IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124202

USE OF MACHINE LEARNING IN HEART DISEASE PREDICTION: A SURVEY

Dinesh Suresh Bhadane[1], Prerana Bedadewar[2], Shital Nalawade[3], Shweta Daphal[4], Shital Gaikwad[5]

Assistant Professor, Dr.D.Y.Patil Institute of Technology, Pimpri, Pune-18[1]

Final Year Students, Dr.D.Y.Patil Institute of Technology, Pimpri, Pune-18[2-5]

Abstract: According to the recent report published by WHO, Cardiovascular diseases are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and includes rheumatic heart disease, cerebrovascular disease, coronary heart disease and other conditions. The basic cause of CVD death is due to heart attacks and strokes and one third of these deaths occur prematurely in folks under 70 years of age. With rapid increase in population, pollution and frequently changing lifestyle of a human being, it becomes a challenge to diagnose a disease and provide the relevant ministration at the right time. With the help of advancements in the technological tools and techniques, machine learning plays a vital role in training and testing the abundant data in the medical field and takes less time in predicting the same with foremost correct and reliable formulas. In this paper we have surveyed the various research papers published in this domain in the recent years and formulated a table which includes various techniques and their corresponding algorithms used with their level accuracy, pros and limitations and also studied the future scope so as to propose a model in the near future which predicts the heart disease with high degree of accuracy and results in robust way of saving the lives at large.

Keywords: KNN, SVM, DECISION TREE, LOGISTIC REGRESSION

I. INTRODUCTION

According to the World Health Organization, 12 million people die each year as a result of heart disease. Early detection of heart disease is critical in making lifestyle changes decisions in high-risk patients. This type of disease can be caused by smoking, high blood pressure, diabetes, obesity, hypertension, cholesterol, and other factors. Cardiovascular disease is a disease that affects people all over the world. This disease can be caused by a variety of heart problems, including coronary artery disease, cardio-vascular disease, stroke, heart failure, and many others**.**

However, machine learning is now a subset model of an artificial intelligence network that employs complex algorithms and deep learning neural networks. Machine Learning is made up of algorithms that use various analytics and statistical techniques to improve the tasks that humans perform naturally and effortlessly on a daily basis. ML is a new AI application that uses various analytics and statistical techniques to improve the performance of specific machine learning from old data. It allows a specific machine to learn from a database and improve its performance through experience. It is beneficial to construct an intelligent machine to solve the specific problem. ML solved a variety of complex problems that statistical algorithms could not solve. ML provides dynamic algorithms that can be used without being explicitly programmed to create an intelligent machine that can solve a variety of difficult problems. ML solved a variety of problems that were divided into three categories. There are three types of problems: supervised, unsupervised, and reinforcement. There are two types of supervised problems: classification problems and regression problems. Unsupervised problems of the clustering type can be solved by ML algorithms. Based on the type of problem, ML assigned different algorithms. The following steps are taken to complete the ML project:

- Establishes a problem statement.
- Subdividing the problem into ML problems.
- Choosing appropriate ML algorithms based on the type of problem.
- Data collection and cleaning.
- Data-driven model training.
- Test the model.
- Evaluate a model based on its accuracy.

DOI: 10.17148/IJARCCE.2023.124202

The existing heart disease database is made up of both numerical and categorical data. These datasets, which are too large for human minds to comprehend, can be explored easily using various machine learning algorithms. Prior to using the algorithm, cleaning and filtering are performed on the database records. As a result, in recent years, the algorithms have proven to be extremely useful in accurately predicting the presence or absence of heart-related diseases.

II. LITERATURE REVIEW

Researchers have proposed various machine learning-based diagnosis techniques in the literature to diagnose HD. In order to explain the significance of the proposed work, this research study presents some existing machine learning-based diagnosis techniques. In medical centers, a number of works have been done on disease prediction systems using various machine learning algorithms.

According to the study, data mining is a subfield of computer science. It is the method by which we extract information from a large amount of data. Every day, new technology emerges, such as artificial intelligence, database management systems, machine learning, and deep learning. The goal of data mining is to find structural data that will provide some reasonable information from a given massive amount of data. The authors proposed using algorithms such as Bayesian and KNN to apply patient data and try to predict heart diseases based on given features. Data classification is one of the most well-known machine learning algorithm tasks. In this case, machine learning is a critical function for extracting knowledge from business activity datasets and transferring it to larger databases.

Some authors elaborated on Diabetes mellitus is a very common disease in many people due to a malfunction in metabolic functionality. As a result, many organs become infected. When it comes to blood veins and nerves. Heart diseases include coronary artery disease, arrhythmias (heart rhythm problems), heart abnormalities (such as congenital heart defects), and a variety of other disorders. This category includes conditions such as cardiomyopathy and heart infections. Chest pain, a symptom of cardiovascular disease, is the most common indicator of heart risk. Following that, it exhibits Nausea, Indigestion, Heartburn, or Stomach Pain. If we can predict early, we may be able to stop any human body from reaching a dangerous stage. Machine Learning techniques provide efficient results for extracting knowledge by creating any predicting model for diagnostic medical datasets collected from patients with various real heart diseases. Using machine learning, we can extract a lot of useful information from this dataset. The authors used popular ML models such as SVM, NB, K-Nearest Neighbours, and c4.5 Decision Tree in this work. In this case DT gives better result in terms of Accuracy or other performance parameters.

The project entails using Python to detect the presence of heart diseases. Chol, treetops, sex, age, and other variables were included in the dataset. The project also made use of a number of other import libraries, including matplotlib, Numpy, Pandas, warnings, and many others. The python programming language was used to assess the outcomes of the specified dataset using a correlation matrix, histogram, support vector classifier, K Neighbours Classifier, Random Forest Classifier, and Decision Tree Classifier. Furthermore, Python is an open-source language that encourages the development of innovative solutions for the health care sectors and provides better outcomes for patients, resulting in improved care delivery.

Where the language also complies with the HIPAA checklist for ensuring medical information security. Diabetes, obesity, an unhealthy diet, being overweight, excessive alcohol consumption, and physical inactivity are the leading causes of heart disease. Some people experience these symptoms during a heart attack. Pain that spreads to the arm, dizziness or light headedness, throat, snoring, and sweating are also possible. Heart attacks, strokes, and coronary heart disease, also known as heart failure and coronary artery disease, are far more common in people over the age of 65 than in those under the age of 65.

DOI: 10.17148/IJARCCE.2023.124202

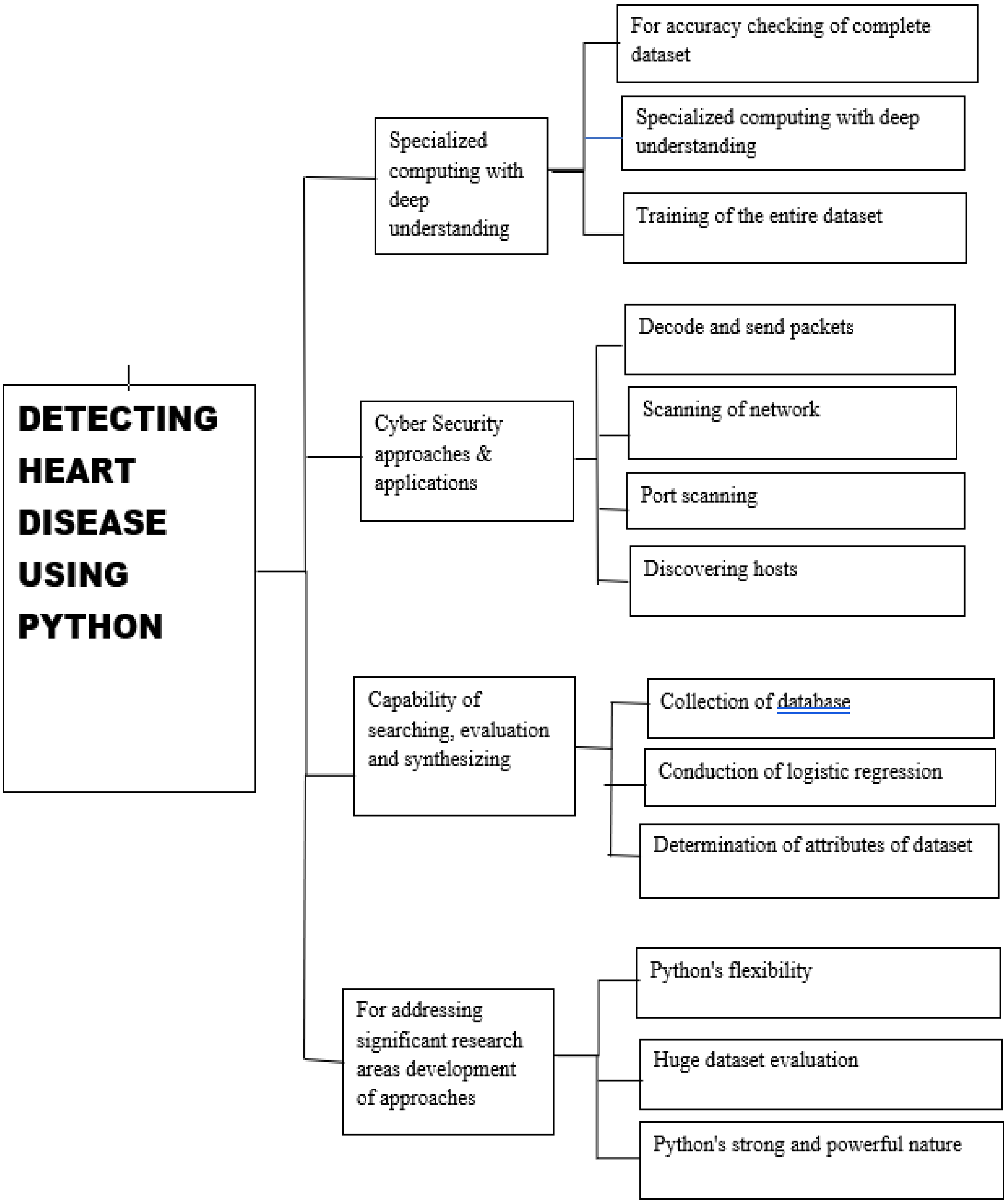
DETECTING HEART DISEASE USING PYTHON
Specialized computing with deep understanding
For accuracy checking of complete dataset
Specialized computing with deep understanding
Training of the entire dataset
Cyber Security approaches & applications
Decode and send packets
Scanning of network
Port scanning
Discovering hosts
Capability of searching, evaluation and synthesizing
Collection of database
Conduction of logistic regression
Determination of attributes of dataset
For addressing significant research areas development of approaches
Python's flexibility
Huge dataset evaluation
Python's strong and powerful nature

Fig. 1 Literature Review

DOI: 10.17148/IJARCCE.2023.124202

TABLE I LITERATURE REVIEW TABLE

Ref.No	*Publisher Name & year*	*Title*	*Techniques Used*	*Accuracy*
1	Mrs.Jayashree L K 2019	Heart disease prediction system	1.KNN	KNN has Better accuracy in android Application.
2	Ranjit Shreshtha. 2019	Heart disease prediction system Using Machine Learning	1.Decision tree 2. Naïve baye	Decision tree has accuracy of 88.163%.
3	Sonam Nikhar 2016	Prediction of heart disease using machine learning algorithms	1.Naïve baye classifier 2.Decision tree	Decision tree has better accuracy when compared to Naïve Baye.
4	Nurul absar 2022	The efficiency of machine learning supported smart system for heart disease Prediction.	1. Decision tree 2.Random Forest 3.KNN 4.Adaboost	Random forest (93.43%) KNN (97.83)
5	Nisha Gupta 2021	Heart disease prediction system Using Machine Learning	1. Decision tree 2.KNN 3.Fuzzy logic 4.SVM 5.ANN 6. Naïve baye	KNN (87%) SVM (83%) DT (79%)
6	V.V Ramalingam 2019	Heart disease prediction system Using Machine Learning techniques	1.SVM 2.KNN 3.Naive Baye 4.DT 5.RF	NB (83.49%) SVM (85.76%) DT (78.46%) RF (97.7%)
7	Prof. (Dr)Kanak Saxena	Efficient Heart disease prediction system	1.SVM	SVM has accuracy 70.59%
8	Hager Ahmad 2019	Heart disease identification from patient social post, Machine learning solution on spark	1.DT 2.Naive baye 3.KNN 4.BPNN 5.LR 6.RF	RF has highest accuracy for real time & can be increased by 11%.
9	Mangesh Limbitote 2020	Prediction technique of heart disease using machine learning.	1.DT 2.KNN 3.LR 4.SVM	Every algorithm has given different result in different situations
10	Mr.Santhana Krishnan.J 2019	Prediction technique of heart disease using machine learning Algorithm.	1.Naive Baye 2.Decision tree	DT (91%)

III. MATERIALS AND METHODS

I. DATASET:

Heart disease data from the Kaggle library was used in this study for testing purposes. We took into account 303 instances, 13 attributes, and one final output label in this dataset. The output label has two classes: one for heart disease presence (labelled as "1") and the other for heart disease absence (labelled as "0").

TABLE II DATASET TABLE

Sr .No	*Attribute*	*Description*	*Type*
1	Age	Display the age of patient	Numerical
2	Sex	Gender of patient(0=female, 1=male)	Nominal
3	Cp	Chest pain type Atypical angina=0 Typical angina=1 Asymptomatic=2 Non-anginal pain=3	Nominal
4	Trest-bps	resting blood pressure (in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl ,values from 126 to 564	Numerical
6	Fbs	Fasting blood sugar > 120mg/dl (true=1,false=0)	Nominal
7	Resting	Resting electro-cardio-graphics result(0 to 1)	Nominal
8	Thali	Maximum heart rate archived(71)	Numerical
9	Exang	exercise included angina(1=yes, 0=no)	Nominal
10	Old peak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slope of the peak exercise ST segment (0 to 1)	
12	Ca	Number of major vessels(0-3) coloured by flourosopy(display the value as integer or float)	Numerical
13	Thal	Display the thalassemia 0 = normal 1 = fixed defect 2 = reversable defect	Nominal
14	Targets	1 or 0 0=No Disease 1= Disease	Nominal

II. PRE PROCESSING OF DATASET: The dataset preparation necessary for effective representation. The dataset has been pre-processed using a variety of methods, including Missing values, Standard Scalar (SS), and Min-Max Scalar.

III. PERFORMANCE METRICS: We employed the Accuracy Score, Confusion Matrix, and Classification Report, three widely used performance evaluation metrics.

Confusion matrix:
To describe how well a classification method performs, a confusion matrix is a table. The output of a classification algorithm is shown and summarized in a confusion matrix.

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positive **TP**	False Positive **FP**
Predicted Negative (0)	False Negative **FN**	True Negative **TN**

Fig. 2 Confusion matrix

Accuracy Score:
This score is used to evaluate the model's performance by comparing the ratio of true positive to true negative predictions. Using a classification report, one can assess the accuracy of the predictions made by a classification algorithm. how many of the forecasts came true and how many didn't. To be more precise, the metrics of a categorization report are predicted using True Positives, False Positives, True Negatives, and False Negatives.

DOI: 10.17148/IJARCCE.2023.124202

$$\text{Accuracy score} = \frac{\text{The number of correct predictions}}{\text{The total number of prediction}} \text{ X } 100$$

$$\text{Accuracy score} = [(TP+TN)/(TP+TN+FP+FN)]*100$$

Precision : Precision is the accuracy of successful forecasts. The number of positive class forecasts that really fall within the positive class is measured by precision.
Precision is TP/(TP + FP).

Recall: Recall measures how many correct class predictions were produced using all of the successful cases in the dataset.
Recall is TP/(TP+FN).

F1 Score is referred to as the harmonic mean of recall and precision.
Support: F1 Score = 2*(Recall * Precision) / (Recall + Precision).

Support = Support is the number of instances of the class that actually appear in the given dataset.

IV.METHODOLOGY

IMPORT LIBRARIES :

1) Numpy:
Numpy is a general-purpose array-processing library that may be used to work with exhibitions. It offers a multidimensional array object with outstanding speed as well as capabilities for interacting with these arrays. It is the cornerstone Python module for scientific computing.
2) Pandas:
The primary focus of the pandas library is dataset manipulation, which includes editing, replacing, and adding new Data-Frame object parts. The calculation of descriptive statistics and the presentation of the columns and rows in a data collection are only two examples of the diverse services that pandas offers.
3) Matplotlib:
This extensive Python package allows users to build interactive, animated, and static visualizations. Matplotlib makes difficult things possible and simple things easy.4) train_test_split: To part the dataset into preparing and testing information. The train_test_split() method is used to split our data into train and test sets.
4) Standard Scaler:
To scale each highlight individually so that the Machine Learning model may more effectively adapt to the dataset.
ADD A DATASET: After obtaining the dataset from Kaggle, I saved it under the name dataset.csv in my working registry. I then read the dataset using read csv() and saved it to the dataset variable.

DATA COLLECTION: We obtained data from Kaggle.com, a source of datasets, using the Cardiovascular Disease dataset, 2019, published by Svetlana Ulianova.70,000 patient records from the collected dataset, which includes 14 characteristics, A dataset is a collection of data or a tool that is necessary for any project or type of research.

PRE-PROCESSING: Segregation of target data and feature data as training and test data. Before training the machine learning models, scale all the values in the data so that they are between 0 and 1.

SUPPORT VECTOR MACHINE: Support A very well-liked supervised machine learning method known as the vector machine can function as both a classifier and a predictor. It locates a hyper-plane in the feature space that distinguishes between the classes for classification.
KNN: One of the simplest but most powerful categorization methods is the K-Nearest Neighbor algorithm. It is typically used for classification jobs where there is little to no prior knowledge about the distribution of the data and makes no assumptions about the data. With this approach, the data point for which a target value is unavailable is located together with the k nearest data points in the training set, and the average value of those data points is then applied to that data point.

DECISION TREE: Most classification-related issues are addressed by this strategy. Both continuous and categorical qualities are performed with ease. Based on the most important predictors, this algorithm splits the population into two

DOI: 10.17148/IJARCCE.2023.124202

or more related groupings. The entropy of each and every attribute is initially calculated by the Decision Tree algorithm. The variables or predictors with the greatest information gain or the lowest entropy are then used to divide the dataset. Recursively applying these two procedures to the remaining properties.

LOGISTIC REGRESSION: One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of independent factors, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. It can be True or False, Yes or No, 0 or 1, etc., but rather than delivering an exact value between 0 and 1, it delivers probabilistic values that are in the range of 0 and 1. In logistic regression, we fit a "S" shaped logistic function, which predicts two maximum values, rather than a regression line (0 or 1). The logistic function's curve shows the likelihood of such an event. based on its weight, a mouse is obese or not, depending on whether the cells are malignant, etc.

V. ALGORITHM

- Step 01: Store Data from Kaggle Repository
- Step 02: Import Prior Libraries:
- Step03: Now Import our Required Dataset
- Step04: Apply Feature Extraction

Data Conversion
Apply Encoding Techniques

- Step 05: Visualize Data for better understanding
- Step06: Applying Machine Learning Algorithms
- Step07: Apply Different Model
- Step08: Repeat Step07 for many times with different Algorithms
- Step09: Finally Compare Results with performance parameters like Accuracy.

VI. SYSTEM ARCHITECTURE

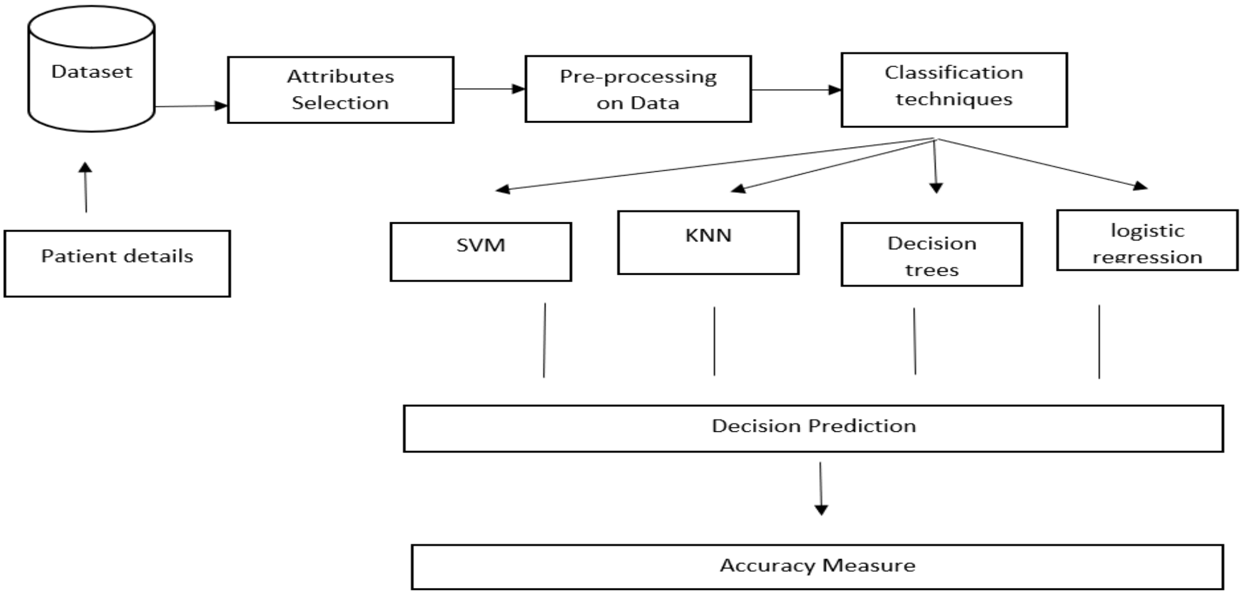

Fig. 3 System Architecture

The comprehensive system architecture of the Heart Disease Prediction System has shown how all of the system's components and their interrelationships are easily understood. Both technical and non-technical users can benefit from understanding a system through its structural representation. The system, database, and interface design make up the architecture, which helps users see things more clearly. To make the system's components, features, and behaviors easier to grasp, the overall design of the system is described.

DOI: 10.17148/IJARCCE.2023.124202

The patient first registers by supplying a few parameters. When he went to check on his health, the collected values or data that has been stored in a database utilizing machine learning approaches, such as data gathering techniques employing some feature extraction approaches, data that has been stored in the database is extracted. When data is extracted, it goes through a number of steps before being used to predict an illness and create a report. This is an overview of the machine learning-based heart disease prediction system.

VII. CONCLUSION

One of the chronic fatal diseases that is rapidly spreading in both economically developed and underdeveloped nations and results in death is heart disease. If The patient is identified early on and given the appropriate care, this damage can be significantly diminished. In this work, we will design a machine learning-based intelligent predictive system for the precise and early prediction and detection of heart disease. The created method was tested using heart.csv datasets imported from Kaggle. We have outlined the various machine learning methods for heart disease prediction and the accuracy metric utilized for system performance evaluation. Here, we elaborated a number of machine learning algorithms and sought to choose the optimal algorithm by examining its aspects.

In this article, the algorithms KNN, SVM, DT, and LR are examined. The suggested system is GUI-based, approachable, scalable, trustworthy, and extendable. By offering early diagnosis, the proposed operating paradigm can also aid in lowering treatment costs.

This tool can be used by general practitioners to make the initial diagnosis of cardiac patients, speeding up the process and preventing unnecessary treatment delays. The scalability and accuracy of this prediction system can both be improved in a number of different ways. We have created a generalized framework that we may now utilize to analyze various data sets in the future.

REFERENCES

[1]. A. L. Bui, T. B. Horwich, and G. C. Fonarow, “Epidemiology and risk profile of heart failure,” Nature Reviews Cardiology, vol. 8, no. 1, p. 30,2011.
[2]. M. Durairaj and N. Ramasamy, “A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,” Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255–260, 2016.
[3]. L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, et al., “Decision making in advanced heart failure: a scientific statement from the american heart association,” Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.
[4]. S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, “Innovative artificial neural networks-based decision support system for heart diseases diagnosis,” 2013.
[5]. Q. K. Al-Shayea, “Artificial neural networks in medical diagnosis,” International Journal of Computer Science Issues, vol. 8, no. 2, pp. 150–154, 2011.
[6]. J. Lopez-Sendon, “The heart failure epidemic,” Medicographia, vol. 33, no. 4, pp. 363–9, 2011.
[7]. . P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, et al., “Forecasting the future of cardiovascular disease in the united states: a policy statement from the american heart association,” Circulation, vol. 123, no. 8, pp. 933–944, 2011.
[8]. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, “Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson’s disease symptom severity,” Journal of the royal society interface, vol. 8, no. 59, pp. 842–855, 2011.
[9]. S. I. Ansarullah and P. Kumar, “A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method,” International Journal of Recent Technology and Engineering, vol. 7, no. 6S, pp. 1009–15, 2019.
[10]. S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, “Fuzzy logic based decision support system for component security evaluation.,” Int. Arab J. Inf. Technol., vol. 15, no. 2, pp. 224–231, 2018.