BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 10/10 pp513–525
Volume 6, Number 4, December 2023
DOI: 10.26599/BDMA.2022.9020052

A Clinical Data Analysis Based Diagnostic Systems for Heart Disease Prediction Using Ensemble Method

Ankit Kumar, Kamred Udham Singh*, and Manish Kumar

Abstract: The correct diagnosis of heart disease can save lives, while the incorrect diagnosis can be lethal. The UCI machine learning heart disease dataset compares the results and analyses of various machine learning approaches, including deep learning. We used a dataset with 13 primary characteristics to carry out the research. Support vector machine and logistic regression algorithms are used to process the datasets, and the latter displays the highest accuracy in predicting coronary disease. Python programming is used to process the datasets. Multiple research initiatives have used machine learning to speed up the healthcare sector. We also used conventional machine learning approaches in our investigation to uncover the links between the numerous features available in the dataset and then used them effectively in anticipation of heart infection risks. Using the accuracy and confusion matrix has resulted in some favorable outcomes. To get the best results, the dataset contains certain unnecessary features that are dealt with using isolation logistic regression and Support Vector Machine (SVM) classification.

Key words: artificial intelligence; support vector machine; logistic regression; cleveland dataset; supervised algorithm; human sensing

1 Introduction

1.1 Importance of heart and heart disease status worldwide

Heart is a vital organ of the human body. It siphons blood to all aspects of our life structures. If it fails to work effectively, the mind and different organs will quit working, and within a couple of moments, the individual will pass on. Changes in the way of life, work-related pressure, and wrong food propensities add to the increment in the pace of a few hearts related diseases. Heart infections are quite possibly the most apparent reason for death from one side of the planet to the other.

Early detection of cardiovascular disease signs is one of doctors' most challenging issues today. Each year, cardiovascular disease kills a large number of people throughout the world. Because of the gravity of the problem, the cardiovascular disease needs prompt attention. Heart disease is typically difficult to detect because of the wide range of potential contributing factors, including but not limited to hypertension, hyperlipidemia, arrhythmia, and other health issues. This means Artificial Intelligence (AI) has the potential to aid in the early diagnosis and management of health problems.

According to the World Health Organization, heart-related infections claim 17.7 million lives yearly, accounting for 31% of all deaths worldwide. Heart disease is now the leading cause of death in India. According to the 2016 Global Burden of Disease report, heart disease killed 1.7 million Indians in 2016. Heart-

- Ankit Kumar is with the Department of Computer Engineering and Applications, GLA University, Mathura 281406, India. E-mail: iiita.ankit@gmail.com.
- Kamred Udham Singh is with the School of Computing, Graphic Era Hill University, Dehradun 248002, India. E-mail: Kamredudhamsingh@gmail.com.
- Manish Kumar is with the Department of Electronics and Communication Engineering, GLA University, Mathura 281406, India. E-mail: manish.kumar@gla.ac.in.

* To whom correspondence should be addressed.

Manuscript received: 2022-07-29; revised: 2022-12-10; accepted: 2022-12-29

related illnesses increase medical care costs and reduce a person's effectiveness. World Health Organization (WHO) estimates that India lost up to 237 billion US dollar between 2005 and 2015 due to heart-related or cardiovascular illnesses.

In this regard, a reasonable and precise forecast of heart-related infections is critical. Clinical associations gather information on various health-related issues from one end of the world to the other. The heart is the largest organ in our body, more significant than any other. On average, a human heart beats about 60 to 100 times per minute, which translates to about 100 000 times per day. Over the course of an average lifespan of around 80 years, the heart will beat approximately 2.5 billion times. Because of the modern way of life and food, coronary disease is on the rise. The diagnosis of heart disease is a difficult task. This arrangement model will predict whether the patient has heart disease based on various conditions/side effects of their body. The data can be investigated using various AI procedures to gain valuable insights. In any case, the amount of information gathered is massive, and it is frequently loud. These datasets, which are too complex for human minds to comprehend, can be quickly investigated using various AI techniques. As a result, these calculations have recently become helpful in precisely anticipating the presence or absence of heart-related illnesses.

Machine Learning (ML) provides dynamic calculations without a specific program to build an intelligent machine that can simplify various troublesome issues. ML addressed a wide range of issues and divided them into three sections. Issues are directed, solo, or supportive in nature. There are two types of issues managed: characterization and relapse. A bunching-type issue can be resolved using ML calculations on its own. ML relegated various calculations based on the nature of the problem.

1.2 Role of artificial intelligence in the medical field

Artificial intelligence is a way of manipulating and extracting implicit, previously unknown/known, and potentially helpful information about data. AI is a vast and diverse field, and its scope and implementation are increasing daily. AI incorporates various classifiers of supervised, unsupervised, and ensemble learning which are used to predict and find the accuracy of the given dataset. Heart disease prediction is an online AI application prepared by a UCI dataset. The client inputs its clinical subtleties to forecast coronary illness for that client. The calculation will compute the likelihood of the presence of heart disease.

1.3 Major contribution

The primary objective of this paper is to develop a straightforward model to use in the medical field. The patient's clinical data will be entered, and based on those details, the algorithm will identify the heart disease and classify it. The system will consist of a website where users will register to receive a report on their heart health based on an analysis of their risk for heart disease. The users must initially fill out a form for registration. The users will then be redirected to the profile page, where they must complete their profile by entering all heart-related information. After submitting their health information, patients can view a report detailing their hearts' status or risk level in percentage form. The system will consist of python code, where users' symptoms will be stored as a dataset, and we will predict whether the person or patient has heart disease using these data.

1.4 Paper outline

Section 1 points to the importance of heart, status of heart disease worldwide, the types of algorithms in machine learning, and how ML projects progress. Section 2 reviewed the existing work, including their objective, algorithm, issue-wise solutions approach, and strengths and weaknesses. In Section 3, we have discussed the step and algorithms used for the proposed work. Section 4 discusses the implementation and analysis of the result, and Section 5 contains the conclusion and future work.

2 Literature Review

This section reviewed the different existing works and identified the existing research gap. Reference [1] meant to foresee coronary illness at the beginning phase so that the disease can be cured on time and we can prevent heart attacks. Usually, heart attacks are not something that takes place suddenly. Instead, it is a result of a particular person's lifestyle for the past few decades. Ali et al.[1] wanted us to detect the initial symptoms that may cause heart disease very early so that specialists and patients get sufficient time to handle the disease. A coronary attack happens when your heart can not siphon sufficient blood and supply it to the body parts instead of not getting enough blood because too-blocked arteries reside in a heart attack. How can information science be utilized to distinguish the issue earlier? Here, we use calculations like Naïve Bayes, SVM, Artificial Neural Networks (ANN), and Hybrid Naïve Bayes, ANN, and SVM.

The attributes used are age, sex, kind of chest torture, resting circulatory strain, serum cholesterol, fasting glucose, resting electrocardiographic results, most noteworthy, heartbeat cultivated, several critical vessels shaded by fluoroscopy, and thallium scan. Different factors utilized are food propensities, history of misery, resting hours, parental history, proactive tasks, way of life, and mental pressure test results.

The parameters by which the above three algorithms are compared are accuracy, specificity, and sensitivity.

Accuracy obtained is as follows:

- Naïve Bayes is 82.97,
- ANN is 85.30,
- SVM is 86.12, and
- Hybrid Naïve Bayes, SVM, and ANN is 88.54.

Ali et al.[2] surveyed predicting and detecting heart disease at an early stage using machine learning algorithms. Heart disease is not something that appears suddenly. It results from a series of events in a person's life due to his/her lifestyle. Some of the supervised learning algorithms like decision trees, SVM, ANN, and Naïve Bayes algorithms are considered in this paperwork.

There are a few kinds of coronary illness on the planet, like coronary conduit infection, vascular sickness, heartbeat issues, primary coronary illness, and cardiovascular breakdown. Coronary artery disease happens since conduits in the heart get hindered. Heart rhythm disorder is nothing but some disturbance in the rhythmic nature of the heart; it either gets too fast or too slow, which is somewhat abnormal to actual behavior. Primary coronary illness is a direct result of the disruption of the complex design heart structure that will cause cardiovascular breakdown. Cardiovascular breakdown is another sickness that happens due to either respiratory failure or hypertension. Traits are sexual orientation, age, resting circulatory strain, chest torment, serum cholesterol, fasting sugar, pulse, Electrocardiogram (ECG) results, old pinnacle, and thalassemia.

Here some expert automated systems are also summarized in the algorithm system used, and the accuracy system is obtained. All are compared so that a better automated system can be created for people to track their health. Information mining and AI assume a vital part in foreseeing coronary illness. Reference [3] proposed another way to deal with anticipated coronary illness or cardiovascular infection utilizing various AI calculations, for example, versatile boosting, strategic relapse, multi-target developmental fluffy classifier, and hereditary fluffy system-logit boost. The precision and consequences of every classifier are contrasted, and it is the best accessible classifier for cardiovascular expectation. Likewise, two free tools were utilized, namely Weka and Keel.

The proposed approach is as follows. Initially, records of the patient and his/her medical records are stored in a medical database. Data are pre-processed, then the feature selection and extraction process is performed on the pre-processed data. After that, classification algorithms are applied to the data using Weka and Keel tools. Two models have been created, and their performance has been evaluated. Out of the two models, the best classification model is selected. This is how we train the algorithm and now test the model. For this, we consider other patients' data not presented in a dataset and then check how our model works and its accuracy.

Characteristics utilized here are age, sex, chest torture type, resting circulatory strain, serum cholesterol, fasting glucose, resting electrocardiographic results, most intense pulse accomplished, workout incited angina, stroke wretchedness initiated by practice comparative with rest, number of significant vessels hued by fluoroscopy, practice thallium scintigraphy, and the anticipated quality. Boundaries on which the exhibition of calculations are assessed depend on precision, affectability, explicitness, and blunder rate utilizing Weka and Keel instruments.

Disease diagnosis identifies the disorder, disease, health issue, or any other condition a person may have. Usually, disease diagnosis is made manually, but the method is prone to errors.

Reference [4] reviewed literature from the last 10 years (2009–2019) and considered 8 most frequently used databases that most of the articles have used. Other techniques like fuzzy logic, machine learning, artificial intelligence, and deep learning are also discussed. This is a review paper that talks about the difference between previous and currently used techniques in diagnosis. Usually, the sequence of events while diagnosing a disease is as follows:

- The clinical history,
- Actual assessment,
- Performing diagnostic tests,
- Drawing conclusion.

Heart disease is one of the significant issues the world is confronting, and it is essential to analyze it at the beginning phase to save individuals' existence[5]. A clinical choice emotionally supportive network (CDSS) can be utilized to analyze the process of decision making in the clinical environment. This investigation depends on Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

Heart Disease Clinical Decision Support System (HDCDSS) is created in Python using Flask as Python Web Server Gateway Interface alongside Bootstrap for information portrayal. The information base utilized was MongoDB. HDCDSS is a straightforward and promising path for clinical staff to analyze the patient's current state. Two public datasets are used in this paper, Stat log and Cleveland.

The calculations utilized are Naïve Bayes, strategic relapse, multi-facet perceptron, decision tree, SVM, and irregular timberland.

The precision obtained by Stat log dataset is 95.90% and by Cleveland dataset is 98.40%.

The assumption of coronary disease using data mining methodology has been a twenty-year-long effort[6]. The majority of papers have completed several data burrowing strategies for the detection of coronary ailment, such as decision tree, Naïve Bayes, neural association, digit thickness, commonly described social affairs, stowing computation, and sponsorship vector machine, with varying degrees of precision on different informational indices of patients around the world. Dimopoulos et al.[6] evaluated various decision tree representations to improve the accuracy of coronary disease prediction using Weka. The decision tree is constructed by analyzing server farms, which are used to evaluate the size of existing features. J48 estimation is an extension of the ID3 algorithm and may produce a small tree. It employs a division and beats method to create decision trees. At each tree node, the estimation selects a characteristic that can partition the models into subsets. Each leaf's center tends to represent a category or choice.

By breaking down the trial results, it is closed that J48, the tree strategy, ended up being the best classifier for coronary illness expectation since it contains more exactness and the ideal opportunity to assemble.

Approach and methodology. The accompanying targets are set for this heart forecast framework.

- The gauge system should not expect any primary data about the patient records. It is taking a gander.
- The picked framework should be versatile to run against an enormous information base with a large amount of information.

The primary goal of this work is to develop a user-friendly interface whereby the patient's clinical details are inputted and the computation is able to identify cardiac ailment and classify it[7].

Coronary illness expectation is an online AI application arranged by a UCI repository. The customer inputs its clinical nuances to get the conjecture of coronary disease for that customer. The estimation will discover the probability of the presence of coronary ailment.

The system includes a website where clients register to receive a report on their hearts' health based on a predictive analysis of their coronary disease. The client should initially fill out an enrollment form. The client will then be redirected to the profile page, where they must complete their profile by entering all heart-related information. Suppose the client's risk is greater than 60%, in that case, he/she will be redirected to a different structure where he/she will be prompted to enter various symptoms so the system can predict the classification of coronary illness from the two most common classifications.

Our framework will carry out the accompanying three calculations:

- Support Vector Machine (SVM) — This one gives 64.4% precision.
- Logistic regression — It gives 61.45% precision.
- Naïve Bayes algorithm — It gives 60% precision.

AI is perhaps the most problematic advancement of this age. AI is, as a rule, vigorously utilized in every one of the areas, including fabricating, medical care, research and development, and so forth. It is additionally a new pattern on the planet and is shaping a significant part of software engineering because of its ability to robotize things, as shown in Table 1.

Here we are anticipating coronary illness events dependent on some essential qualities which are the most appropriate informational index that we have gathered. The renowned choice tree model is being carried out in our task dependent on the indications explicitly the characteristics needed for the forecast.

The decision tree is based on the property with the highest data gain, and the Gini index is determined based on the best divided trait with the highest split. The decision tree calculation attempts to solve the problem

Table 1 Parameters[8] used for heart diseases.

Parameter	Description
ID (Integer)	For record number
Gender (Integer)	(1) male and (2) female
Age (Integer)	Current age
Systolic Blood Pressure (SBP) (Integer)	−150 mmHg to 16 020 mmHg
Height (Integer)	55 cm to 255 cm
Weight (Integer)	10 kg to 200 kg
Cholesterol (Integer)	(N) standard, (H) above average, and (W) well above normal
Diastolic blood pressure (Integer)	70 mmHg to 11 000 mmHg
Cardiovascular disease (Integer)	Based on the case as user data
Smoking (Integer)	Based on the case as user data
Glucose (Integer)	(N) normal, (H) above normal, and (W) well above normal
Physical activity (Integer)	Based on the case as user data
Alcohol intake (Integer)	Based on the case as user data

Note: 1 mmHg=0.133 kPa

by employing tree representation. Each inner hub of the tree corresponds to a characteristic, while each leaf hub corresponds to a class name. Additionally, we utilize the support vector machine to organize the datasets based on the class mark. The support vector machine handles the piece and employs a hyperplane to group datasets.

If scikit is familiar with the apparent qualities in the class, the mark should be converted into the twofold qualities, 1 and 0. To accomplish this, we imported a name encoder. We then applied the cross-validation to a portion of the first dataset, dividing it into a training set and a test set in the proportions of 75% and 25%, respectively. Then, we utilized the decision tree and the support vector machine to process the outcome, and we discovered that the decision tree performed better than SVM, as the precision for expectation on the test set was 100% for the decision tree and 55% for SVM.

Figure 1 shows the architecture of the model proposed. First import dataset, then process the data, and use classify tab perform predictions. Later, apply the ML algorithm to the dataset to create a model. Segregate test and train data from the taken dataset.

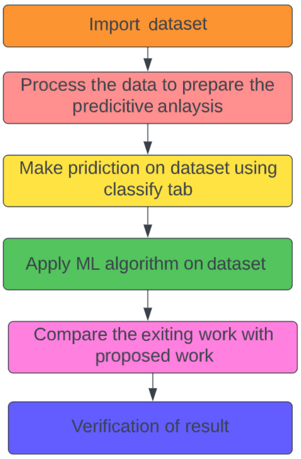

Import dataset
Process the data to prepare the predicitive anlaysis
Make pridiction on dataset using classify tab
Apply ML algorithm on dataset
Compare the exiting work with proposed work
Verification of result

Fig. 1 Architecture of the proposed model.

Data about cardiovascular disorders are used in this analysis. This dataset, which we believe to be the third study to use, has information on around 70 000 patients and 11 features. We also try out several machine learning and deep learning algorithms to see which is the most effective at making accurate predictions about heart disease.

Elshafie et al.[8] used an existing approach on the ADNI dataset that optimizes the computational cost over other existing prediction methods. Other existing work[9] has higher computational cost in terms of time complexity as the number of parameters increases the time complexity of the model. In this work, we have deployed a combination of strategies that work together to accomplish a single goal: feature extraction from datasets, 3D-to-2D grayscale picture conversion, and image size reduction by clipping[10].

3 Design and Implementation

In this section, we have discussed the proposed work's design and implementation.

3.1 Architectural design of proposed work

Here we implemented the model using ML algorithms in which we have implemented Support Vector Machine (SVM) and different algorithms.

Figure 2 explains the architectural design of the model: the dataset is stored, the data stored are pre-processed, and then the data classification is done using support vector machines and logistic regression to determine whether the patients have heart disease or not. The final stage is to predict the disease in a patient using the symptoms we had in the dataset.

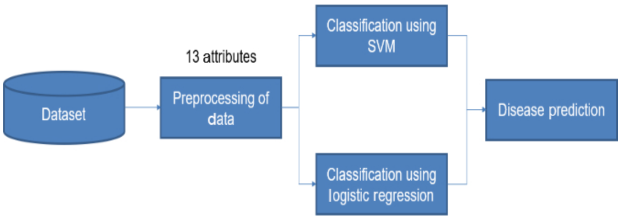


Fig. 2 Architectural design of proposed work.

3.2 Details of dataset used

There are 14 independent features and 1 target feature.

(1) age;

(2) sex;

(3) cp (chest pain type) (4 values);

(4) trestbps (resting blood pressure);

(5) chol (serum cholesterol in mg/dl);

(6) fbs (fasting blood sugar) > 120 mg/dl;

(7) restecg (resting electrocardiographic results) (values 0, 1, and 2);

(8) thalach (maximum heart rate achieved);

(9) exang (exercise-induced angina);

(10) oldpeak = ST depression induced by exercise relative to rest;

(11) slope of the peak exercise ST segment;

(12) ca (number of major vessels (0–3) colored by fluoroscopy);

(13) thal: 3 = normal; 6 = fixed defect; and 7 = reversible defect;

(14) Target is the binary target variable. 0 indicates that the patient has heart disease, and the value is 1 if not.

3.3 Proposed work

In addition to the traditional categories of malignant (M) and benign (B), premalignant has been added to the existing benign cancer classification. This group of patients is eligible for individualized care and treatment options. Concerning data security, categories such as copied, transferred, or retrieved data must remain vigilant for suspicious activity. Classification is the process of labeling data to make it more accessible and valuable. Minimizing duplicate data can help us save money on storage and backup costs[11]. This means that in certain situations, processing time may be drastically reduced.

SVM is an algorithm that teaches machines to distinguish independently between things belonging to different groups. The instance closest to the hyperplane determines its margin of separation, so hyperplanes with the most significant separation between classes in an SVM model would have the most separation margins.

Maximizing profits is a fundamental objective for SVM. The classification is supplied by $y_i \in \{-1, 1\}$, where $i = 1, 2, \ldots, N$ with training data N, and the classification is carried out given the training dataset and a feature vector of size n[12].

Weight vectors are constructed by subtracting rescaled hyperplane classifiers that fulfill Eq. (1) from the original weight vector, and this weight vector is used in Eq. (1) to produce a linear classifier in Eq. (2).

It is feasible to solve Formula (3) by utilizing Formula (4)'s Lagrange function, resulting in the values:

$$\omega = \sum\nolimits_i \alpha_i y_i x_i \tag{1}$$

$$b = y_i - \omega^{\mathrm{T}} x_i \tag{2}$$

The Lagrange multipliers denote inequality constraint multipliers for each inequality constraint[13].

$$\begin{cases} f(x) = \operatorname{sign}(\omega x + b); \\ y_i(\omega^{\mathrm{T}} x_i + b) \geqslant 1,\ i = 1, 2, \ldots, N \end{cases} \tag{3}$$

$$\min \frac{1}{2}\omega^{\mathrm{T}}\omega \text{ s.t. } y_i(\omega^{\mathrm{T}} x_i + b) \leqslant 1,\ i = 1, 2, \ldots, N \tag{4}$$

$$L(\omega, b, \alpha) = \frac{1}{2}\omega^{\mathrm{T}}\omega + \sum_i \alpha_i (1 - y_i(\omega^{\mathrm{T}} x_i + b)) \tag{5}$$

In Formula (3), a slack variable is added to Eq. (2) since no separable data can be separated by Eq. (5). 0.5^{T} is multiplied by the penalty constant C to get an Eq. (5). A penalty constant C can be added to your account if you classify a training example incorrectly. A hyperplane with a wide margin of safety will be selected for optimizations with lower values of C, which may lead to more points being incorrectly categorized[14].

There are various choices for parameter C, and the ideal performance is assessed using either a separate validation set or cross-validation, depending on which approach is used.

$$y_i(\omega^{\mathrm{T}} x_i + b) \geqslant 1 - \xi_i, i = 1, 2, \ldots, N \tag{6}$$

$$\text{minimize} \frac{1}{2}\omega^{\mathrm{T}}\omega + C\xi_i \text{ set } y_i(\omega^{\mathrm{T}} x_i + b) \leqslant 1 - \xi_i,$$
$$i = 1, 2, \ldots, N \tag{7}$$

where ξ_i is the margin.

A linear hyperplane can separate support vectors that are not linearly separable when used in conjunction with kernel functions. The Radial Basis Function (RBF) kernel is often used as the kernel function in nonlinear SVM classifiers. From the feature map, we may deduce $\phi(x)$, which is $K(x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j)$.

The RBF kernel may be found in Formula (7). To decrease the number of terms in Formula (7), we utilize the formula $\gamma = 1/2\sigma^2$. It is possible to get Formula (7) by using the Lagrange function, which answers the two-part issue in Formulas (6) and (7). Finally reached the conclusion promised by Formula (7), we may now take a breather.

3.4 Confusion matrix

The confusion matrix is the representation of actual data values versus predicted data values. It is the tabled structure and is used to measure the performance of our classification model in machine learning[15].

Components of the confusion matrix are

• **True positive:** The values in original data were positive and are also predicted positive.

• **False positive:** The values were negative originally but are predicted positive, which is incorrect.

• **False negative:** The values were positive in actual data but are predicted wrongly, that is, predicted negative.

• **True negative:** The values in the original data were negative and are also predicted negative.

Other evaluation terminologies we come across using the confusion matrix are

• Accuracy is the ratio between total correct predictions and all the predictions done by the classification model.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\%.$$

• Recall or sensitivity[15] is the ratio of the true positive and the total number of positive outcomes we get in the classification model evaluation.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%.$$

• Precision is used to know the actual positive outcomes out of all the positive outcomes predicted by the model[16].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%.$$

We have discussed the theoretical aspects of the model, including an explanation of SVM, logistic regression, libraries used in python code like matplotlib, seaborn, pandas, and random libraries, the concept of confusion matrix, etc.

3.5 Discuss input and variable

We have used the UCI dataset[17] of heart disease prediction.

The steps involved in this process are shown below:

(1) Read the input image.

(2) Convert the image to black and white using Otsu Binarization.

(3) Apply the discrete wavelet transform. It allows the analysis of images at various levels of resolution.

(4) Extract useful features using principal component analysis.

(5) Apply a grey-level co-occurrence matrix to find the homogeneity, mean, standard deviation, variance, smoothness energy, entropy, RMS, contrast, correlation, kurtosis, skewness, IDM, mean square error, etc., utilized for the performance evaluation[18].

The algorithm determines which pixel goes below the foreground and background. A picture or image with an abundant grey level is regenerated into fewer grey-level pictures, and comparisons are completed on each pixel intensity with a reference worth (threshold)[19]. If the input picture $f(x, y)$ and the binary version is $g(x, y)$, then $g(x, y) = 1$ if $f(x, y) \geqslant T$ or 0 otherwise. Principal Component Analysis (PCA) has been used to shrink the number of surplus characteristic sets from the dataset, as shown in Table 2.

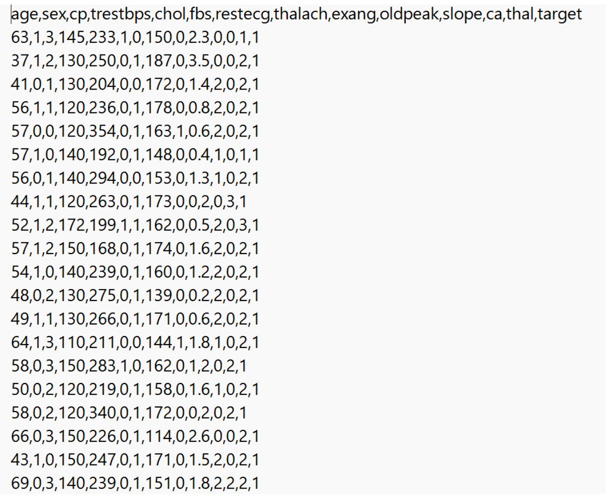
4 Experimental Scenario

We have worked upon two datasets, namely heart dataset and CHD dataset. We have stored these datasets in CSV files, that is, separated value files.

Figure 3 provides a visual representation illustrating the composition of a dataset. Dataset 1, the heart

Table 2 Dataset description of UCI dataset of heart disease prediction of 303 patient record.

Metrics	Age	Sex	cp	trestbps (mmHg)	chol (mg/dL)	fbs (mg/dL)	restecg	thalach (beats per minute)	exang	oldpeak	Slope
Mean	54.366 34	0.683 168	0.966 997	131.6238	246.264	0.148 515	0.528 053	149.6469	0.326 733	1.039 604	1.399 34
Std	9.082 101	0.466 011	1.032 052	17.538 14	51.830 75	0.356 198	0.525 86	22.905 16	0.469 794	1.161 075	0.616 226
Min	29	0	0	94	126	0	0	71	0	0	0
25% of total datasets	47.5	0	0	120	211	0	0	133.5	0	0	1
50% of total datasets	55	1	1	130	240	0	1	153	0	0.8	1
75% of total datasets	61	1	2	140	274.5	0	1	166	1	1.6	2
Max	77	1	3	200	564	1	2	202	1	6.2	2

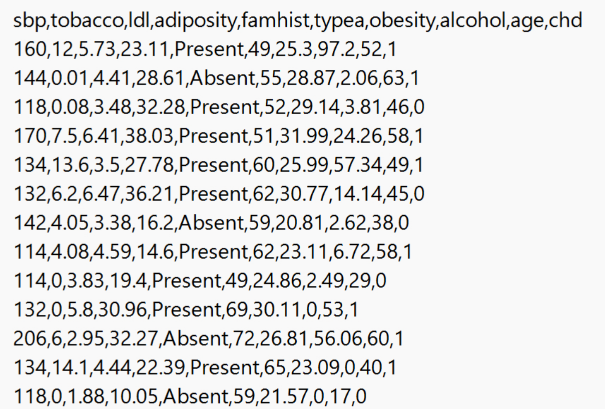


```
age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target
63,1,3,145,233,1,0,150,0,2.3,0,0,1,1
37,1,2,130,250,0,1,187,0,3.5,0,0,2,1
41,0,1,130,204,0,0,172,0,1.4,2,0,2,1
56,1,1,120,236,0,1,178,0,0.8,2,0,2,1
57,0,0,120,354,0,1,163,1,0.6,2,0,2,1
57,1,0,140,192,0,1,148,0,0.4,1,0,1,1
56,0,1,140,294,0,0,153,0,1.3,1,0,2,1
44,1,1,120,263,0,1,173,0,0,2,0,3,1
52,1,2,172,199,1,1,162,0,0.5,2,0,3,1
57,1,2,150,168,0,1,174,0,1.6,2,0,2,1
54,1,0,140,239,0,1,160,0,1.2,2,0,2,1
48,0,2,130,275,0,1,139,0,0.2,2,0,2,1
49,1,1,130,266,0,1,171,0,0.6,2,0,2,1
64,1,3,110,211,0,0,144,1,1.8,1,0,2,1
58,0,3,150,283,1,0,162,0,1,2,0,2,1
50,0,2,120,219,0,1,158,0,1.6,1,0,2,1
58,0,2,120,340,0,1,172,0,0,2,0,2,1
66,0,3,150,226,0,1,114,0,2.6,0,0,2,1
43,1,0,150,247,0,1,171,0,1.5,2,0,2,1
69,0,3,140,239,0,1,151,0,1.8,2,2,2,1
```

Fig. 3 Attribute of Dataset 1—heart dataset.

dataset, was composed of 13 attributes, similar to the initial dataset in Fig. 3, along with one target attribute. Therefore, the total number of attributes in this dataset remained consistent at 14. The heart dataset encompassed 304 record values, allowing for an extensive examination and exploration of the data. Dataset 2, the CHD dataset, presented a different set of attributes, totaling 9 in number. Similar to the other datasets, it included one target attribute, resulting in a total of 10 attributes for analysis. In terms of record values, the CHD dataset contained 463 records, offering a substantial amount of data to be processed and studied. By utilizing these diverse datasets, we were able to access and leverage a wide range of attributes and record values, enabling comprehensive analysis and investigation of the data from different perspectives.

In Fig. 4, it offers a comprehensive description of a distinct dataset. This dataset, unlike the previous one, consisted of a larger set of attributes, totalling 76 in number. Additionally, it included 14 subset values, which may represent specific subsets or categories within the dataset. The detailed information provided in Fig. 4 allows for a more comprehensive understanding of the dataset's structure and composition. During our analysis, we employed two different datasets: Dataset 1, also known as the heart dataset[20], and Dataset 2, referred to as the CHD dataset[21].



```
sbp,tobacco,ldl,adiposity,famhist,typea,obesity,alcohol,age,chd
160,12,5.73,23.11,Present,49,25.3,97.2,52,1
144,0.01,4.41,28.61,Absent,55,28.87,2.06,63,1
118,0.08,3.48,32.28,Present,52,29.14,3.81,46,0
170,7.5,6.41,38.03,Present,51,31.99,24.26,58,1
134,13.6,3.5,27.78,Present,60,25.99,57.34,49,1
132,6.2,6.47,36.21,Present,62,30.77,14.14,45,0
142,4.05,3.38,16.2,Absent,59,20.81,2.62,38,0
114,4.08,4.59,14.6,Present,62,23.11,6.72,58,1
114,0,3.83,19.4,Present,49,24.86,2.49,29,0
132,0,5.8,30.96,Present,69,30.11,0,53,1
206,6,2.95,32.27,Absent,72,26.81,56.06,60,1
134,14.1,4.44,22.39,Present,65,23.09,0,40,1
118,0,1.88,10.05,Absent,59,21.57,0,17,0
```

Fig. 4 Attribute of Dataset 2—CHD dataset.

Figure 5 shows the graphical analysis of accuracy we got in our reviewed research papers. In this section, we have seen the paper's design and implementation details, including architectural design, the dataset used, the software used, performance evaluation, etc. In Section 5 we will walk through the experimental result and its analysis.

5 Experimental Result and Analysis

This section will discuss the analysis done by considering all four scenarios. Also, we will see visual comparisons between various attributes of Dataset 1 that is the primary dataset we are working on.

(1) For attribute ca, the maximum is 0 and the minimum is 4.

(2) For attribute thal, the maximum is 2 and the minimum is 0.

(3) There are more males than females in the dataset.

(4) Typical and non-typical angina is the most

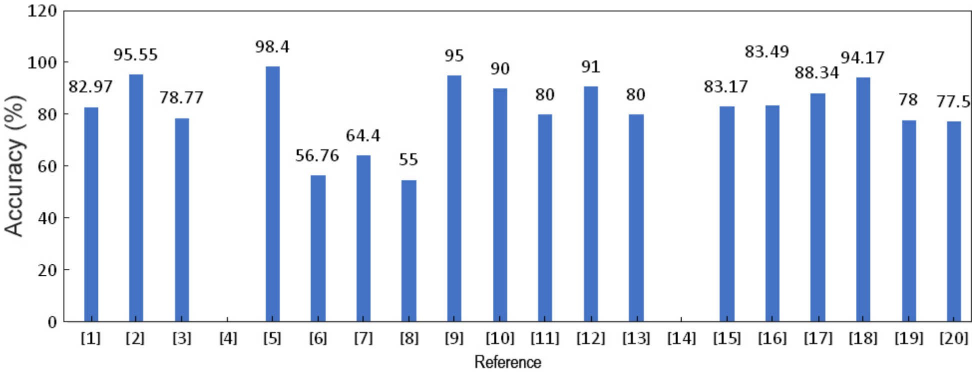


Fig. 5 Graphical analysis of accuracy of exiting work.

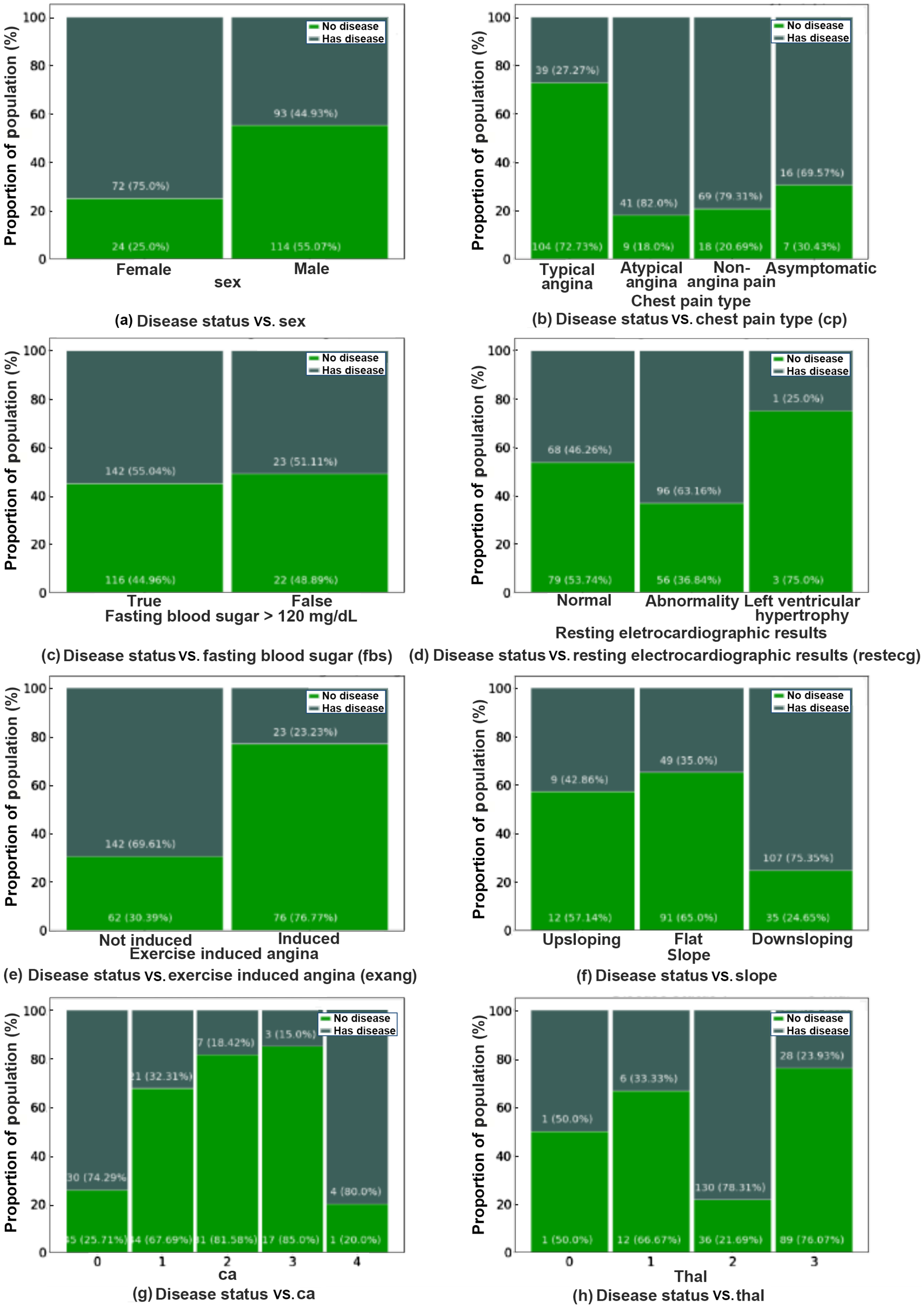
common chest pain[22].

Figure 6 concludes that

(1) The sample is biased when considering the attribute “sex”.

(2) Typical and non-typical chest pains are joint.

It analyzes different attributes and their weight in



(a) Disease status vs. sex

(b) Disease status vs. chest pain type (cp)

(c) Disease status vs. fasting blood sugar (fbs)

(d) Disease status vs. resting electrocardiographic results (restecg)

(e) Disease status vs. exercise induced angina (exang)

(f) Disease status vs. slope

(g) Disease status vs. ca

(h) Disease status vs. thal

Fig. 6 Stacked bar charts of attribute.

predicting heart disease.

Chest pain (cp) and target have a positive link in the heat map above. Those with a high risk of experiencing chest discomfort are more likely to develop heart disease. The goal is positively correlated with thalach, slope, and resting, in addition to discomfort in the chest[23].

In other words, when we exercise, our hearts need more blood, but because our arteries are restricted, blood flow is slowed down. There is a strong negative association between the goal and ca, old peak, and thal.

Figure 7 concludes that no attribute has a value greater than 0.5.

Figure 8 concludes that the balance target variable and attributes like age, chol, and treetops are distributed throughout the dataset.

Figures 9–12 show the classification technique using different algorithms.

(1) Heart disease dataset (SVM) vs. heart disease dataset (LR)

(a) By SVM — 67%;

(b) By logistic regression — 82%.

(2) CHD dataset (SVM) vs. CHD dataset (LR)

(a) By SVM — 72%;

(b) By logistic regression — 72.6%.

(3) Heart disease dataset (SVM) vs. CHD dataset (SVM)

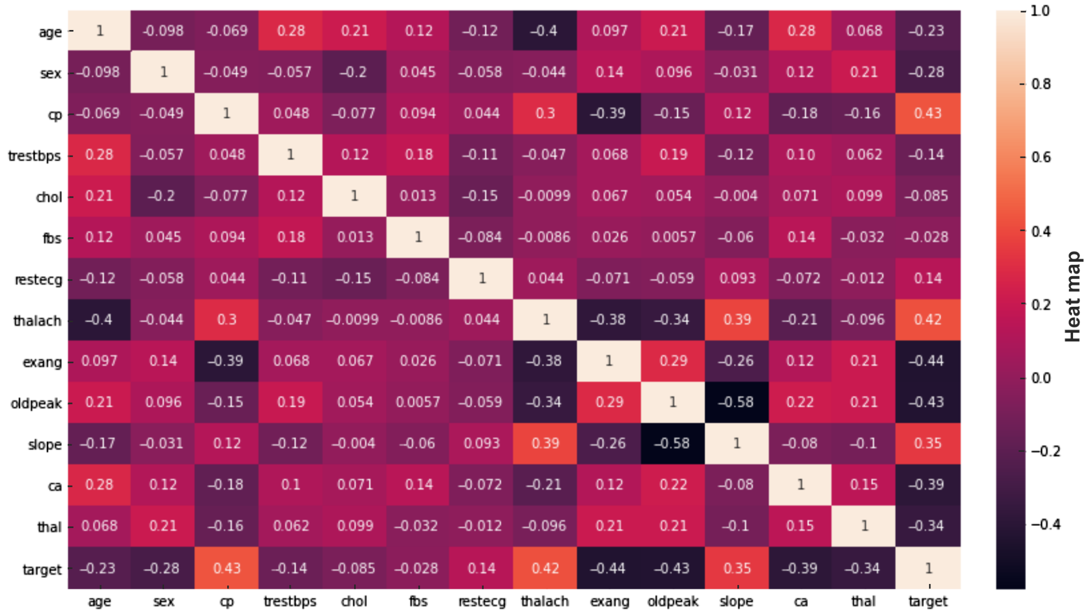


Fig. 7 Heatmap of attributes.

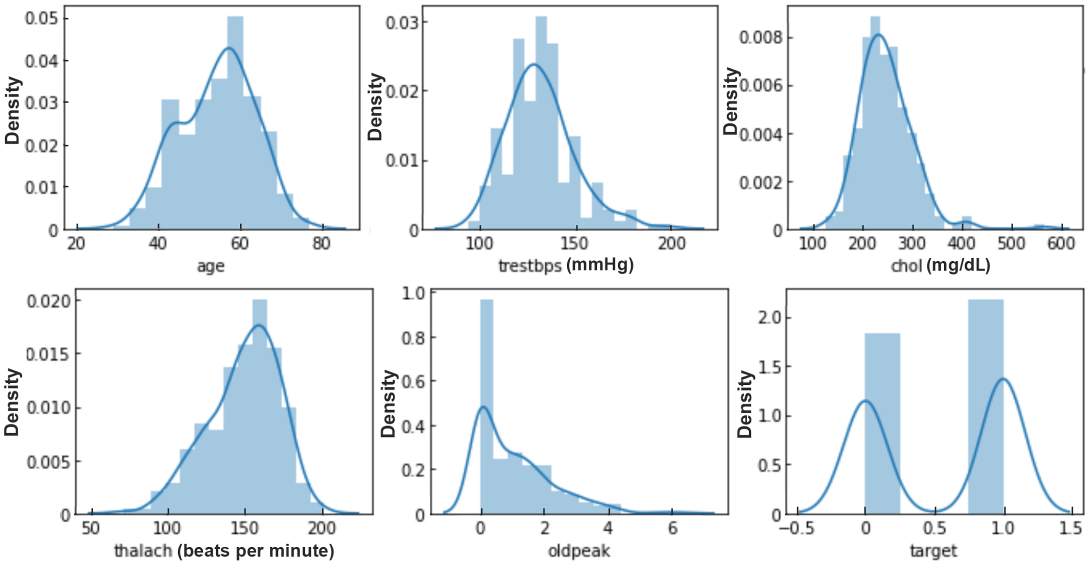


Fig. 8 Distance plot of attributes.

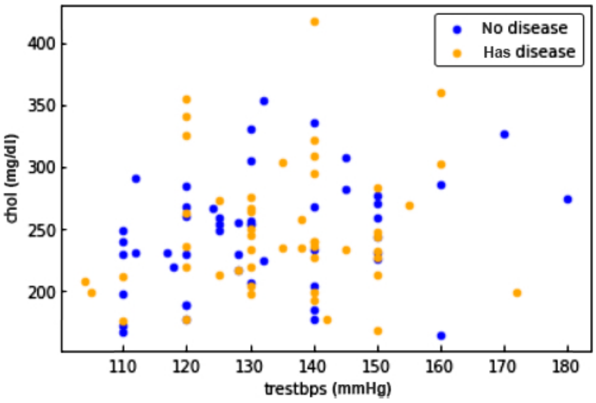


Fig. 9 Dataset 1 graphical snapshot for SVM.

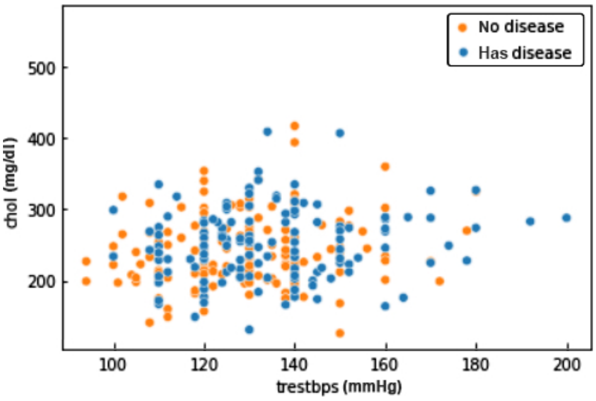


Fig. 10 Dataset 1 graphical snapshot for LR.

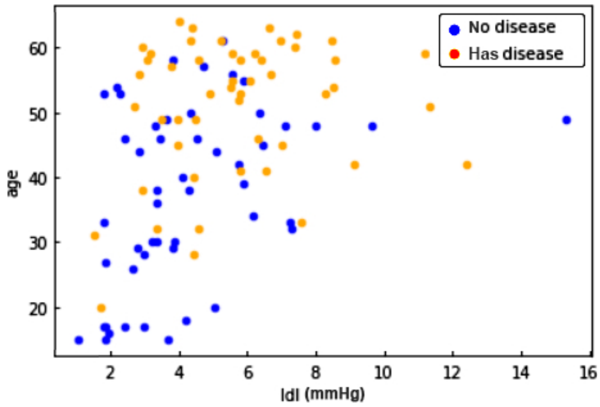


Fig. 11 Dataset 2 graphical snapshot for SVM (ldl is a parameter, which identifies that person is infected by disease or not.)

(a) By SVM for Dataset 1— 67%;

(b) By SVM for Dataset 2— 72%.

(4) Heart disease dataset (LR) vs. CHD dataset (LR)

(a) By logistic regression for Dataset 1— 82%;

(b) By logistic regression for Dataset 2— 72.6%.

We have discussed the analysis done by considering all four scenarios and also have seen visual comparisons between various attributes of Dataset 1.

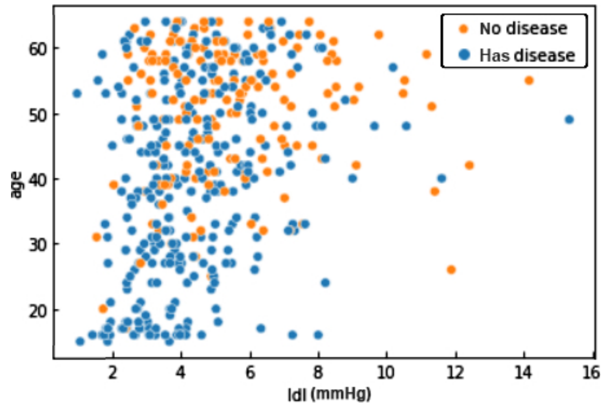


Fig. 12 Dataset 2 graphical snapshot for LR.

6 Conclusion and Future Scope

Heart disease is one of the leading problems arising worldwide as people's lifestyle is degrading daily because of their work and their routine. So, there is a need we must and be able to detect heart disease so that we can save a good amount of people from losing their lives. As computer science engineers, we try to play our role in this medical field world, and the problem will grow slowly and gradually. We have taken a dataset and performed the classification algorithms, namely support vector machine and logistic regression, which classify whether the patient has heart disease. So that doctors can further work on a person's health and help them cure and move toward a better lifestyle.

We have taken two datasets: the heart dataset and the CHD dataset. Both are in the form of a CSV file, i.e., a comma-separated value file. We applied support vector machine and logistic regression on both datasets and found graphical classification and accuracy for each.

We compared the accuracy by considering four different scenarios.

(1) Heart disease dataset (support vector machine) vs. heart disease dataset (logistic regression)

(a) By SVM — 67%;

(b) By logistic regression — 82%.

(2) CHD dataset (support vector machine) vs. CHD dataset (logistic regression)

(a) By SVM — 72%;

(b) By logistic regression — 72.6%.

(3) Heart disease Dataset (support vector machine) vs. CHD dataset (support vector machine)

(a) By SVM for Dataset 1— 67%;

(b) By SVM for Dataset 2— 72%.

(4) Heart disease dataset (logistic regression) vs. CHD dataset (logistic regression)

(a) By logistic regression for Dataset 1— 82%;

(b) By logistic regression for Dataset 2— 72.6%.

So, we conclude that we have successfully implemented both the SVM and LR on different datasets and compared their accuracy according to their key attributes. After reviewing 20 papers on the topic of heart disease prediction, we learned the methods and different algorithms to predict whether the patient has heart disease. We have selected SVM and logistic regression for classification. Table 3 shows the comparative study of exiting methods with proposed work and our proposed work outperformed with every parameter shown in Table 3. Specificity is calculated by dividing the number of true negative predictions by the sum of true negatives and false positives. It focuses on the proportion of correctly identified negative instances out of all the actual negative instances in the dataset. F1-score is a commonly used performance metric for classification models that combines precision and recall into a single value. It provides a balanced measure of a model's accuracy by considering both the model's ability to correctly identify positive instances (precision) and its ability to capture all positive instances (recall).

Everything needs to be updated or upgraded quickly to cope with the current technology. Apart from gradation, there can be many other features that could add up to the proficiency and ability of our prediction model. For universal applicability, we may easily add more functionality. We can also fuse various algorithms to get desired features in a single algorithm, which can help us attain better accuracy. Also, we fuse these two features to attain better accuracy to develop an algorithm more efficient than SVM.

To evaluate the effectiveness of the models, their precision was considered. The limitation of the models is to only focus on limited parameters. The outcomes indicated that the ML ensemble model best predicted cardiovascular illness. We have implemented a few new processes to have an analytically viable dataset. In the future, we may use several methods to determine which features are the most useful for our projects. Better and more precise evaluations may be achieved by using more datasets. Last but not least, deep learning and reinforcement learning methods may help solve the problem of prediction to improve the efficiency with which cardiovascular illness is diagnosed.

Table 3 Comparative study of exiting methods.

Model	Accuracy (%)	Precision (%)	Specificity (%)	F1-score (%)
CNN model[16]	96.11	94.44	93.35	95.25
XGBoost	93.47	92.25	93.26	92.15
Random forest[24]	87.88	89.45	87.38	88.35
Proposed work	97.25	92.45	87.22	91.22

References

[1] F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, M. Imran, and K. S. Kwak, A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, *Information Fusion*, vol. 63, pp. 208–222, 2020.

[2] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, *Computers in Biology and Medicine*, vol. 136, p. 104672, 2021.

[3] M. S. Amin, Y. K. Chiam, and K. D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, *Telematics and Informatics*, vol. 36, pp. 82–93, 2019.

[4] K. Andjelkovic, D. K. Ostric, and I. Andjelkovic, Prediction of heart failure in adults with congenital heart disease, *European Journal of Heart Failure*, vol. 16, p. 87, 2014.

[5] T. H. S. Dent, Predicting the risk of coronary heart disease. II: The role of novel molecular biomarkers and genetics in estimating risk, and the future of risk prediction, *Atherosclerosis*, vol. 213, no. 2, pp. 352–362, 2010.

[6] K. Dimopoulos, K. Muthiah, R. Alonso-Gonzalez, S. J. Wort, G. P. Diller, M. A. Gatzoulis, and A. Kempny, Heart or heart and lung transplantation for patients with congenital heart disease in England: Outcomes and future predictions, *Circulation*, vol. 134, p. A17841, 2016.

[7] A. S. Dontas, A. Menotti, C. Aravanis, A. Corcondilas, D. Lekos, and F. Seccareccia, Long-term prediction of coronary heart disease mortality in two rural Greek populations, *European Heart Journal*, vol. 14, no. 9, pp. 1153–1157, 1993.

[8] M. G. Elshafie, A. Hagag, E. S. A. El-Dahshan, and M. A. Ismail, A hybrid bidirectional LSTM and 1D CNN for heart disease prediction, *International Journal of Computer Science and Network Security*, vol. 21, no. 10, pp. 135–144, 2021.

[9] O. E. Emam, A. Abdo, and M. M. Mahmoud, An adaptive heart disease behavior-based prediction system, *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, pp. 374–383, 2019.

[10] L. Harel-Sterling, F. Wang, S. Cohen, A. Liu, J. Brophy, G.

Paradis, and A. Marelli, Risk predictions in adult congenital heart disease patients with heart failure: A systematic review, *Journal of the American College of Cardiology*, vol. 73, no. 9, p. 656, 2019.

[11] Y. L. Huang and A. Sajid, Prediction model of pathogenic gene of coronary heart disease based on machine learning, *Basic & Clinical Pharmacology & Toxicology*, vol. 122, p. 32, 2018.

[12] R. Katarya and S. K. Meena, Machine learning techniques for heart disease prediction: A comparative study and analysis, *Health and Technology*, vol. 11, no. 1, pp. 87–97, 2021.

[13] M. Kavousi, S. Elias-Smale, J. H. Rutten, R. V. Proenca, M. Oudkerk, M. P. D. Maat, A. Hofman, E. Steyerberg, A. H. V. D. Meiracker, and J. C. Witteman, Comparison of novel markers in prediction of coronary heart disease risk: The rotterdam study, *Circulation*, vol. 122, no. 21, p. A17770, 2010.

[14] Erica Research Group, Prediction of coronary heart-disease in Europe. The 2nd report of the WHO-ERICA project, *European Heart Journal*, vol. 12, no. 3, pp. 291–297, 1991.

[15] C. Mufudza and H. Erol, Poisson mixture regression models for heart disease prediction, *Computational and Mathematical Methods in Medicine*, vol. 2016, p. 4083089, 2016.

[16] K. Saaksjarvi, P. Knekt, and S. Mannisto, The prediction of a diet quality index on cardiovascular disease and coronary heart disease mortality, *Annals of Nutrition and Metabolism*, vol. 67, p. 327, 2015.

[17] S. C. Siu, J. Grewal, M. Sermer, J. Mason, M. Kiess, R. Wald, J. Colman, and C. Silversides, Comprehensive prediction of cardiac outcomes in pregnant women with heart disease, *Circulation*, vol. 136, p. A15606, 2017.

[18] C. Sowmiya and P. Sumitra, A hybrid approach for mortality prediction for heart patients using ACO-HKNN, *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 5, pp. 5405–5412, 2021.

[19] H. T. Cate, F. Kontny, and D. W. Nilsen, Editorial: Novel and potential markers for prediction of outcome in patients with acute and chronic coronary heart disease, *Frontiers in Cardiovascular Medicine*, vol. 6, p. 66, 2019.

[20] M. Tomasdottir, N. Hadziosmanovic, C. Held, P. E. Aylward, A. Budaj, C. P. Cannon, J. Engdahl, C. B. Granger, W. Koenig, A. J. Manolis, et al., Biomarkers improve the prediction of incident atrial fibrillation in patients with stable coronary heart disease, *Circulation*, vol. 138, p. A15437, 2018.

[21] S. Udhan and B. Patil, A systematic review of Machine learning techniques for heart disease prediction, *International Journal of Next-Generation Computing*, vol. 12, no. 2, pp. 229–239, 2021.

[22] X. M. Yuan, J. H. Chen, K. Zhang, Y. Wu, and T. T. Yang, A stable AI-based binary and multiple class heart disease prediction model for IoMT, *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 2032–2040, 2022.

[23] D. Q. Zhang, Y. Y. Chen, Y. X. Chen, S. Y. Ye, W. Y. Cai, J. X. Jiang, Y. C. Xu, G. F. Zheng, and M. Chen, Heart disease prediction based on the embedded feature selection method and deep neural network, *Journal of Healthcare Engineering*, vol. 2021, p. 6260022, 2021.

Ankit Kumar is working as an assistant professor at the Department of Computer Engineering and Applications, GLA University, India. His research area is wireless sensor networks. His articles are published in 54 international journals and 11 national journals. He has received 8 patents. His work has been profiled broadly, such as in information security, cloud computing image processing, neural network, and network. His research interests include computer network information security, computational model, compiler design, and data structure. He is a reviewer and editor of many reputed journals.

Kamred Udham Singh received the PhD degree from Banaras Hindu University, India in 2019. From 2015 to 2016, he was a junior researcher, and from 2017 to 2019, he was a senior researcher with University Grant Commission (UGC), India. In 2019, he became an assistant professor at the School of Computing, Graphic Era Hill University, India. His research interests include image security and authentication, deep learning, medical image watermarking, and steganography.

Manish Kumar is presently working as an associate professor at the Department of Electronics and Communication Engineering, GLA University, Mathura, India. He has vast experience in the field of teaching and research. He has published more than 20 research papers in reputed international and national journals. His areas of research include vehicular network, digital image processing, machine learning, and time series data analysis.