Data & Knowledge Engineering 140 (2022) 102042

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset

Karthick Kanagarathinam [a,*], Durairaj Sankaran [b], R. Manikandan [c]

[a] *Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, Andhra Pradesh, India*
[b] *Department of Mechatronics Engineering, K.S. Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India*
[c] *Department of Electronics & Communication Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India*

ARTICLE INFO

Keywords:
Heart disease
Machine learning classifier
Feature selection
Hybrid heart disease dataset
CatBoost

ABSTRACT

CVD (cardiovascular disease) is one of the most common causes of death in the world today. CVD prediction allows health professionals to make an informed decision about their patients' health. Data mining is the process of transforming large amounts of medical data in its raw form into actionable insights that can be used to make intelligent forecasts and decisions. Machine learning (ML) based prediction models provide a better solution to help patients' health diagnoses in the health care industry. The objective of this research is to create a hybrid dataset to aid in the development of a best CVD risk prediction model. The Hungarian, the Switzerland, the Cleveland, and the Long Beach datasets are the most commonly used datasets in heart disease (HD) prediction. These datasets have a maximum of 303 instances with missing values in their features, and the presence of missing values reduces the accuracy of the prediction model. So, in this article, we created the "Sathvi" dataset by combining these datasets, and it has 531 instances with 12 attributes with no missing data. The Pearson's correlation method was used to eliminate redundant features during the feature selection process. The Naive Bayes (NB), XGBoost, k-nearest neighbour (k-NN), multilayer perceptron (MLP), support vector machine (SVM), and CatBoost ML classifiers have been applied for prediction. The CatBoost ML classifier was validated with 10-fold cross validation, and the best accuracy ranged from 88.67% to 98.11%, with a mean of 94.34%.

1. Introduction

Heart disease (HD) is a prevalent disease that afflicts many people in their middle or old age, and it frequently results in fatal complications. According to 2008 health survey, stroke accounted for about one in 18 deaths in the United States (US). In US, 6,55,000 people per year are died by HD. CVDs affect the cardiovascular system. Approximately one in every eighteen Americans died as a direct result of a stroke in 2008, according to government statistics. [1]. To manage CVD, lifestyle changes are necessary, or the healthcare provider may prescribe medications. The earlier CVD is detected, the easier it is to treat. The common symptoms of CVD are chest pain, an irregular heartbeat, nausea, etc. The most frequently identified possible CVD cause remained BMI. Having high cholesterol and high blood pressure were the second and third most common risk factors for CVD. According to the 2011 survey, men were 1.64 times more likely than women to have CVD [2]. Faced with a global viral pandemic like COVID-19 [3], We must emphasize international objectives to reduce the early mortality led by CVD, which limits healthy and sustainable development in all countries around the world. There is an abundance of research data and hospital patient records available. There are many

[*] Corresponding author.
E-mail addresses: karthick.k@gmrit.edu.in (K. Kanagarathinam), durairajeeebe@gmail.com (D. Sankaran), money_kandan2004@yahoo.co.in (R. Manikandan).

https://doi.org/10.1016/j.datak.2022.102042
Received 27 December 2021; Received in revised form 8 May 2022; Accepted 21 May 2022
Available online 1 June 2022

open resources available to access healthcare information, and research can be conducted to determine how various information and communication technologies can be utilized to predict/ diagnose HD before it turns fatal. ML-based techniques are becoming more common in business and society, and they are now being employed to healthcare [4]. ML is a scientific discipline that studies how machines acquire knowledge from data and develop themself. It is primarily based on statistics and probability [5]. However, when it comes to decision making process, it outperforms standard statistical methodologies. The information gathered from a dataset and fed into the algorithm is referred to as features. The quality of the features offered to the algorithm determines the model's prediction accuracy.

The job of the ML developer is to identify the subset of attributes that will best fit the objective, thereby boosting the model's accuracy. There are three basic steps to take in developing the ML prediction model, namely training, testing, and validation [6]. Training is essential because the prediction or classification model's accuracy is dependent on the training data. The algorithm's performance will be evaluated using the test dataset. The k-fold validation is required to determine the stability of the model [7]. The primary aim of this research is to build the best early-stage CVD prediction model based on the most optimal attributes. Among the sub-goals are a review of existing approaches for detecting CVD; creating a hybrid dataset with no missing values; determining the best features using the Pearson's r coefficient of correlation feature selection technique; building various prediction models on a "Sathvi" dataset using different ML algorithms; and evaluating the performance of the best ML algorithm using k-fold cross validation.

2. Related work

An algorithm's ability to learn from its own data and experience is known as ML. It is regarded as a component of artificial intelligence. It has a wide range of applications in the fields of electrical [8], health care [9], agriculture [10], meteorology [11], and so on. The HD risk prediction model was developed by Shah et al. [12] using 14 essential attributes. They used NB, decision trees, k-NN, and random forests for data mining classification. They discovered that the k-NN classifier has the highest accuracy of 90.789%.

Sequential minimal optimization (SMO) was used by Reddy et al. [13] to achieve an accuracy of 85.148% using the entire Cleveland heart dataset attributes and 86.468% using the best attribute set found from the Chi-square feature selection, respectively. The algorithms' performance was assessed using 10-fold Cross-validation. The Ensemble approach was used by Ibomoiye DM et al. [14] to achieve classification accuracy of 93% on the Cleveland dataset and 91% on the Framingham dataset. k-NN, SVM, and Logistic Regression are just a few of the ML classifier models that Ramya et al. [15] used to create an effective computational intelligent system (LR). In order to verify the model's accuracy, specificity, error rate, and sensitivity, performance metrics such as the Mathews Correlation Coefficient are employed. Nearly identical accuracy values (87 and 85%, respectively) were found by them between LR and SVM, while k-NN had 69%.

Heart rate variability (HRV) and pulse transit time variability (PTTV) were analysed concurrently in healthy and heart failure patients by Zhao et al. [16] with the goal of investigating the improvement of HRV-based heart failure detection with the assistance of PTTV analysis. A SVM classifier produced the best classification results with a sensitivity of 0.93, a specificity of 0.88, and an accuracy of 0.90 when HRV, PTTV, and predicted probabilities from distance distribution matrix-based convolutional neural network models were combined.

Patients with heart failure can benefit from a clinical decision support system developed by G. Guidi et al. [17], which includes an assessment of the severity of the condition, the prediction of the type of heart failure, as well as the ability to track the progress of individual patients over time. They aimed to simplify monitoring scenarios by automatically producing outputs about the severity and type of HF that could be read by non-cardiologist physicians and nurses. They compare various types of ML techniques to provide these outputs, ultimately deciding on the classification and regression tree (CART) method as the best fit for their requirements. Cross-validation accuracy in severity assessment is 81.8%, and type prediction accuracy is 87.6%, with CART providing a human-friendly decision-making process. Due to the small sample size, they were unable to extrapolate the results.

Classifying HD with six machine-learning techniques (Logistic Regression, SVM, kNN, artificial neural network, non-parametric Bayesian network, and random forest) was the goal of a study conducted by Tougui et al. [18]. There are 13 features, one target variable, and 303 instances in the Cleveland dataset used in this study, 139 of which are CVD patients and the other 164 of which are healthy. After analysing all the data, they found that Matlab's Artificial Neural Network model had the highest accuracy (85.86%) and sensitivity (83.94%) of among other tools or techniques.

Most studies have found that age and sex (gender) are the most significant factors in predicting cardiovascular disease risk. The existing datasets from the UCI ML database [19] Hungarian, Switzerland, Cleveland, and Long Beach have more missing values in each instance, which is undesirable when developing an ML-based prediction model. If a data feature has more than 50% missing data, it will be excluded from model development, even if it has a positive correlation with the target. As a result, the independent use of the datasets other than Cleveland may be ineffective for developing risk prediction models. As an outcome, a large dataset is required for developing the best HD prediction model. In this article, the created "Sathvi" dataset meets this requirement with 531 instances with 11 attributes. These 531 instances have been obtained by pre-processing the existing four datasets.

3. Materials and methods

The following stages are involved in the development of a CVD risk prediction model. It begins with the creation of the "Sathvi" dataset, followed by pre-processing the data, feature selection, application of ML classification algorithms, identification of the best ML algorithm, and k-fold cross validation of the selected model.

Table 1
Heart disease dataset description.

Dataset	No. of instances	No. of features	Missing values (in Rows-instances)
Hungarian	294	14	292
Switzerland	123	14	124
Cleveland	303	14	6
Long Beach	200	14	199
Hybrid dataset	920	14	621
Sathvi dataset	531	12	–

Table 2
Features description.

Attribute name	Attribute description
Age	Age in years
Sex	1 denotes a male and 0 denotes a female.
CP	Chest pain type 1 — typical angina, type 2 — atypical angina, type 3 — non-anginal pain and type 4 — asymptomatic
trestbps	Resting blood pressure (in mmHg at entry to the health centre)
chol	Serum lipid level in mg/dl
fbs	1 denotes true i.e., the fasting blood sugar level >120 mg/dl; 0 denotes false.
restecg	Resting ECG results: Null — Normal, 1 — ST-T wave abnormality and 2 — probable or definite left ventricular hypertrophy.
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes; Null = no)
oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment (1, 2 & 3) 1 – Upsloping, 2 – flat & 3 - Downsloping
ca	Number of major vessels (0–3) coloured by flourosopy
thal	Thalassemia: 3 = Normal, 6 = Fixed defect and 7 = Reversible defect
target	Null = No risk of CVD, 1, 2, 3 & 4 = Risk of CVD

Table 3
Missing values in % for each feature of the "Hybrid dataset".

Feature	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Missing data in %	6.41%	3.26%	9.78%	0.22%	5.98%	5.98%	6.74%	33.59%	66.41%	52.83%

3.1. Data

Hungarian, Switzerland, Cleveland, and Long Beach datasets in Table 1 are collected from the UCI Machine Learning repository [19]. Most of the researchers [20–23] used the Clevland dataset to develop the HD prediction model. UCI's database [19] contains 76 attributes, and the processed dataset has 14 features. Moreover, each instance's total number of missing values is shown in Table 1. The developed datasets "Hybrid" and "Sathvi" have 920 and 531 instances, respectively, and can be found as supplementary material with this article. The "Sathvi" dataset contains no missing values and was used to create the CVD risk prediction model. The feature descriptions are available in Table 2 [19].

3.1.1. "Sathvi" dataset

The "Hybrid" dataset combines all four datasets (Hungarian, Switzerland, Cleveland, and Long Beach Dataset) and consists of 920 instances, 14 attributes, and 621 missing features (each instance).

Table 3 shows the percentage of missing values for each attribute/feature. For the total of 920 instances of the hybrid dataset, both 'ca' and 'thal' have more than 50% missing values. These two features (columns) have been removed from the "Hybrid" dataset and formed the "Sathvi" dataset. In addition, any instances with missing values in the "Hybrid" dataset are discarded. Finally, the "Sathvi" dataset has 531 instances and 12 features with no missing values. The processed "hybrid" and "Sathvi" datasets are available in the supplementary file. Table 4 illustrates the statistical information of the "Sathvi" dataset.

3.2. Data pre-processing

Fig. 1 depicts the pre-processing stage of the "Sathvi" dataset. Fig. 1(a) depicts the target values ranging from 0 to 4. A value of zero indicates that the patient is not at risk of CVD, whereas a value of one or higher indicates that the patient has HD. The "Sathvi" dataset contains 39% of instances that are not at risk of HD, while the other target values 1, 2, 3, and 4 have a risk of CVD with

Table 4
Statistical information of the "Sathvi" dataset.

Features	Age	Sex	CP	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target
Count	531	531	531	531	531	531	531	531	531	531	531	531
Mean	54.844	0.761	3.352	133.407	216.855	0.160	0.744	138.463	0.497	1.218	1.765	1.130
Std	8.824	0.427	0.912	18.969	99.014	0.367	0.894	25.834	0.500	1.105	0.601	1.183
Min	29	0	1	0	0	0	0	60	0	−1	1	0
25%	49	1	3	120	197	0	0	120	0	0.1	1	0
50%	56	1	4	130	233	0	0	140	0	1	2	1
75%	61	1	4	142	273	0	2	159	1	2	2	2
Max	77	1	4	200	603	1	2	202	1	6.2	3	4

30.3%, 13.4%, 13.4%, and 3.95% of the total instances, respectively. The targets were divided into two categories to develop the risk prediction model: the presence of HD '1' and the absence of HD '0'. The target values 1–4 have been converted to '1'. Fig. 1(b) illustrates this. Fig. 1(c) depicts the distribution of age and target. The minimum and maximum age ranges are 29 to 77. HD is a common ailment in people over the age of 60, as well as those between the ages of 41 and 60. The number of occurrences is shown on the 'y' axis, and the affected age group is visualized on the top in terms of target values. Fig. 1(d) depicts the gender data, which contains 76.1% male data and 23.9% female data. Fig. 1(e) and (f) show the cholesterol and slope levels in relation to the target, respectively. The cholesterol level is the serum lipid level. In terms of target values, the slope of the peak exercise of 1 indicates an upsloping, 2 indicates a flat, and 3 indicates a down sloping. The pairs plot has been shown in Fig. 2 and shows the distribution of single variables as well as the relationships between two variables. It allows us to quickly investigate distributions and relationships in a dataset. The pair plot gives us a comprehensive first look at the data.

3.3. Feature selection

Due to the sparsity of the data, a massive number of features adds complexity to a model and may deteriorate its performance. Not all features are equally useful or significant for target prediction [24]. The presence of certain characteristics may be detrimental. Certain characteristics may be highly correlated with one another. Feature selection is the method of reducing the attributes to the required number in a model without sacrificing its performance significantly. For feature selection, the Pearson's correlation coefficient (r) is used [25]. It is a correlation coefficient that indicates the linear relationship between two variables. Its value is between −1 and +1, −1 indicates negative linear association, 0 indicates no linear relation, and 1 indicates strong correlation. Additionally, r is invariant when the location and scale of the two variables are changed independently, implying that for a linear function, the angle to the x-axis has no effect on r. Pearson's correlation coefficient (r) is expressed as in Eq. (1).

$$r_{pq} = \frac{n\sum p_i q_i - (\sum p_i * \sum q_i)}{\sqrt{n\sum {p_i}^2 - (\sum p_i)^2} * \sqrt{n\sum {q_i}^2 - (\sum q_i)^2}} \tag{1}$$

where, r_{xy} = Pearson's correlation coefficient between x and y
n = The number of total observations
p_i = p's value (for ith observation)
q_i = q's value (for ith observation)

Fig. 3 shows the heat map of Pearson's correlation coefficient. The threshold correlation value of 0.5 has been used to find significant attributes. It is found that the 'slope' attribute is not significant in developing the model. It is decided to discard the 'slope' attribute from the pre-processed "Sathvi" dataset. The other ten attributes, namely 'Age', 'Sex', 'CP', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', and 'oldpeak', were considered for prediction model development.

4. Proposed machine learning classifiers

The NB, XGBoost, k-NN, SVM, MLP, and CatBoost ML classifiers have been applied for prediction. It is described in this section.

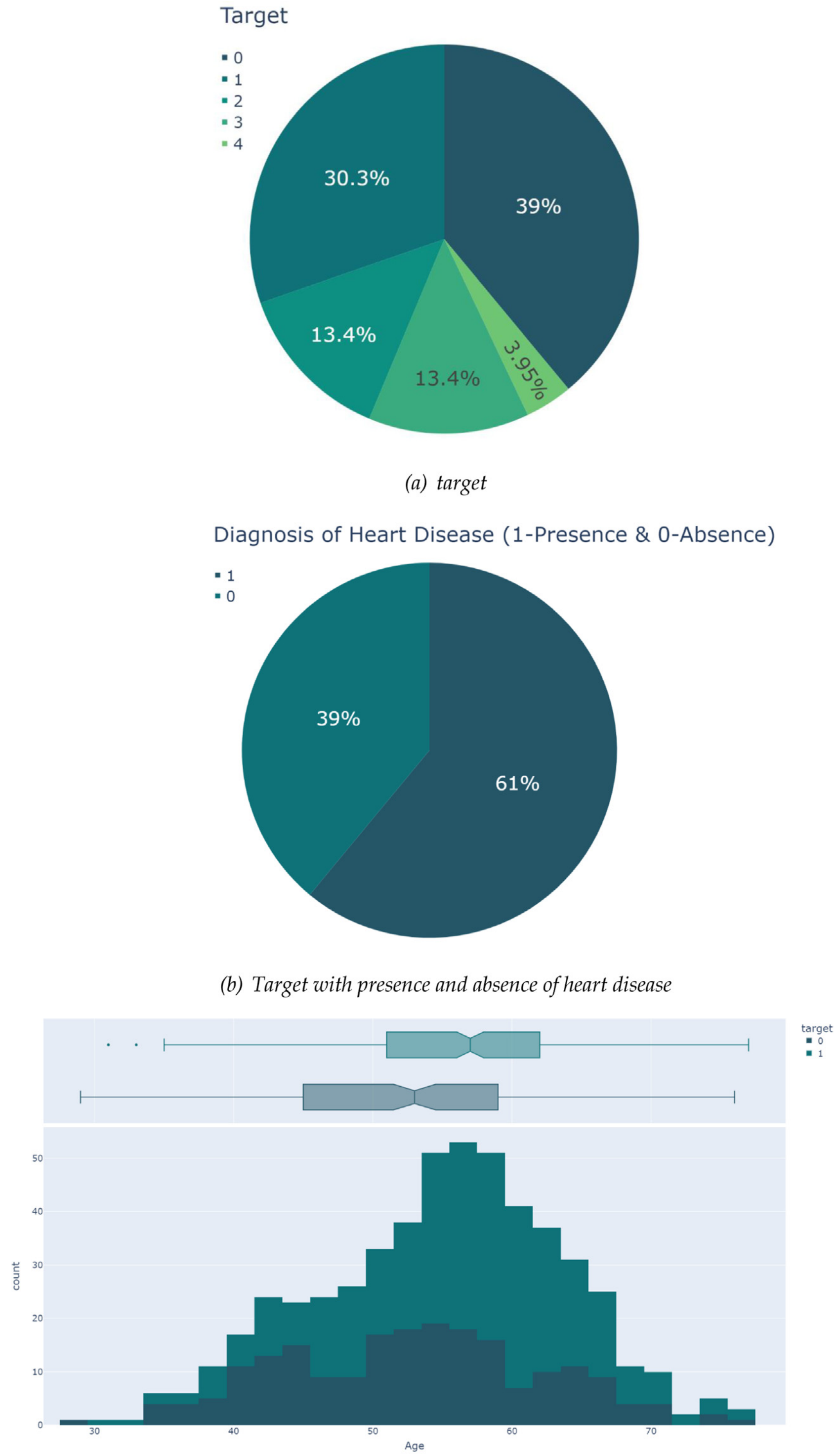
4.1. Naive Bayes classifier

Binary (two-class) and multi-class classification problems can be solved using Naive Bayes [26]. The method is most easily understood if the input values are binary or categorical. In comparison to more sophisticated methods, NB classifiers are extremely fast and, using Bayes' theorem, make predictions about new data. It is stated in Eq. (2).

$$\mathrm{P}\left(\mathrm{c}|\mathrm{d}\right) = \frac{(\mathrm{P(d|c)} * \mathrm{P(c)})}{\mathrm{P(d)}} \tag{2}$$

where

P(c| d) - Likelihood of hypothesis 'c' being true specified the data 'd'
P(d| c) - The likelihood that data 'd' is true if hypothesis 'c' is true.
P(c) - Likelihood where hypothesis 'a' is true
P(d) - Probability of the data

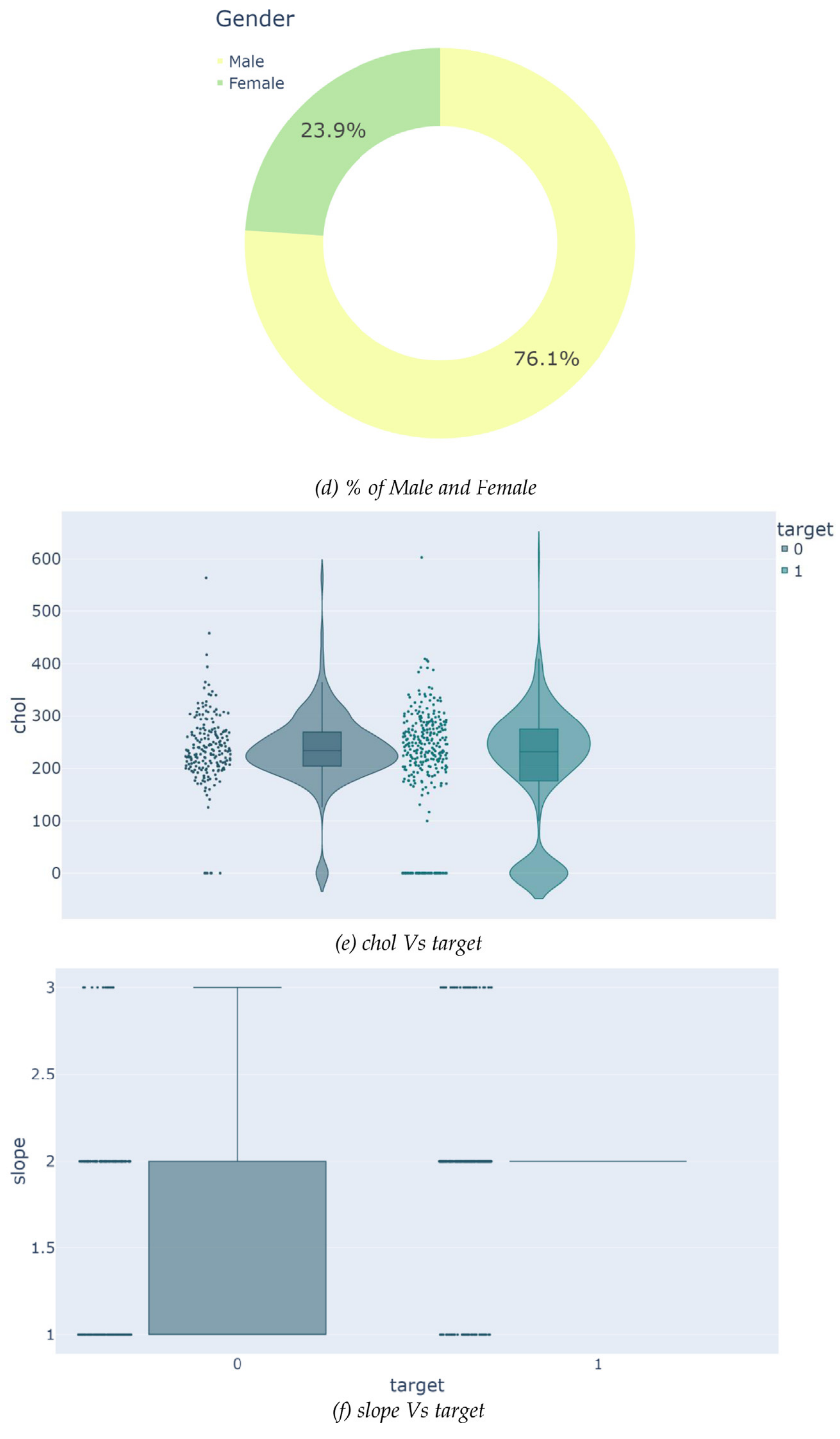


(a) target

(b) Target with presence and absence of heart disease

(c) Age Vs target

Fig. 1. Data visualization of “Sathvi” dataset.



(d) % of Male and Female

(e) chol Vs target

(f) slope Vs target

Fig. 1. (*continued*).

4.2. *Extreme gradient boost*

XGBoost [27] is a gradient boosting machine (GBM) implementation, a well-known algorithm for supervised learning. It is applicable to regression as well as classification problems.

If DS is the dataset that has $\boldsymbol{m}$ features, then for n instances,

$$\mathrm{DS} = \{\left(x_i, y_i\right) : i = 1 \ldots ..n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\} \tag{3}$$

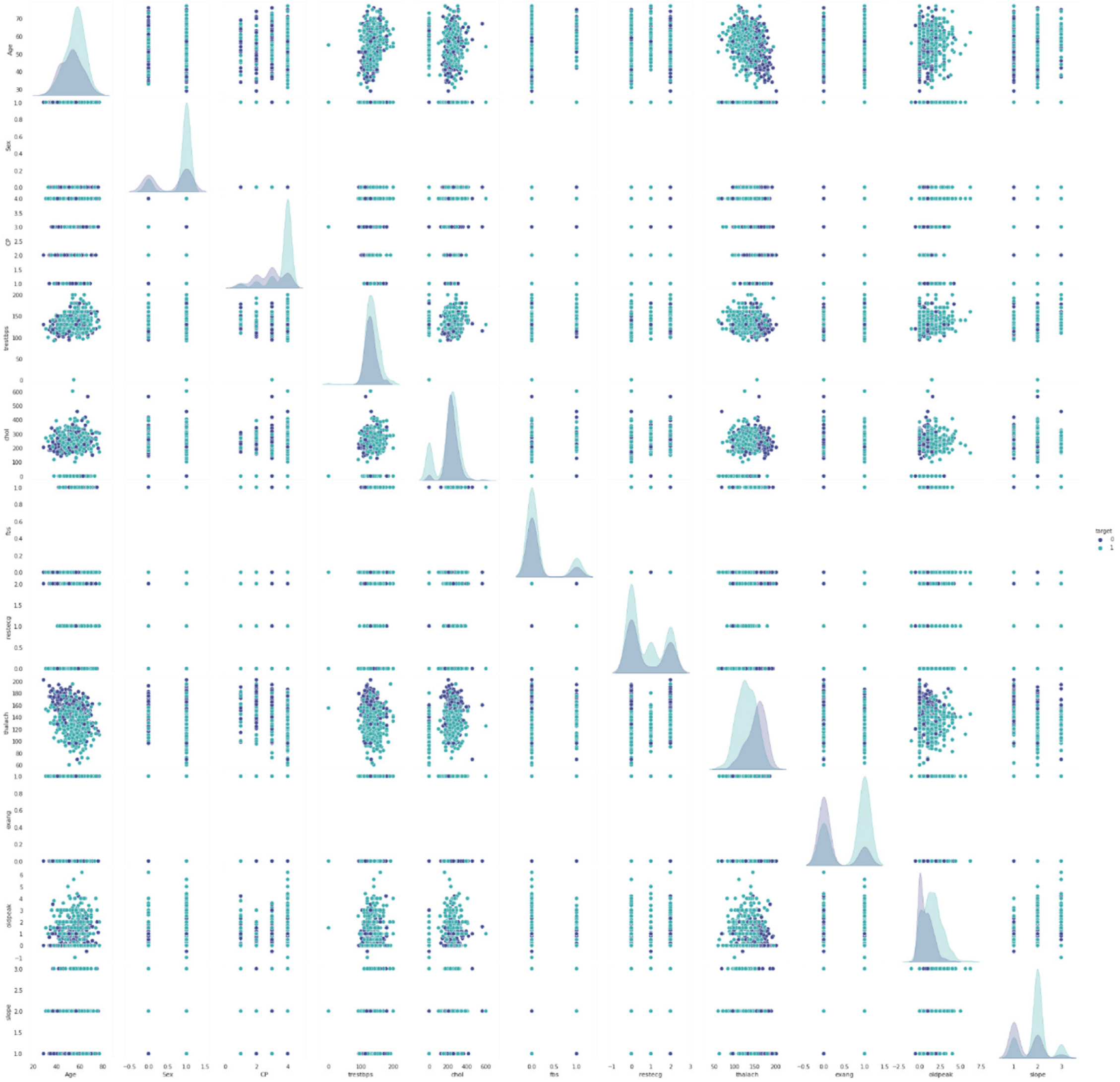


Fig. 2. Pair plot.

Let $\hat{y}_{\mathrm{i}}$ be the target value of the ensemble tree model, which is built using the Eq. (4).

$$\hat{y}\mathrm{i} = \phi\left(x_i\right) = \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in \mathcal{F} \tag{4}$$

Here K denotes the model's total number of trees and f_k denotes the model's kth tree.

4.3. K-NN classifier

The k-NN algorithm [28] makes the assumption that similar objects exist close together. That is, similar objects were located close together. The pseudo code of the k-NN classifier is shown in Fig. 4.

4.4. Support vector machine

RBF kernel-based SVM [29] classifiers can transform nonlinear problems into linear ones in multi-dimensional space by using an RBF kernel. RBF kernel in the SVM classifier is defined as follows:

$$K\left(y, y'\right) = e^{-\gamma\|y-y'\|^2} \tag{5}$$

where, $\|y-y'\|^2$ - The Euclidean squared distance between two feature vectors and γ is a scalar.

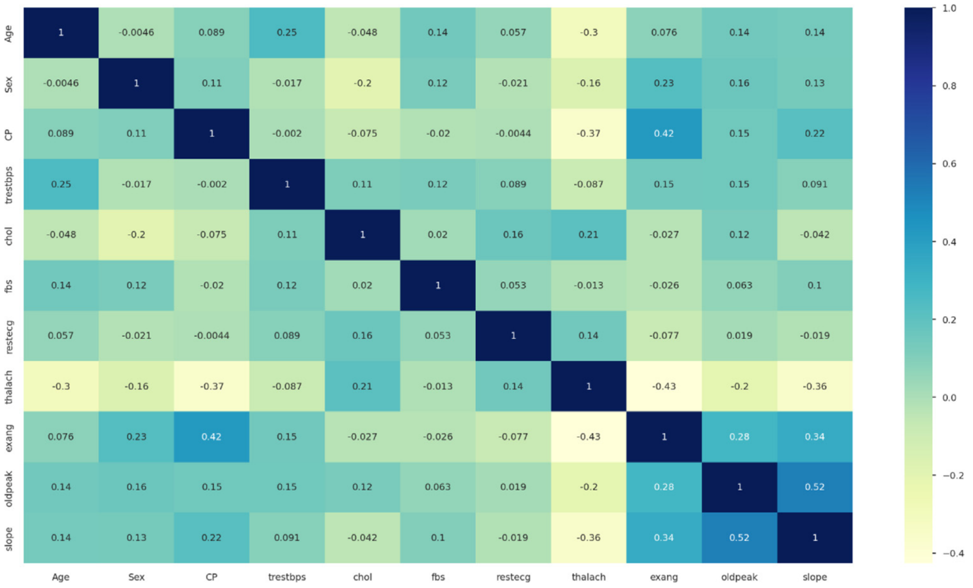


Fig. 3. Heat map of Pearson's correlation coefficient.

```
$k$-Nearest Neighbor
Classify $(\mathbf{X}, \mathbf{Y}, x)$ // $\mathbf{X}$: training data, $\mathbf{Y}$: class labels of $\mathbf{X}$, $x$: unknown sample
for $i = 1$ to $m$ do
  Compute distance $d(\mathbf{X}_i, x)$
end for
Compute set $I$ containing indices for the $k$ smallest distances $d(\mathbf{X}_i, x)$.
return majority label for $\{\mathbf{Y}_i$ where $i \in I\}$
```

Fig. 4. Pseudo code of k-NN classifier.

```
Import MLP Classifier
Hidden_layer_sizes = (64, 32), Activation = 'logistic'
Maximum Iterations=2000
Fit the classification model with (X_train, y_train)
Prediction with X_test
Plot the Confusion Matrix - mlp_conf_matrix
Estimate the Accuracy- accuracy_score(y_test, mlp_predict)
```

Fig. 5. Pseudo code of MLP.

4.5. Multilayer perceptron

A MLP [30] is an artificial neural network model that maps input data sets to appropriate output data sets by using a feed-forward mechanism. An MLP is made up of many layers of nodes in a directed graph, and each layer is fully connected to the next. The neurons (or processing elements) in each node (with the exception of the input nodes) have a nonlinear activation function, with the exception of the input nodes. The MLP Pseudo code is shown below in Fig. 5.

4.6. CatBoost

CatBoost is a gradient boosting algorithm [31] for decision trees. Yandex's open-sourced ML algorithm. For more descriptive data formats, it provides robust out of the box support Category and boosting are the roots of the name "CatBoost". The name "Boost" is derived from the gradient boosting ML algorithm because this library is built on the gradient boosting library. CatBoost can convert categories into numbers without any explicit pre-processing. Hyper-parameter tuning is simplified, and the risk of overfitting is

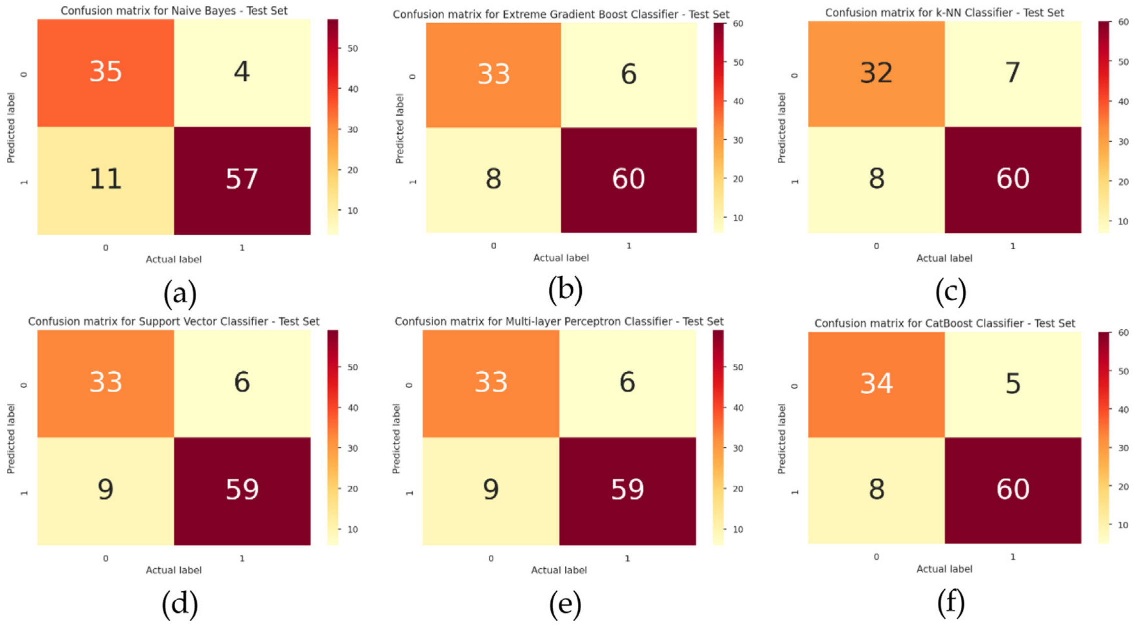


Fig. 6. Confusion matrix of (a) NB (b) XGBoost (c) k-NN (d) SVM (e) MLP (f) CatBoost.

decreased, resulting in more generalized models as a result of this approach. For developing the model, the number of iterations has been set to 1000 and the random strength is set to 0.1.

5. Results and discussion

5.1. Training and test dataset

The modelling step infers a representative model from the data. Training datasets are collections of data used to construct models, and they contain known features as well as target. Validation of the created model will also require comparison to another well-known dataset referred to as the test dataset or validation dataset. To facilitate this process, it is feasible to partition the entire known dataset into a training and a test set [32]. The "Sathvi" dataset has an 80:20 split between training and testing sets, with 424 and 107 instances each having 10 features, with attribute 11 serving as the "target" for the model. Among 424, the target '0' has 168 and the target '1' has 256 instances. This 80:20 dataset was used to assess all six classifiers. And the best performing model was subjected to k-fold cross validation.

5.2. Accuracy of the model

The performance metric accuracy which is defined in Eq. (6) is to calculate the ratio of precisely predictable samples to the total sample count. If the model used is highly accurate, it can be considered the best model.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

Fig. 6 shows the confusion matrix of all six models. The left top of the matrix represents true positive, the right top of the matrix represents the number of false positives, the left bottom of the matrix represents false negative, and the right bottom of the matrix represents true negative. The accuracy of each model is computed using Eq. (6). Fig. 7 shows the receiver operating characteristic (ROC) curve of all the six models. The ROC curve shows us how well the ML classifier is performing. Table 5 shows the CVD risk prediction accuracy of NB, XGBoost, k-NN, SVM, MLP, and CatBoost ML classifiers. It is observed that CatBoost provides the best accuracy compared to all other models with an 80:20 data split of the "Sathvi" dataset.

5.3. k-fold cross validation

Dataset is randomly divided into 'k' mutually exclusive subgroups or "folds" from the original dataset as F_1, F_2, ...F_k, each being about the same in size, in cross-validation with k folds. There are k iterations of training and testing. The CatBoost classifier has undergone a 10-fold cross validation. In iteration 'i' the test set is partition F_i, and the rest of the segments or subgroups or folds are utilized to train the model together. That is, for the first iteration of Fold 1, Fold 1 will be the test set and the other Folds

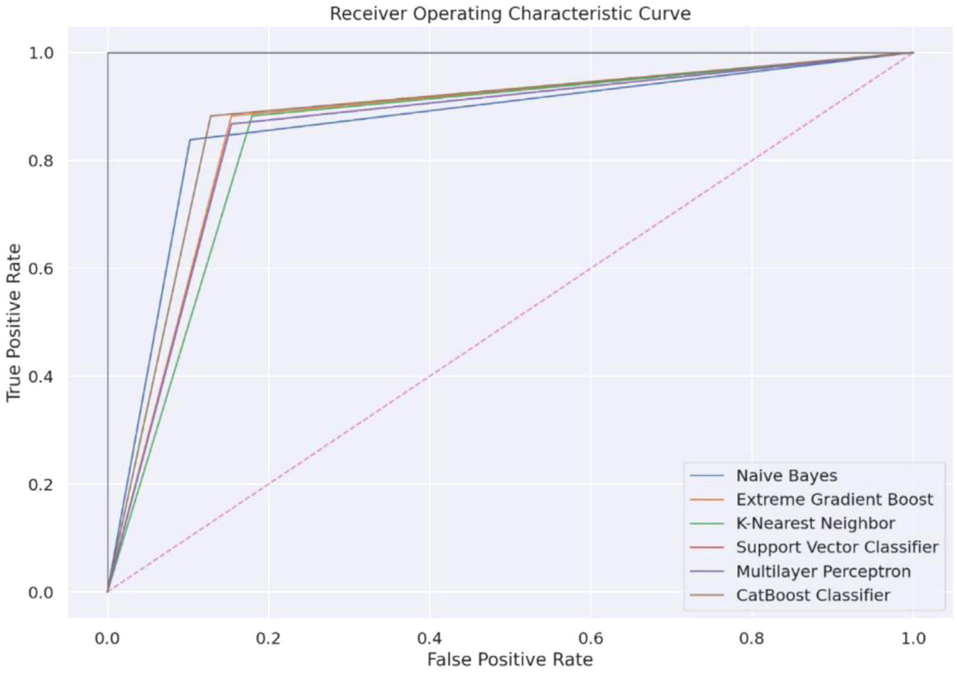


Fig. 7. ROC curve of ML classifiers.

Table 5
Proposed ML CVD risk prediction model accuracy.

Model	Accuracy (%)
Naive Bayes	85.98
XGBoost	86.91
k-NN	85.98
SVM	85.98
MLP	85.98
CatBoost	87.85

Table 6
10-fold cross validation accuracy of CatBoost ML algorithm.

Fold number	Training set	Test set	Accuracy (%)
Fold 1	477	54	92.59
Fold 2	478	53	98.11
Fold 3	478	53	96.22
Fold 4	478	53	96.22
Fold 5	478	53	94.33
Fold 6	478	53	88.67
Fold 7	478	53	94.33
Fold 8	478	53	94.33
Fold 9	478	53	96.22
Fold 10	478	53	92.45
Mean			**94.34%**

2 to 10 will act as the training set. Table 6 and Fig. 8 show the 10-fold cross validation accuracy of the CatBoost ML algorithm. It is observed that the CatBoost ML classifier obtains the best accuracy range of 88.67%–98.11% with a mean of 94.34%. Fold 1 has 477 and 54 instances for training and test set, while the other 9 folds have 478 and 53 instances for training and test set, respectively. Table 6 and Fig. 8 show the 10-fold cross validation accuracy of the CatBoost ML algorithm. Table 7 provides the comparison of accuracy with existing ML CVD risk prediction models. It is found that the developed model outperforms with a mean accuracy of 94.34%.

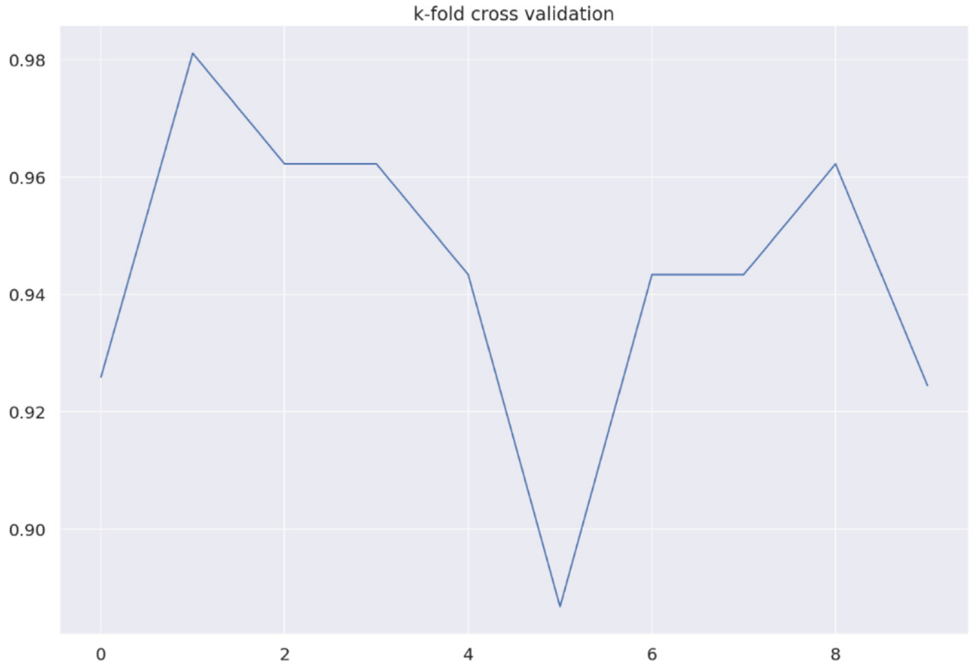


Fig. 8. The CatBoost ML algorithm's 10-fold cross validation accuracy.

Table 7
Comparison of accuracy with the existing ML CVD risk prediction model.

Author	ML Algorithm	Accuracy %
Shah et al. [12]	k-NN classifier	90.789
Reddy et al. [13]	SMO+ Chi-squared	86.468
Ibomoiye DM et. al., [14]	Ensemble Approach (Clevland dataset)	93
	Ensemble Approach (Framingham dataset)	91
Ramya et al. [15]	Logistic regression	87
L. Zhao et al. [16]	SVM	90
Proposed method	***CatBoost***	**94.34**

6. Conclusion

In this research, the "Sathvi" dataset has been created using the existing four CVD datasets with 531 instances. It does not have any missing values. The "hybrid" and "Sathvi" datasets are available as supplementary files for public use. The risk prediction model was developed with six ML classifiers and identified that the CatBoost ML classifier performs better with a mean accuracy of 94.34% by performing 10-fold cross validation. The risk prediction model was developed with 10 attributes. Pearson's correlation coefficient (r) has been employed to select the optimal features. The CatBoost ML classifier has the merit of lowering the chances of overfitting. When compared to existing models, the proposed model is significantly more accurate and has a higher number of instances. The developed model will help health practitioners make timely decisions on CVD prediction. Further, diagnostic assistance using ML can be developed in the future.

CRediT authorship contribution statement

Karthick Kanagarathinam: Conceptualization, Data curation, Writing – original draft, Investigation, Methodology. **Durairaj Sankaran:** Supervision, Validation, Writing – review & editing. **R. Manikandan:** Formal analysis, Software, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset is available as supplementary file with this article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.datak.2022.102042.

References

[1] . Roger, L. Véronique, et al., Heart disease and stroke statistics–2012 update: a report from the American Heart Association, Circulation 125 (1) (2012) e2–e220, http://dx.doi.org/10.1161/CIR.0b013e31823ac046.

[2] DMT. Tran, N. Lekhak, K. Gutierrez, S. Moonie, Risk factors associated with cardiovascular disease among adult Nevadans, PLoS One 16 (2) (2021) e0247105, http://dx.doi.org/10.1371/journal.pone.0247105.

[3] Karthick Kanagarathinam, Ebrahem A. Algehyne, Kavaskar Sekar, Analysis of 'earlyR' epidemic model and time series model for prediction of COVID-19 registered cases, Mater. Today: Proc. (ISSN: 2214-7853) (2020) http://dx.doi.org/10.1016/j.matpr.2020.10.086.

[4] Thomas Davenport, Ravi Kalakota, The potential for artificial intelligence in healthcare, Future Healthc. J. 6 (2) (2019) 94–98, http://dx.doi.org/10.7861/futurehosp.6-2-94.

[5] S. Makridakis, E. Spiliotis, V. Assimakopoulos, Statistical and machine learning forecasting methods: Concerns and ways forward, PLoS One 13 (3) (2018) e0194889, http://dx.doi.org/10.1371/journal.pone.0194889.

[6] A. Vabalas, E. Gowen, E. Poliakoff, AJ. Casson, Machine learning algorithm validation with a limited sample size, PLoS One 14 (11) (2019) e0224365, http://dx.doi.org/10.1371/journal.pone.0224365.

[7] K. Pal, B.V. Patel, Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques, in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 83–87, http://dx.doi.org/10.1109/ICCMC48092.2020.ICCMC-00016.

[8] K. Sekar, K. Karthick, Power quality disturbance detection using machine learning algorithm, in: 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), 2020, pp. 1–5, http://dx.doi.org/10.1109/ICADEE51157.2020.9368939.

[9] Fuzhe Ma, Tao Sun, Lingyun Liu, Hongyu Jing, Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network, Future Gener. Comput. Syst. 111 (2020) 17–26, http://dx.doi.org/10.1016/j.future.2020.04.036ISSN0167-739X.

[10] A. Sharma, A. Jain, P. Gupta, V. Chowdary, Machine learning applications for precision agriculture: A comprehensive review, IEEE Access 9 (2021) 4843–4873, http://dx.doi.org/10.1109/ACCESS.2020.3048415.

[11] Diego Gómez, Pablo Salvador, Julia Sanz, Jorge Gil, Juan Fernando Rodrigo, José Luis Casanova, Machine learning approach to predict leaf colour change in fagus sylvatica L. (Spain), Agricult. Forest Meteorol. 310 (2021) 108661, http://dx.doi.org/10.1016/j.agrformet.2021.108661ISSN0168-1923.

[12] D. Shah, S. Patel, S.K. Bharti, Heart disease prediction using machine learning techniques, SN Comput. Sci. 1 (2020) 345, http://dx.doi.org/10.1007/s42979-020-00365-y.

[13] KVV. Reddy, I. Elamvazuthi, AA. Aziz, S. Paramasivam, HN. Chua, S. Pranavan, Heart disease risk prediction using machine learning classifiers with attribute evaluators, Appl. Sci. 11 (18) (2021) 8352, http://dx.doi.org/10.3390/app11188352.

[14] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang, An improved ensemble learning approach for the prediction of heart disease risk, Inform. Med. Unlocked 20 (2020) 100402, http://dx.doi.org/10.1016/j.imu.2020.100402ISSN2352-9148.

[15] Ramya Perumal, AC. Kaladevi, Early prediction of coronary heart disease from cleveland dataset using machine learning techniques, Int. J. Adv. Sci. Technol. 29 (06) (2020) 4225–4234, http://sersc.org/journals/index.php/IJAST/article/view/16428.

[16] L. Zhao, C. Liu, S. Wei, C. Liu, J. Li, Enhancing detection accuracy for clinical heart failure utilizing pulse transit time variability and machine learning, IEEE Access 7 (2019) 17716–17724, http://dx.doi.org/10.1109/ACCESS.2019.2895230.

[17] G. Guidi, M.C. Pettenati, P. Melillo, E. Iadanza, A machine learning system to improve heart failure patient assistance, IEEE J. Biomed. Health Inf. 18 (6) (2014) 1750–1756, http://dx.doi.org/10.1109/JBHI.2014.2337752.

[18] I. Tougui, A. Jilbab, J.El. Mhamdi, Heart disease classification using data mining tools and machine learning techniques, Health Technol. 10 (2020) 1137–1144, http://dx.doi.org/10.1007/s12553-020-00438-1.

[19] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, M.D. James Beckerman, Heart Disease, UCI Machine Learning Repository, 1988.

[20] M.M.A. Rahhal, Y. Bazi, H. Alhichri, N. Alajlan, F. Melgani, R.R. Yager, Deep learning approach for active classification of electrocardiogram signals, Inform. Sci. 345 (2016) 340–354.

[21] D. Wettschereck, T.G. Dietterich, An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms, Mach. Learn. 19 (1) (1995) 5–27.

[22] S. Nalluri, R.V. Saraswathi, S. Ramasubbareddy, K. Govinda, E. Swetha, Chronic heart disease prediction using data mining techniques, advances in intelligent systems and computing, in: Data Engineering and Communication Technology, Springer, Singapore, 2020, pp. 903–912.

[23] I.M. Pires, G. Marques, N.M. Garcia, V. Ponciano, Machine learning for the evaluation of the presence of heart disease, Procedia Comput. Sci. 177 (2020) 432–437.

[24] R. Spencer, F. Thabtah, N. Abdelhamid, M. Thompson, Exploring feature selection and classification methods for predicting heart disease, Digit. Health (2020) http://dx.doi.org/10.1177/2055207620914777.

[25] B. Frey, The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation (Vols. 1-4), SAGE Publications, Inc, Thousand Oaks, CA, 2018, http://dx.doi.org/10.4135/9781506326139.

[26] S. Xu, Bayesian Naïve Bayes classifiers to text classification, J. Inf. Sci. 44 (1) (2018) 48–59, http://dx.doi.org/10.1177/0165551516677946.

[27] Ahmedbahaaaldin Ibrahem Ahmed Osman, Ali Najah Ahmed, Ming Fai Chow, Yuk Feng Huang, Ahmed El-Shafie, Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia, Ain Shams Eng. J. 12 (2) (2021) 1545–1556, http://dx.doi.org/10.1016/j.asej.2020.11.011.

[28] Bunheang Tay, Jung Keun Hyun, Sejong Oh, A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images, Comput. Math. Methods Med. (2014) 276589, http://dx.doi.org/10.1155/2014/276589, 8 pages, 2014.

[29] Y. Zhang, Support vector machine classification algorithm and its application, in: C. Liu, L. Wang, A. Yang (Eds.), Information Computing and Applications. ICICA 2012, in: Communications in Computer and Information Science, vol. 308, Springer, Berlin, Heidelberg, 2012, http://dx.doi.org/10.1007/978-3-642-34041-3_27.

[30] Soo See Chai, Whye Lian Cheah, Kok Luong Goh, Yee Hui Robin Chang, Kwan Yong Sim, Kim On Chin, A multilayer perceptron neural network model to classify hypertension in adolescents using anthropometric measurements: A cross-sectional study in Sarawak, Malaysia, Comput. Math. Methods Med. (2021) 2794888, http://dx.doi.org/10.1155/2021/2794888, 11 pages, 2021.

[31] Guomin Huang, Lifeng Wu, Xin Ma, Weiqiang Zhang, Junliang Fan, Xiang Yu, Wenzhi Zeng, Hanmi Zhou, Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions, J. Hydrol. 574 (2019) 1029–1041, http://dx.doi.org/10.1016/j.jhydrol.2019.04.085ISSN0022-1694.

[32] V. Roshan Joseph, Akhil Vakayil, SPlit: An optimal method for data splitting, Technometrics (2021) http://dx.doi.org/10.1080/00401706.2021.1921037.

Dr. K. Karthick is working as an Associate Professor in Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, India. He received his B.E. degree in Electrical and Electronics Engineering from Periyar University, Salem, India and a M.E. degree in Power Electronics and Drives from Anna University, Chennai, India. He completed his Doctorate in Electrical Engineering from Anna University, Chennai. He has more than 16 years of experience in teaching. He is the member of ISTE. His research interests include data analytics, machine learning, text detection and recognition, image processing, and Electrical Drives.

Dr. S. Durairaj is working as an Assistant Professor in the Department of Mechatronics Engineering at K S Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India. He completed his Doctorate in May 2017 at Anna University, Chennai, India. He completed his M.E. degree in Power Electronics and Drives in 2009. He has more than 12 years of experience in teaching. His research interest includes Green Energy, Power Electronics and drives, machine learning, etc.

Dr. R. Manikandan received his B.E degree in Electronics and Instrumentation Engineering from Annamalai University, Chidambaram in 2002. He obtained his M.E degree in Applied Electronics from Anna University, Chennai in 2008 and his Ph.D. degree in Image/Video Processing from the Department of Advanced Sports Training and Technology at Tamil Nadu Physical Education and Sports University, Chennai in 2014. His main research interests include automation, computer vision and image/video processing. He is now a Professor in the Department of Electronics and Communication Engineering at Panimalar Engineering College, Chennai.