International Journal of Advance Research in Science and Engineering
Volume No. 13, Issue No. 06, June 2024
www.ijarse.com
IJARSE
ISSN 2319 - 8354

A Novel Approach For Predicting The Heart Disease Using Machine Learning

D. Sai Naga Mahesh[(1)], B. Surendranath Singh[(2)] , I. Yadu Kondalu[(3)], B. Naveen[(4)]
MS. T. L. K. Prasanna M.Tech[(5)], MRS. Ch. Hema Sri M.Tech[(6)]
[(1),(2), (3),(4)] UG Students [(5),(6)] Associate Professor Department of Electronics and Communication Engineering,
Tirumala Engineering College, Palnadu (Dist), JNTU Kakinada, India,522601

Abstract: The healthcare industries of today generate massive volumes of data regarding patients, illnesses, and other relevant subjects. Data mining offers a plethora of techniques for revealing hidden patterns or similarities in data. Medical disorders can be diagnosed using these patterns. However, the raw medical data that are currently available are widely distributed, plentiful, and varied. This data must be gathered in an orderly fashion. A medical information system can be developed using this data once it has been collected. Data mining offers a user-friendly way to discover novel and hidden patterns in data. Data mining methods and tools are useful in the healthcare industry for predicting various illnesses and answering business-related questions. Disease prediction is crucial to data mining.

Keywords — *Cardiovascular Disease, Random Forest, SVM, Naïve Bayes*

I. INTRODUCTION

In addition, heart attacks and strokes account for 80% of mortality caused by CVDs. As a result, early detection of cardiac irregularities and tools for heart disease prediction can save many lives and assist medical professionals in creating treatment plans that are helpful in lowering the death rate from cardiovascular illnesses.

As a result of the advancement of sophisticated healthcare systems, a vast amount of patient data known as big data in electronic health record systems is currently accessible and can be utilized to create predictive models for cardiovascular illnesses. A discovery technique called data mining or machine learning is used to analyze large amounts of data from multiple angles and turn it into information that is helpful.

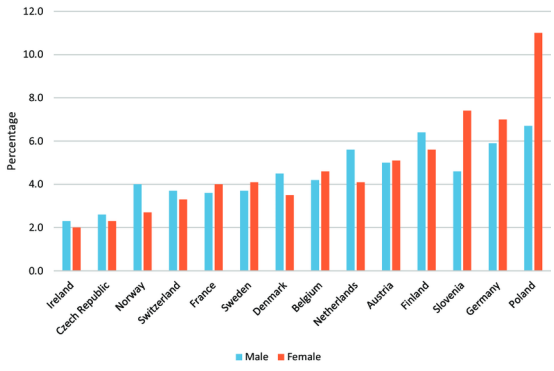


Fig.1: Cardiovacular disease of male and female in various countries

Today's healthcare sectors produce enormous amounts of data about patients, diseases, and other related topics. Numerous strategies are available in data mining to uncover hidden patterns or similarities in data. These patterns can be used to diagnose medical conditions. Nonetheless, the raw medical data that are currently accessible are dispersed broadly, abundant, and diverse. It is necessary to gather this data in an organized manner. After gathered, this data can be used to create a medical information system.

A user-oriented method for finding new and hidden patterns in data is provided by data mining. In the world of healthcare, data mining techniques and technologies are helpful in forecasting different diseases and providing answers to business queries. In data mining, disease prediction is important.

II. LITERATURE SURVEY

In 2018, Bo Jin and Chao Che presented a model called "Predicting the Risk of Heart Disease With EHR," which was created using artificial neural networks. This study analyzed and predicted heart illness using electronic health record data from real-world datasets pertaining to individuals' heart conditions. We put into practice a one-hot encryption model that uses victimization of heart failure episodes and diagnostics them; these are the fundamental ideas behind the neural network model's enlarged memory. We anticipated that examining the data would highlight how crucial it is to honor the natural world's outcomes as documented in the records[1].

Anjan Nikhil Repaka et al. created a system that employs the Advanced Encryption Standard (AES) algorithm for secure data transfer in disease prediction and NB (Naïve Bayesian) techniques for dataset classification[2].

Fast rule based cardiac disease prediction with associative approach is a model developed and applied by K. Prasanna Lakshmi and Dr. C.R.K. Reddy (2015). The author used the chi-square test to predict the disease using some associative techniques from the model[3].

Pure classifier association rule is the name given to the model that M. Satish and colleagues developed and applied to forecast the heart disease model using naïve

bayes and decision tree models. For this approach, he employed a data warehousing dataset on heart illness[4].

2018 saw the introduction of "Heart Disease Prediction using Evolutionary Rule Learning" by Aakash Chauhan. This study lessens the amount of manual labor, which also aids in directly extracting data (information) from electronic records. Using data mining on the patient's dataset, we must apply a certain frequency of pattern growth in order to derive this kind of rule. This will assess the situation, make an effort to cut costs for services, and demonstrate that most guidelines contribute to the most accurate forecast of heart disease[5].

The main purpose of this model's optimization is to use algorithmic grid search programs. This prediction of cardiovascular illness makes use of two different types of experiments. A random forest model is created and used to predict the model in the first form[6]. A random forest model based on the Random Search Algorithm is generated in the second form. Compared to traditional random forest models, this methodology is less complicated and more efficient. It achieves 3.3% greater accuracy than the random search algorithm when compared to the traditional random forest approach. The suggested learning approach can assist medical professionals in enhancing the accuracy of heart failure diagnosis[7][8].

A large amount of data about patients and their illnesses has been gathered by clinical databases. A set of records with health-related characteristics was acquired from the Cleveland Heart Disease database[8]. The dataset is used to extract the patterns important to the diagnosis of heart attack. The training dataset and testing dataset each received an equal portion of the records. A total of 76 medical attributes were retrieved from 303 records. Each and every attribute has a numerical value. We are focusing on a smaller collection of characteristics—just 14 in all[9][10][11].

III. DESIGN

The suggested work or process flow diagram is depicted in the image below. The Cleveland Heart Disease Database was first gathered from the UCI website. The information was then pre-processed, and 13 key features were chosen.

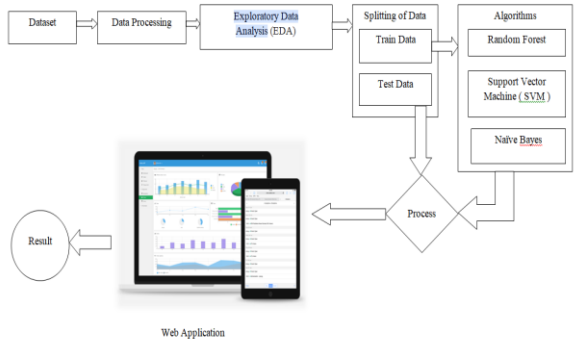


Fig.3: System Arcitecture

Using the Chi2 technique and the Recursive feature Elimination Algorithm, we were able to select 16 top features for feature selection. Next, use the logistic algorithm and ANN separately, then calculate the accuracy. Ultimately, we computed the optimal technique for diagnosing heart disease using the suggested Ensemble Voting approach.

Data Collection

The accuracy of classification metrics is heavily dependent on the quality of the dataset used for statistical predictions. For our research, we have picked the following datasets to both highlight the significance of the dataset and to assess its generalizability.

Table.1 Dataset

Attribute	Description	Domain of value
Age	Age in years	29 to 77
Sex	Sex	Male (1) Female (0)
Cp	Chest pain type	Typical angina (1) Atypical angina (2) Non-anginal (3) Asymptomatic (4)
Trestbps	Resting blood sugar	94 to 200 mm Hg
Chol	Serum cholesterol	126 to 564 mg/dl
Fbs	Fasting blood sugar	>120 mg/dl True (1) False (0)
Resteeg	Resting ECG result	Normal (0) ST-T wave abnormality (1) LV hypertrophy (2)
Thalach	Maximum heart rate achieved	71 to 202
Exang	Exercise induced angina	Yes (1) No (0)
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2
Slope	Slope of peak exercise ST segment	Upsloping (1) Flat (2) Downsloping (3)
Ca	Number of major vessels coloured by fluoroscopy	0–3
Thal	Defect type	Normal (3) Fixed defect (6) Reversible defect (7)
Num	Heart disease	0–4

This project's whole effort is broken down into four components. These consist of:

a. Data Pre-processing;

b. Capability

c. Classification and

d. Prediction

a . **Data Pre-processing**: All the pre-processing functions required to process all incoming texts and documents are contained in this file. We read the train, test, and

validation data files first, and then we carried out several preprocessing operations like stemming and tokenization. Some exploratory data analysis is carried out, such as distribution of answer variables and quality checks for null or missing values, among other things.

b. **Capability**: Extracting We used sci-kit learn Python modules for our feature extraction and selection processes in this file. We have employed simple bag-of-words and n-grams for feature selection, followed by term frequency techniques like tf-tdf weighting. Although they haven't been employed yet in the project, word2vec and POS tagging have also been utilized by us to extract the features.

c. **Classification**: All of the classifiers for the diagnosis of breast cancer diseases are constructed here. Several classifiers get the extracted features. We have utilized classifiers from sklearn, including Random Forest, Naive-Bayes, and SVM. Every retrieved feature was applied to every classifier. After the model was fitted, the confusion matrix was examined and the f1 score was compared.

Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression .But however ,it is mainly used for classification problems .As we know that a forest is made up of trees and more trees means more robust forest.

Similarly ,random forest creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting .It is ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result .

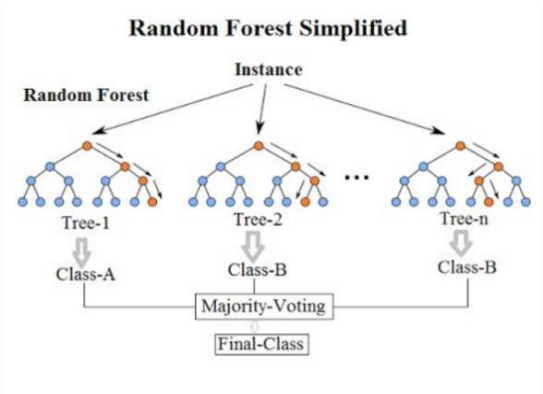


Fig.4: Random Forest

Support Vector Machine(SVM)

A supervised machine learning technique called Support Vector Machine (SVM) is utilized for regression as well as classification. Although we also state that regression issues are best suited for categorization. The SVM algorithm's primary goal is to locate the best hyperplane in an N-dimensional space that may be used to divide data points into various feature space classes.

The hyperplane attempts to maintain the largest feasible buffer between the nearest points of various classes. The number of features determines the hyperplane's dimension. A line represents the hyperplane when there are just two input characteristics. The hyperplane transforms into a 2-D plane if there are three input characteristics.

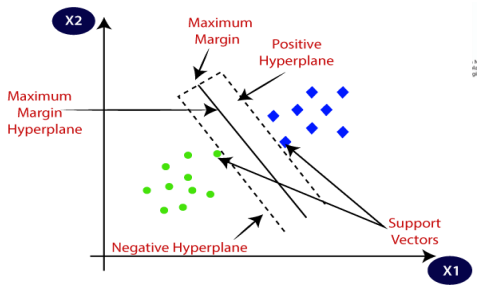


Fig.5: SVM

Naïve Bayes

A group of classification algorithms built on Bayes' Theorem are known as naive Bayes classifiers. Instead of being a single algorithm, it is a family of algorithms that are united by the idea that each pair of characteristics being categorized stands alone.

The Naïve Bayes classifier, one of the most straightforward and efficient classification algorithms, facilitates the quick creation of machine learning models with quick prediction skills.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A/B) - Conditional Probability of A given B
P(BIA) = Conditional Probability of A given B
P(A) = Probability of event A
P(B) = Probability of event A

d. **Prediction**

The method that we ultimately decided upon and found to be the best performer was stored to disk under the name final_model.sav. The prediction.py file will utilize this model to categorize heart conditions when you close this repository, and it will also be transferred to the user's computer.

The model is used to produce the final classification result, which is displayed to the user along with the likelihood of truth, after receiving a news story from the user.

Fig.6: Sequence Diagram

An example of a static structural diagram in the Unified Modeling Language (UML) is a class diagram, which shows the classes, attributes, operations (or methods), and interactions between objects in a system. It offers a fundamental notation for additional UML-recommended structural diagrams.

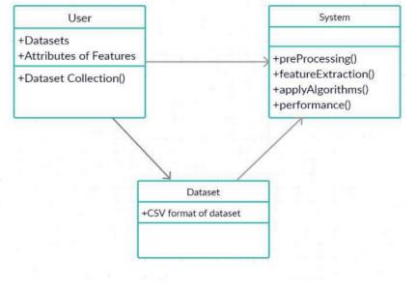


Fig.7: Class Diagram

IV. RESULTS AND DISCUSSION

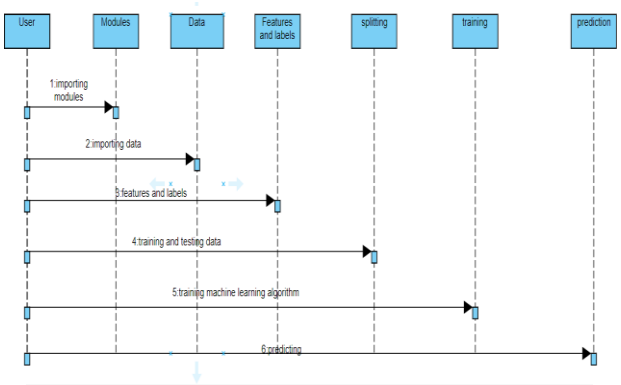
In order to choose the most significant and instructive characteristics for the purpose of creating predictive models, feature selection techniques are essential tools in data analysis and machine learning.

Table.1: Statistical analysis of dataset

```
[ ] dataset.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

The image below illustrates how the dataset information is enlarged based on categorical or numerical data.




```
[ ] dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Fig.8: Data Info

Based on a few factors, the person in the dataset under consideration is categorized as having heart disease or not.

```
[ ] print("Percentage of patience without heart problems: "+str(round(target_temp[0]*100/1025,2)))
    print("Percentage of patience with heart problems: "+str(round(target_temp[1]*100/1025,2)))

    #Alternatively,
    # print("Percentage of patience with heart problems: "+str(y.where(y==1).count()*100/1025))
    # print("Percentage of patience with heart problems: "+str(y.where(y==0).count()*100/1025))

    # #Or,
    # countNoDisease = len(df[df.target == 0])
    # countHaveDisease = len(df[df.target == 1])

Percentage of patience without heart problems: 48.68
Percentage of patience with heart problems: 51.32
```

Fig.9: Information of With heart disease and without heart disease

Once the correctness of each algorithm has been established through assessment, the machine learning algorithms are tested. The model is being trained, and after that training is complete, test data is sent.

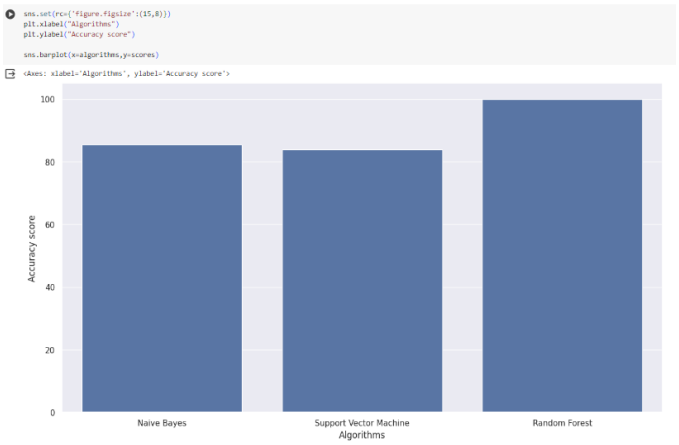


Fig.9: Algorithm Accuracy Score

Table.2: Algorithm Accuracy scores

Algorithm	Accuracy
K Nearest Neighbors	67.21%
Decision Tree	81.97%
Support Vector Machine	83.9%
Logistic Regression	85.25%
Naïve Bayes	85.37%
Random Forest	**100%**

Fig.10: Home Screen for Heart Disease Detection

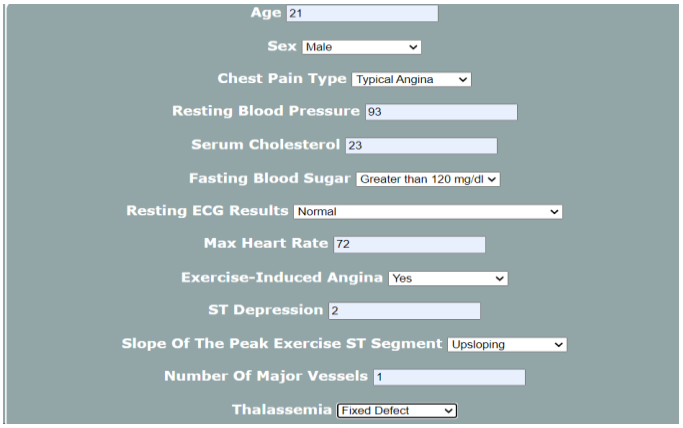


Fig.11: Patient Details

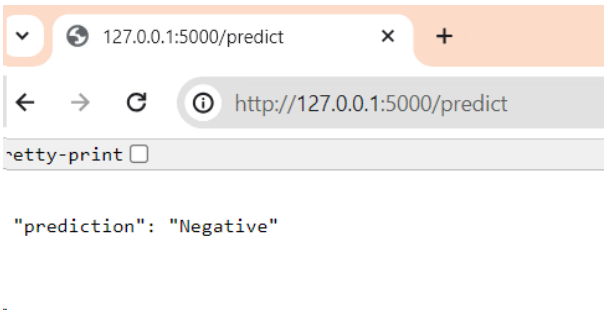


Fig.12: Output for particular patient details For Without Heart Disease

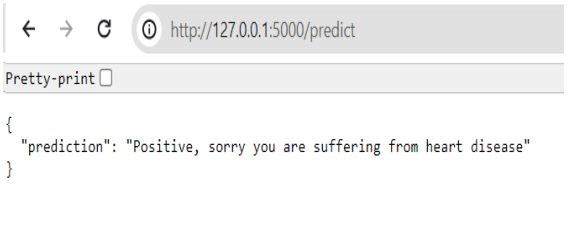


Fig.13: Output for particular patient details With Heart Disease

V. CONCLUSION

In this research, we provide a heart disease prediction system that uses various classifier algorithms to forecast the risk of heart disease. The methods include Random Forest, Support Vector Machines, and Naive Bayes. Our analysis shows that Random Forest outpe forms Random Forest in terms of accuracy. By eliminating extraneous and irrelevant features from the dataset and selecting just the most informative ones for the classification job, we hope to increase the Random Forest's performance.

Future Scope

As illustrated before the system can be used as a clinical assistant for any clinicians. The disease prediction through the risk factors can be hosted online and hence any internet users can access the system through a web browser and understand the risk of heart disease. The Future work may be combination of Embedded systems and Machine learning can be used from anywhere to see the paralyzed patients also.

REFERENCES

[1]. Tahseen Ullah , Syed Irfan Ullah, Khalil Ullah, Muhammad Ishaq, Ahmad Khan, Yazeed Yasin Ghadi, And Abdulmohsen Algarni , "Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection ",' in Proc. IEEE Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN), Jan 2024,DOI: 10.1109/ACCESS.2024.3359910

[2] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, ''An artificial intelligence model for heart disease detection using machine learning algorithms,'' Healthcare Anal., vol. 2, Nov. 2022, Art. no. 100016, doi: 10.1016/j.health.2022.100016.

[3] M. Ganesan and N. Sivakumar, ''IoT based heart disease prediction and diagnosis model for healthcare using machine learning models,'' in Proc. IEEE Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN), Mar. 2019, pp. 1–5, doi: 10.1109/ICSCAN.2019.8878850.

[4] D. P. Isravel, S. V. P. Darcini, and S. Silas, ''Improved heart disease diagnostic IoT model using machine learning techniques,'' Int. J. Sci. Technol. Res., vol. 9, no. 2, pp. 4442–4446, 2020.

[5] I. S. G. Brites, L. M. da Silva, J. L. V. Barbosa, S. J. Rigo, S. D. Correia, and V. R. Q. Leithardt, ''Machine learning and IoT applied to cardiovascular diseases identification through heart sounds: A literature review,'' Informatics, vol. 8, no. 4, p. 73, Oct. 2021, doi: 10.3390/informatics8040073.

[6] D. T. Thai, Q. T. Minh, and P. H. Phung, ''Toward an IoT-based expert system for heart disease diagnosis,'' in Proc. 28th Mod. Artif. Intell. Cogn. Sci. Conf. (MAICS), 2017, pp. 157–164.

[7] B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, and E. K. R. Patro, ''Early and accurate prediction of heart disease using machine learning model,'' Turkish J. Comput. Math. Educ., vol. 12, no. 6, pp. 4516–4528, 2021.

[8] S. Anitha and N. Sridevi, Heart Disease Prediction Using Data Mining Techniques S Anitha, N Sridevi to Cite This Version, document HAL Id Hal02196156, 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02196156/document

[9] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, ''Prediction of heart disease using a combination of machine learning and deep learning,'' Comput. Intell. Neurosci., vol. 2021, pp. 1–11, Jul. 2021, doi: 10.1155/2021/8387680.

[10] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, ''Heart disease prediction using machine learning algorithms,'' IOP Conf., Mater. Sci. Eng., vol. 1022, no. 1, Jan. 2021, Art. no. 012072, doi: 10.1088/1757- 899x/1022/1/012072.

[11] B. Pavithra and V. Rajalakshmi, ''Heart disease detection using machine learning algorithms,'' in Proc. Int. Conf. Emerg. Current Trends Comput. Expert Technol., vol. 35, 2020, pp. 1131–1137, doi: 10.1007/978-3-030- 32150-5_115

[12]. N. Louridi, S. Douzi, and B. El Ouahidi, ''Machine learning-based identification of patients with a cardiovascular defect,'' J. Big Data, vol. 8, no. 1, pp. 1–5, Dec. 2021, doi: 10.1186/s40537-021-00524-9.

[13] P. Singh, G. K. Pal, and S. Gangwar, ''Prediction of cardiovascular disease using feature selection techniques,'' Int. J. Comput. Theory Eng., vol. 14, no. 3, pp. 97–103, 2022, doi: 10.7763/ijcte.2022.v14.1316.

[14] M. Swathy and K. Saruladha, ''A comparative study of classification and prediction of cardio-vascular diseases (CVD) using machine learning and deep learning techniques,'' ICT Exp., vol. 8, no. 1, pp. 109–116, Mar. 2022, doi: 10.1016/j.icte.2021.08.021.

[15] D. Vaddella, C. Sruthi, B. K. Chowdary, and S.-R. Subbareddy, ''Prediction of heart disease using machine learning techniques,'' Restaur. Bus., vol. 118, no. 1, pp. 125–129, 2019, doi: 10.26643/rb.v118i1.7621.

[16] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, ''Heart disease prediction using machine learning techniques: A survey,'' Int. J. Eng. Technol., vol. 7, no. 2, p. 684, Mar. 2018, doi: 10.14419/ijet.v7i2.8.10557.

[17] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, ''Heart disease identification method using machine learning classification in E-healthcare,'' IEEE Access, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[18] P. Kalpana, S. S. Vignesh, L. M. P. Surya, and V. V. Prasad, ''Prediction of heart disease using

machine learning,'' J. Phys., Conf. Ser., vol. 1916, no. 1, May 2021, Art. no. 012022, doi: 10.1088/1742-6596/1916/1/012022.

[19] A. Ed-Daoudy and K. Maalmi, ''Real-time machine learning for early detection of heart disease using big data approach,'' in Proc. Int. Conf. Wireless Technol., Embedded Intell. Syst. (WITS), Apr. 2019, pp. 1–5, doi: 10.1109/WITS.2019.8723839.

[20] I. D. Mienye, Y. Sun, and Z. Wang, ''An improved ensemble learning approach for the prediction of heart disease risk,'' Informat. Med. Unlocked, vol. 20, Jan. 2020, Art. no. 100402, doi: 10.1016/j.imu.2020.100402.

[21] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, ''MIFH: A machine intelligence framework for heart disease diagnosis,'' IEEE Access, vol. 8, pp. 14659–14674, 2020, doi: 10.1109/ACCESS.2019.2962755.

[22] R. Atallah and A. Al-Mousa, ''Heart disease detection using machine learning majority voting ensemble method,'' in Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS), Oct. 2019, pp. 1–6, doi: 10.1109/ICTCS.2019.8923053.

[23] M. Bheemalingaiah, G. R. Swamy, P. Vishvapathi, P. V. Babu, E. N. Rao, and P. N. Rao, ''Detection of heart disease by using reliable Boolean machine learning algorithm,'' J. Theor. Appl. Inf. Technol., vol. 99, no. 15, pp. 3856–3880, 2021, doi: 10.5281/zenodo.5353586.

[24] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, ''Risk prediction of cardiovascular disease using machine learning classifiers,'' Open Med., vol. 17, no. 1, pp. 1100–1113, Jun. 2022.

[25] S. I. Ayon, M. M. Islam, and M. R. Hossain, ''Coronary artery heart disease prediction: A comparative study of computational intelligence techniques,'' IETE J. Res., vol. 68, no. 4, pp. 2488–2507, Jul. 2022, doi: 10.1080/03772063.2020.1713916.

About Authors

Ms.T.L.k. Prassana Currently working as Assistant Professor in Tirumala Engineering College. She received her M.Tech (Computers & Communications) Degree from Jawaharlal Nehru Technological University Hyderabad. She is a life member of technical association in IETE.

D. Sai Naga Mahesh currently studying B.Tech (zz Engineering) in Sai Tirumala NVR Engineering College, State Board of Technical Education and Training , Andhra Pradesh in 2021. He is a life member of Technical Association IETE.

B. Surendranth Singh currently studying B.Tech (Electronics & Communication Engineering) in Tirumala Engineering College, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh in the year 2024. He completed his Diploma (Electronics & Communication Engineering) in Sai Tirumala NVR Engineering College, State Board of Technical Education and Training , Andhra Pradesh in 2021. He is a life member of Technical Association IETE.

I. YaduKondalu currently studying B.Tech (Electronics & Communication Engineering) in Tirumala Engineering College, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh in the year 2024. He completed his Intermediate in Narayana Junior College in 2020. He is a life member of Technical Association ISTE.

B. Naveen currently studying B.Tech (Electronics & Communication Engineering) in Tirumala Engineering College, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh in the year 2024. He completed his Intermediate in Narayana Junior College in 2020. He is a life member of Technical Association ISTE.