba et al., 2011)Also achieved the same 85% accuracy in the same method. (Patel et al., 2021) Proposed a KNN model for the same dataset and achieved an accuracy of 90.1%. (Mohan et al., 2019) calculated 86.1% accuracy in random forest model and we calculated 87.8% accuracy for the same model. And graph 2 shows that the suggested method is the best among them, with an accuracy of 95.61 percent.

In table 3 it observed that from the proposed model we are getting sensitivity of 92.3 percent, Precision 83.7 percent, and specificity of 93.8. Percent. The same parameter is implemented by author (Mohan et al., 2019) and record maximum sensitivity 90.5 percent in naïve bias and maximum specificity 100 percent in SVM. Also record maximum precision 82.6 percent.

V Conclusion-

In conclusion, this research has proven that it is possible to classify heartbeats. Some characteristics of heartbeats and machine learning classification methods are used for the classification task. The suggested model has the potential to be implemented in clinical practice as a guide to aid cardiologists in reviewing ECG heartbeat signals and deeper understanding them. Other types of datasets, like stress and clinical datasets can stand to gain from the classification stage. We used datasets from the UCI machine learning directory to diagnose heart disease in this paper.

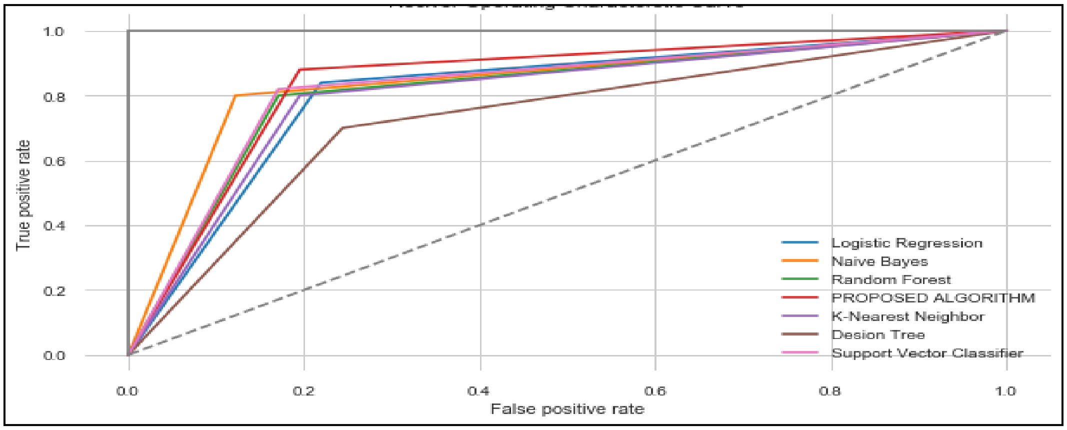
Classification methods, such as Naive Bayes, SVM, Decision Tree, Random Forest, and AdaBoost algorithms, are used to Train clinical data. Gradient Boost has the highest accuracy rate of 95.62 percent in the case of heart disease, including an experiment. Only 14 essential attributes are employed in this model. After pre-processing the data, it was observed that the Gradient boost delivered the finest outcomes in this model. Our objective is to achieve robust and convenient estimation with a relatively small quantity of features and tests. Other statistical methods, like clustering and association rubrics, and genetic algorithm, can also be used to broaden this research. In future, this research can be formulated more diverse and integrated frameworks to improve the precision of early prediction of heart disease.

TABLE 3: Comparison of Accuracy with previous algorithm

Sr. No.	Authors	Naive Bayes	SVM	KNN	Logistic Regression	Decision Tree	Random Forest	Proposed Algorithm
1.	Ashok et. al.(2016)	83%	82	80	85	-	-	-
2	Kumar et. al.(2018)	-	-	-	85%			
3.	Devansh et.al. (2020)	88.15%	-	90.78%	-	80.26%	86.84%	-

4.	S.Mohan et.al.(2019)	75.8%	86.1%	-	82.9%	85%	86.1%	-
5.	Proposed	83.51%	84.79%	80.21%	85%	77.5%	87.94%	95.6%

Graph 1: ROC curve of accuracy for all and proposed algorithm



Graph 2: Bar Plot for Accuracy of Prediction

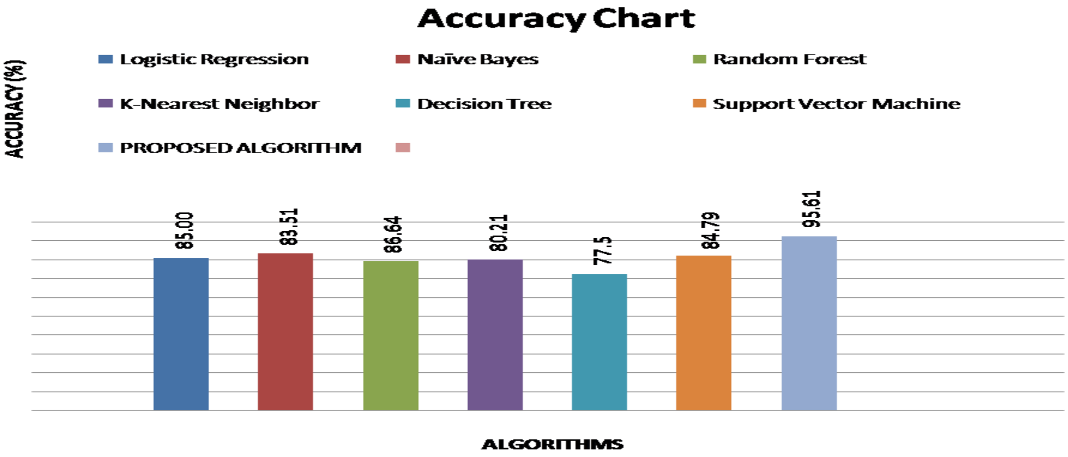


Table 4: Comparative analysis of Sensitivity, Specificity, and Precision in %

S.No.	Method	Sensitivity	Specificity	Precision
1	Naïve Bias(Mohan et al.,	90.5	79.8	60.0

	2019)			
2	Logical Regression(Mohan et al., 2019)	89.6	91.1	25.0
3	Decision Tree(Mohan et al., 2019)	86	98.8	0.0
4	Random Forest(Mohan et al., 2019)	87.1	98.8	10.0
5	SVM(Mohan et al., 2019)	86.1	100	0.0
6	HRFLM(Mohan et al., 2019)	90.1	92.8	82.6
7	Proposed Method	92.3	93.8	83.7

References

Akyildiz, I. F., Lee, W. Y., Vuran, M. C., & Mohanty, S. (2006). NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer Networks*, *50*(13), 2127–2159.

Alarsan, F. I., & Younes, M. (2019). Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0244-x

Azar, A. T. (2013). Fast neural network learning algorithms for medical applications. *Neural Computing and Applications*, *23*(3–4), 1019–1034. https://doi.org/10.1007/s00521-012-1026-y

Bahrami, B., & Shirvani, M. H. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, *2*(2), 3159–3199.

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications*, *19*(8), 1165–1195. https://doi.org/10.1007/s00521-010-0362-z

Bansal, A., Agarwal, R., & Sharma, R. K. (2015). Determining diabetes using iris recognition system. *International Journal of Diabetes in Developing Countries*, *35*(4), 432–438. https://doi.org/10.1007/s13410-015-0296-1

Bhramaramba, R., Allam, A. R., Kumar, V. V., & Sridhar, G. R. (2011). Application of data mining techniques on diabetes related proteins. *International Journal of Diabetes in Developing Countries*, *31*(1), 22–25. https://doi.org/10.1007/s13410-010-0001-3

Blake, R. (2007). Breaking the "Invisible-profession" paradigm. In *Journal of Environmental Health* (Vol. 70, Issue 3).

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(1), 262–267. https://doi.org/10.1073/pnas.97.1.262

C, K., & G.M, N. (2015). Classification and Prediction of Heart Disease from Diabetes Patients using Hybrid Particle Swarm Optimization and Library Support Vector Machine Algorithm. *International Journal of Computing Algorithm*, *4*(2), 54–58. https://doi.org/10.20894/ijcoa.101.004.002.001

Celesti, F., Celesti, A., Carnevale, L., Galletta, A., Campo, S., Romano, A., Bramanti, P., & Villari, M. (2017). Big data analytics in genomics: The point on Deep Learning solutions. *Proceedings - IEEE Symposium on Computers and Communications*, *Iscc*, 306–309. https://doi.org/10.1109/ISCC.2017.8024547

Deepika, K., & Seema, S. (2017). Predictive analytics to prevent and control chronic diseases. *Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, ICATccT 2016*, *November*, 381–386. https://doi.org/10.1109/ICATCCT.2016.7912028

Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, *29*(10), 685–693. https://doi.org/10.1007/s00521-016-2604-1

Farran, B., Channanath, A. M., Behbehani, K., & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait-a cohort study. *BMJ Open*, *3*(5), 1–10. https://doi.org/10.1136/bmjopen-2012-002457

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000).

Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*(10), 906–914. https://doi.org/10.1093/bioinformatics/16.10.906

Heydari, M., Teimouri, M., Heshmati, Z., & Alavinia, S. M. (2016). Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. *International Journal of Diabetes in Developing Countries*, *36*(2), 167–173. https://doi.org/10.1007/s13410-015-0374-4

Ho, T. K. (1995). Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Ho, T. K. (1998). *00709601.Pdf*. *20*(8), 832–844.

Hoptroff, R. G. (1993). The principles and practice of time series forecasting and business modelling using neural nets. *Neural Computing & Applications*, *1*(1), 59–66. https://doi.org/10.1007/BF01411375

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*(1). https://doi.org/10.1088/1757-899X/1022/1/012072

Kanmani, S., Uthariaraj, V. R., Sankaranarayanan, V., & Thambidurai, P. (2007). Object-oriented software fault prediction using neural networks. *Information and Software Technology*, *49*(5), 483–492. https://doi.org/10.1016/j.infsof.2006.07.005

Khan, S. I., Choubey, S. B., Choubey, A., Bhatt, A., Naishadhkumar, P. V., & Basha, M. M. (2021). Automated glaucoma detection from fundus images using wavelet-based denoising and machine learning. *Concurrent Engineering Research and Applications*. https://doi.org/10.1177/1063293X211026620

King, R. D. (n.d.). *Statlog Project Data Set*. 1–6.

Kumari, M., & Godara, S. (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. *International Journal of Computer Science Trends and Technology*, *2*(2), 304–308. http://www.ijcst.com/vol22/2/milan.pdf

Learning, P. M. L., & Kidwell, D. A. (n.d.). *Book Tracking Login Change Home*. 1–7.

Mariot, A., Sgoifo, S., & Sauli, M. (1964). I gozzi endotoracici: contributo casistico-clinico (20 casi). *Il Friuli Medico*, *19*(6).

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Patel, J., Khaked, A. A., Patel, J., & Patel, J. (2021). Heart Disease Prediction Using Machine Learning. *Lecture Notes in Networks and Systems*, *203 LNNS*, 653–665. https://doi.org/10.1007/978-981-16-0733-2_46

Policy, P. (2019). *We use cookies to personalise content and ads , to provide social media features and to analyse our traffic . We also share information about your use of our site with our social media , advertising and analytics partners in accordance with our Privacy St*. 1–9.

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. *Procedia Computer Science*, *85*, 962–969. https://doi.org/10.1016/j.procs.2016.05.288

Ranawana, R., & Palade, V. (2005). A neural network based multi-classifier system for gene identification in DNA sequences. *Neural Computing and Applications*, *14*(2), 122–131. https://doi.org/10.1007/s00521-004-0447-7

S.Dangare, C., & S. Apte, S. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *International Journal of Computer Applications*, *47*(10), 44–48. https://doi.org/10.5120/7228-0076

Sharma, V., Yadav, S., & Gupta, M. (2020). Heart Disease Prediction using Machine Learning Techniques. *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, *29*(3), 177–181. https://doi.org/10.1109/ICACCCN51052.2020.9362842

Vikas, C., & Saurabh, P. (2013). Data Mining Approach to Detect Heart Diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, *2*(4), 56–66.

Wu, Y., & Vapnik, V. N. (1999). Statistical Learning Theory. *Technometrics*, *41*(4), 377. https://doi.org/10.2307/1271368

Yasdi, R. (2000). A literature survey on applications of neural networks for human-computer interaction. *Neural Computing and Applications*, *9*(4), 245–258. https://doi.org/10.1007/s005210070002