International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Cardiovascular Syndrome Prediction Using Machine Learning Algorithms

[1]Dr. K. Sreenivasulu, [2]Ms. B. Anuradha, [3]Dr. A. Chandra Obula Reddy, [4]Dr. Vikram Neerugatti, [5]Appawala Jayanthi, [*6]K. K. Baseer, [7]Dharmesh Dhabliya

Submitted: 01/12/2023 **Revised**: 25/01/2024 **Accepted**: 03/02/2024

Abstract—: Cardiovascular disease can be caused by a variety of factors. Researchers can predict cardiovascular infirmity using a variety of methods, regardless of whether a person has the condition or not. The heart disease has been placed via extracting significant qualities and most relevant features using a variety of research methods, such as pulse, cholesterol levels, and other symptoms. The major goal of the study is to use the data to forecast whether the person has a cardiovascular condition. As a result, data mining is employed, which makes it simple to analyse the data collection. Null values and duplicate values are eliminated. The data is subjected to regression analyses utilising decision trees with party and rpart, random forests, linear regression, and logistic regression. The data set is trained and tested using regressions. The regressions are compared, and the outcome for the data set is reliable. All comparisons within the data set are then made using the regression. Therefore, the findings indicate whether or not the individual will eventually develop cardiovascular disease.

Keywords— *cardiovascular disease, Decision Tree, Data Mining , rpart Random Forest, Linear and Logistic Regressions.*

I. Introduction

Stroke occurs when the flow of blood to the brain is interrupted and can lead to cardiovascular disease. Numerous variables, including smoking, obesity, an unhealthy lifestyle, physical activity, high cholesterol, and binge drinking, can contribute to cardiovascular disease. Fatigue, chest pain, and heartburn are examples of cardio vascular symptoms. These are a few cardiovascular disease symptoms. Several data mining methods may be employed to predict the onset of cardiovascular syndrome.

Machine learning is similar to human learning systems in how it uses input data or information. Machine learning uses cutting edge approaches to alter and retrieve features or important data. Machine learning employs historical information or facts to find original patterns and apply algorithms to generate useful outcomes [1]. The primary focus of the proposed project is a decision support system for the identification of heart disease. After characteristics engineering and preprocessing techniques have been applied, machine learning techniques like random forest, decision trees, gradient boosted trees, linear support vector classifier, logistic regression, one-vs-rest, and multilayer perception are used to perform binary and multi classification on the data stream [2]. Heart disease identification needs to be exact and accurate in order to prevent human loss. While other

[1]*Professor*
Department of Computer Science and Engineering
G.Pullaiah College of Engineering and Technology, Kurnool AP INDIA
1sreenu.kutala@gmail.com
[2]*Assistant Professor*
Department of Electronics and Communication Engineering
Sri Eshwar College of Engineering
Coimbatore
2saianu08@gmail.com
[3]*Associate Professor*
Department of CSE (AI & ML)
Sri Venkateswara College of Engineering. Kadapa. AP, INDIA
3acoreddy@gmail.com
[4]*Associate Professor*
Department of CSE,
Faculty of Engineering and Technology
Jain (Deemed - to - be) University, Bangalore, Karnataka.
4vikram.n@jainuniversity.ac.in
[5]*Department of Computer Science and Engineering,*
Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India.
5jayanthi.a33@gmail.com
[6]*Associate Professor of CSE,*
GITAM School of Technology,
GITAM (Deemed to be University),
Bengaluru, Karnataka, INDIA
**6drkkbaseer@gmail.com*
[7]*Department of Information Technology*
Vishwakarma Institute of Information Technology,
Pune, India
7Dharmesh.dhabliya@viit.ac.in
**Corresponding Author: *6drkkbaseer@gmail.com*

approaches are quick but accurate, previous study studies have a number of disadvantages, including the fact that they take a long time to compute [3].

Researchers seeks for Data correlations that could significantly increase the accuracy of prediction rates when utilizing different machine learning models in order to identify cardiac sickness using past medical data [4]. Understanding the facts of heart disease is essential to improving prediction accuracy. This study [5] has used an investigational evaluation to assess the effectiveness of models constructed using categorization algorithms and relevant attributes selected using various feature selection procedures.

The data sets must be cleaned up and all null values deleted. Next, regression is performed using the cleaned data. DT, RF, Linear, Logistic, and other types of regression are employed. Different formulas are wheedled to calculate the regressions. The data sets are utilised for training and testing the regressions after being imported into R studio. These regressions are employed for regression analysis and comparison. Regressions are trained and evaluated against one another. The data set is examined using the more effective regression.

The suggested approach is intended to determine whether or not a person is exhibiting signs of heart disease. We took into account earlier predictions made and created a new architecture that will outperform the old models. Figure 1 shows the architecture, which is made up of the repository for heart disease and its symptoms. The data preprocessor verifies that the data are in the correct order and fills in any gaps in the data if necessary. When the data are examined for analysis, they are sent to the feature selection stage, where several techniques like filtering, wrapping, and embedding are used to reduce the features [5].The subsequent phase of data division involves training and testing data. On the data, the fusion method is employed. Based on the information provided by the model's fusion algorithm, the suggested model assists in determining whether the data contains symptoms of heart disease or not [6].

Cardiovascular syndrome prediction through machine learning involves a systematic approach to leveraging data for accurate risk assessment. Initial steps encompass the collection of diverse datasets, including patient demographics, medical histories, and clinical measurements. Following this, preprocessing techniques are applied to handle missing data, normalize numerical features, and encode categorical variables. Feature selection becomes pivotal in identifying relevant factors affecting predictions. Different AI calculations, for example, strategic relapse, support vector machines, arbitrary woods, and brain organizations, are then utilized, with preparing including the advancement of hyper boundaries and thorough assessment through measurements like exactness and accuracy. Interpretability of the model is achieved through techniques like SHAP values or LIME. Rigorous validation and testing on separate datasets ensure the model's generalizability, and subsequent deployment requires attention to security and privacy in handling healthcare data. Continuous monitoring and updates, alongside collaboration with healthcare professionals, ensure the model's ongoing accuracy and ethical use in predicting cardiovascular syndromes.

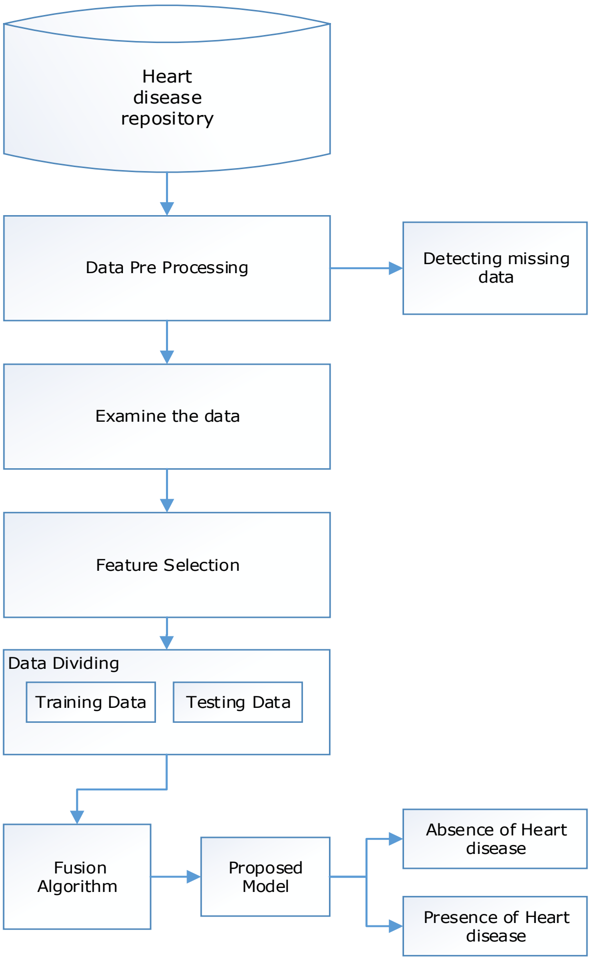

Fig 1. Architecture for predicting the Heart syndrome

The order of the sections in the manuscript is as follows: The literature review conducted on cardiovascular disease is covered in part II. The methods used to collect data from diverse domains and organize it are described in Section III. In section IV, algorithms are put into practice using data sets of different symptoms, and section V presents the results as graphs of different algorithms over the data sets. Section VI explains the conclusion and upcoming work.

II. **Literature Survey**

Cardio-vascular syndrome prediction using machine learning requires an in-depth analysis of existing research articles, papers, and studies. Researchers have increasingly explored the application of machine learning algorithms in predicting cardiovascular syndromes, aiming to enhance early diagnosis and risk assessment.

Several studies have focused on the utilization of diverse datasets, incorporating patient information, lifestyle factors, and clinical measurements. Researchers often emphasize the importance of data preprocessing techniques to address issues like missing data, normalization, and feature encoding. Feature selection methodologies have been a subject of investigation, with researchers exploring the most effective ways to identify key predictors of cardiovascular risk.

There have been a number of different machine learning algorithms used, and which one is used most often is determined by the nature of the prediction task and the particulars of the dataset. Calculated relapse, support vector machines, irregular timberlands, and brain networks are among the normally utilized calculations, with specialists ceaselessly refining and looking at their exhibition.

Interpretability has emerged as a critical aspect of these models, and studies have proposed methods such as SHAP values and LIME to make machine learning models more transparent and understandable for healthcare professionals.

Validation and testing procedures have been a focus in the literature, with researchers addressing the importance of robust evaluation metrics and techniques like cross-validation. The deployment of models in real-world healthcare settings has also been discussed, highlighting the need for ethical considerations and privacy safeguards.

Continuous monitoring and updates for model improvement, as well as collaborations with healthcare professionals, have been emphasized in the literature to ensure the practicality, accuracy, and ethical use of machine learning models in predicting cardiovascular syndromes. Overall, the literature reflects a dynamic field with ongoing efforts to refine methodologies and contribute to the advancement of cardiovascular risk prediction using machine learning.

Various strategies are used for finding the symptoms of cardiovascular disease are discussed in [7]. In order to obtain accurate results, this research utilizes various data mining techniques such classification and clustering learning methods to determine the causes of cardio vascular disease. The unsupervised learning method is used to evaluate the outcomes: Clustering using K-means. The primary risk factors for cardiovascular syndrome, such as smoking, alcohol consumption, high cholesterol, age, and weight, were their initial emphasis.

Without considering the patient's medical history, [8] developed a way to predict which patients are more likely to develop cardiac syndrome. The main emphasis is on identifying patients who are more likely to comply based on various medical criteria. Machine learning methods like logistic regressions and KNN were used to forecast and classify the heart disease patient. When compared to Naive bayes, this logistic and KNN study revealed correct results.

The national institutes of health have published statistics regarding heart disease with reports that include key health behaviors including smoking, physical activity, food, age, and weight, which were described by the author in [9]. They have made a statistics report about cardiovascular health available.

In [10] discussed about heart disease leading the severe health issues like kidneys failure, diabetes, and lungs failure. There are so many factors for causing the heart disease for the people consuming the tobacco, regular of drinking alcohol, overweight, breathing problem, chest pain. And lack of physical activity this they have prescribed about the medication.

In [11] divides the population into four separate age groups and details the statistics showing which age groups are more frequently impacted by heart attacks. The Bayesian statistical model was employed. People who are between the ages of 35 and 44 are said to be more susceptible to heart disease.

In [12] offers a survey on the medical background, diet, and health of people with heart disease. For the health measures, they have utilized machine learning algorithms like regressions.

In [13] it covers many aspects of the cardiovascular infirmity they have use the supervised learning techniques such as DT, RF, and Logistic Regression. They have used the uci data to forecast whether the individuals are getting the cardiovascular disease.

In [14] author has discussed about heart disease and very serious health issues caused by different lifestyle and food habits. They have detected the factors causing heart disease. They used the algorithms to calculate its enactment using the metrics for evaluation have trained the data and validated the machine learning.

In [15] explains heart disease can be anticipated utilizing ML algorithms like Random Forest, SVM, Navies Bayes and the Decision Tree are used. Suggests the best algorithm to predict the heart disease.

In [16] author uses the machine learning because it uses the accurate results and quick diagnosis of disease. They further classifies for the analysis of the data for analysis and performance compared to others. In this they used the algorithm DNN.

In [17] authors used a method for drug-target relationship prediction based on groups by Appling Bayesian, helps in providing the better medication for the cardiovascular diseases.

III. Model Planning

In model planning it gives the basic idea of how to gather the data from the patients using the devices and arrange them in the database based on the data sets and domains. The steps in data discovery for the syndrome are as follows.

a. Data Discovery

We have diseases, health-related difficulties, and health-care practices like physical health care, diet, and exercise in this domain. The resources for causing cardiovascular disease depends on the factors like age, weight, cholesterol, smoke, alcohol, ap_hi, ap_lo, height, Blood Pressure etc., The problem statement is to predict which algorithm gives the best accurate for predicting cardiovascular disease depends on the factors like age, weight, cholesterol. Caretakers of the patient, healthcare organizations, government officials of the health care industry, physicians, and patients are some of our major stakeholders. The problem for cardiovascular syndrome. What is the involvement of the team in predicting cardio vascular disease? What are the changes made for predicting cardiovascular disease? What features mainly need to be considered? What are the major factors and solutions for cardiovascular disease? Using classification and regression algorithms, to predict cardiovascular disease. To predict a better model and Review the raw data about cardio vascular disease. Once the data discovery is done need to preprocess the data in the following manner.

b. Data Preprocessing

The term "data preprocessing," which is a component of "data preparation," refers to any kind of processing that is carried out on the raw data to prepare it for a subsequent step in the data processing process. It has generally been an essential stage in the information mining process.

The first step involves establishing a methodical sandbox for organized experimentation. ETLT (Extract, Transform, Load, Transform) processes are diligently executed to manage data flows effectively. Data exploration is undertaken through in-depth study to gain insights into its characteristics and patterns. Rigorous information conditioning is applied to enhance the quality and relevance of the data.

A comprehensive survey and imaginative analysis contribute to refining strategies and envisioning potential insights. The initial phase of the data analysis process involves acquiring the dataset, a crucial step that sets the foundation for subsequent exploration. Once obtained, the dataset undergoes meticulous scrutiny for missing data values, necessitating a thorough identification and handling process to ensure data integrity. Encoding the dataset follows, transforming categorical variables into a format suitable for analysis. The dataset is then judiciously split, creating subsets for training and testing purposes. Importing the dataset into the R environment facilitates seamless integration for further analysis. In this R environment, the data undergoes comprehensive scrutiny, and the work process unfolds as various statistical and machine learning techniques are applied to glean meaningful insights and draw informed conclusions. This structured approach from dataset acquisition to analysis within the R environment ensures a systematic and thorough exploration of the data, laying the groundwork for robust results. Organizing methodical sandbox.

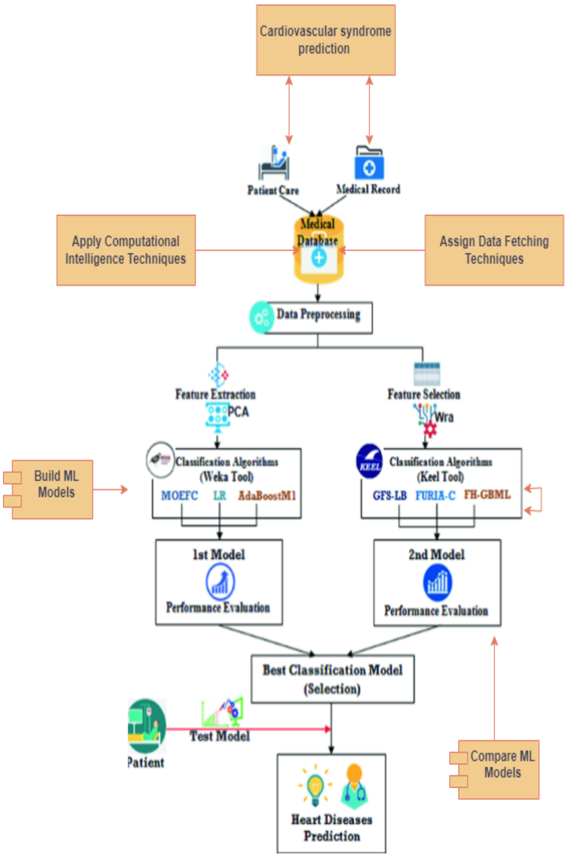

Fig 2. Model Planning for predicting the Heart syndrome

To train and analyze the data, the dataset is imported and put into the R environment. The information is gathered, taken from the cardio vascular dataset, and handled with regard to missing information.

The cardiovascular dataset contains 1000 rows and 13 columns that are related medical health care features. Such as age, weight blood pressure, cholesterol, smoke, alcohol. The cardio vascular dataset is cleaned and transformed into R environment as per the formula given to generate. The survey can be done by analyzing the dataset and review the data if any negative values or unwanted data present, and then removed from the dataset. The work process had done by checking what values need to be visualizing by taking active for cardiovascular dataset.

c. Model Building

In supervised learning algorithm Decision Tree is used. In the decision tree we have the root node of the and further classified into sub tree as the child nodes. In the decision tree we use to split the facts in to the train facts and the test facts. In the decision tree child nodes are also called as the leaf nodes or terminal nodes .entire tree is called branch or sub tree. Nodes are divided into two or more sub nodes.

Random forest is similar to that of the decision tree. Random forest is machine learning algorithm. It is used for the both classification and regression problems. Random forest contains number of decision trees and uses the predictive accuracy of the dataset. In the decision tree we have the greater number of trees. As uses the training set and have different decision tree and predict [18-19]. Linear Regression is supervised learning model.it is one of the easiest and it depends on the individual variables. Linear Regression is the predictive analysis. Linear Regression makes predictions for continuous variables. It shows the relationship between a dependent and independent variables.

Logistic Regression is supervised learning model. It uses for predicting the categorical dependent variable. It predicts the output of a categorical dependent variable. uses the exact value 0 and 1.probabilistic values lies between 0 and 1.

Calculated Relapse is a measurable technique broadly utilized for twofold characterization undertakings, filling in as a primary device in the domain of AI. Logistic regression, in contrast to linear regression, which predicts continuous outcomes, is made to predict whether or not an event will occur. When the dependent variable is a binary one, such as presence or absence, success or failure, this method of predictive modeling excels. A linear combination of input features is transformed by logistic regression into a probability score between 0 and 1. The model provides an estimate of the likelihood that an observation belongs to a particular class, making it a useful option in a variety of fields, such as healthcare for the prediction of disease or finance for credit scoring. Its simplicity, interpretability, and efficiency contribute to the widespread adoption of Logistic Regression, making it an essential tool for predictive modeling and decision-making across diverse domains.

IV. Implementation and Results

In Implementation based on the models data is processed in the following manner by training the data sets then testing it. Develop the datasets for the training and testing. Develop the analytical model on teaching facts and test on test testing facts.

a) **Decision Tree:** In the decision tree we have used the r part and party packages. The parameters from data set used in the decision tree are in age, gender, gluc, smoke ,ap_hi, ap_lo, alco ,alcohol in the formula.

b) **Random Forest:**

In the random forest we have used the package named random forest from the library of r- studio; in this the parameters are id and plotted.

c) **Linear Regression:**

In the Linear Regression, we have used the parameters age and height.

d) **Logistic Regression:**

In the logistic regression we have used the parameters like age, gender, p_lo, ap_hi, alco, gluc, and smoke, active. In the logistic Regression we have used family gaussain ("log").

The parameters are tabulated in table 1.

Table 1: Performance comparison for various models

Models	Accuracy	Specificity	Sensitivity
Decision Tree	85	84	83
Regression	89	87	88
Random forest	87	86	87
Proposed Model	93	87	85

By the evaluation of various models it concludes the accuracy of the proposed model is better than other models. In this paper we have tried to provide a model where the predicting of the cardiovascular disease is done based on the health condition of the patients. The following graphs shows the results after applying the data sets to calculate the performance of the algorithm in predicting the diseases .here the p-value is considered as the normal reading values of the health person.

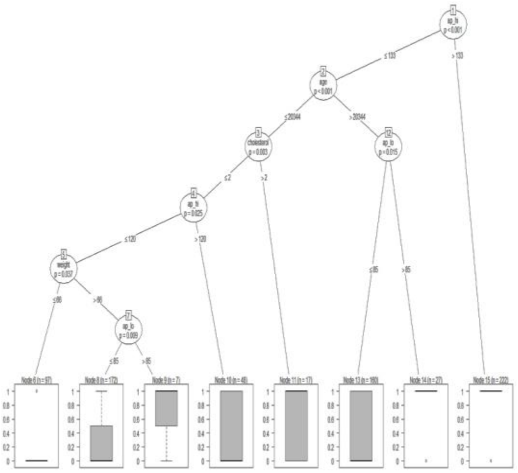

Fig 3. Decision tree for heart disease categorize symptoms

The figure 2 & 3 explains the p-values of the person which helps in taking the decision for the predicting of the disease. It also helps in categorize the symptoms of the disease caused the various factors like blood pressure, smoking, alcohol and cholesterol levels in the blood.

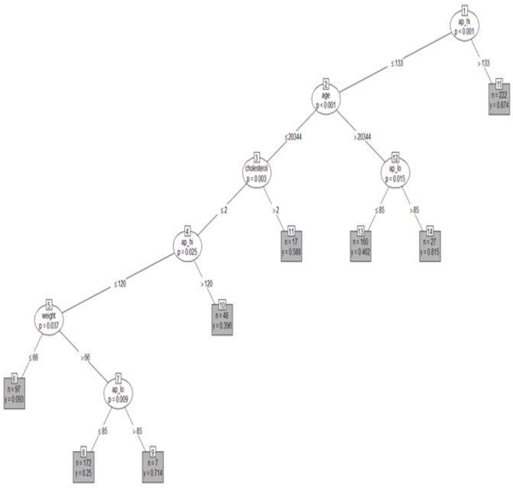

Fig 4. Decision tree for categorize with and without heart disease symptoms

In Figure 4 it gives the relation between the observed and the predicted values of the cardiovascular disease. The graph in figure 5 show the errors that occur doing the calculating the values and noting the nearest values for better prediction. In figure 6 comparisons of regression algorithm results which help in predicting the disease is shown.

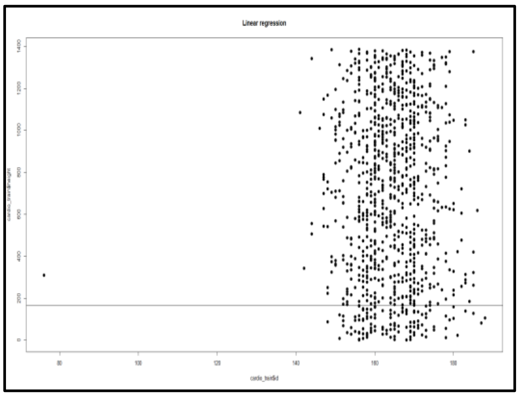

Fig 5. Linear Regression for prediction the chances of disease

The comparative analysis of the models reveals that the Proposed Model outperforms others with an impressive accuracy of 93%, showcasing its superior predictive capability. The model also maintains a balanced performance in specificity (87%) and sensitivity (85%), indicating its effectiveness in correctly identifying both true negatives and true positives. Regression follows closely with an accuracy of 89% and comparable specificity and sensitivity values (87% and 88%, respectively). Random Forest demonstrates balanced performance, but Decision Tree lags behind in sensitivity with the lowest value at 83%. Overall, the Proposed Model emerges as the top-performing model, making it a compelling choice for its robust predictive accuracy and well-balanced classification metrics.

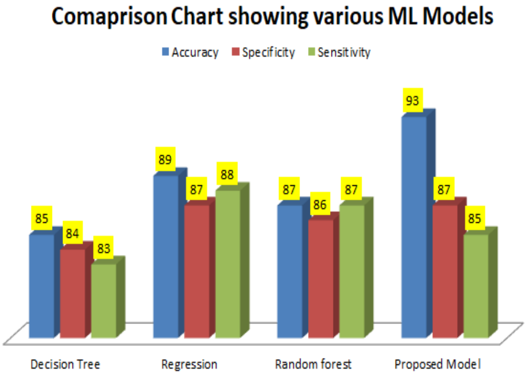

Fig 6 .Comparison chart for Prediction of cardiovascular values

V. Conclusion And Future Work

The classification modeling approaches and modeling tools used in this cardiovascular disease model are used. We used methods like logistic regression, linear regression, random forests, and decision trees. Based on traits including age, gender, smoking, and physical activity, this model forecasts the likelihood that a person will have cardiovascular disease. Many medical databases employ these machine learning techniques to monitor patient health because they can predict outcomes more accurately than people. With the use unsupervised learning the suggested model over performance is linear regression, which offers the highest level of accuracy out of all of these techniques.

References

[1] Abdul Saboor, Muhammad Usman, Sikandar Ali, Ali Samad, Muhmmad Faisal Abrar, Najeeb Ullah, "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol. 2022, Article ID 1410169, 9 pages, 2022.

[2] Danish Hamid, Syed Sajid Ullah, Jawaid Iqbal, Saddam Hussain, Ch. Anwar ul Hassan, Fazlullah Umar, "A Machine Learning in Binary and Multi classification Results on Imbalanced Heart Disease Data Stream", Journal of Sensors, vol. 2022, Article ID 8400622, 13 pages, 2022.

[3] Farhat Ullah, Xin Chen, Khairan Rajab, Mana Saleh Al Reshan, Asadullah Shaikh, Muhammad Abul Hassan, Muhammad Rizwan, Monika Davidekova, "An Efficient Machine Learning Model Based on Improved Features Selections for Early and Accurate Heart Disease Predication", Computational Intelligence and Neuroscience, vol. 2022, Article ID 1906466, 12 pages, 2022.

[4] K. S. Archana, B. Sivakumar, Ramya Kuppusamy, Yuvaraja Teekaraman, Arun Radhakrishnan, "Automated Cardio ailment Identification and Prevention by Hybrid Machine Learning Models", Computational and Mathematical Methods in Medicine, vol. 2022, Article ID 9797844, 8 pages, 2022.

[5] Kaushalya Dissanayake, Md Gapar Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 5581806, 17 pages, 2021.

[6] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine

Learning Algorithms", Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018. https://doi.org/10.1155/2018/3860146.

[7] Avijit Chaudhuri et.al , “EarlyPrediction of Heart Disease Using the Most Significant Features of Diabetes by Machine Learning Techniques” . drafted on,may-2021, https://www.researchgate.net/publication/351437170_E.

[8] Harshit Jindal et al,” Heart disease predictin using machine learning algorithms”, 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072 , Doi: 10.1088/1757-899X/1022/1/012072.

[9] Salim S.Virani et.al,” Heart Disease and Stroke Statistics—2021 Update”, published on 27 jan 2021, https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000000950.

[10] K. K. Baseer, S. B. A. Nas, S. Dharani, S. Sravani, P. Yashwanth and P. Jyothirmai, "Medical Diagnosis of Human Heart Diseases with and without Hyperparameter tuning through Machine Learning," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India,2023,pp.1-8,doi: 10.1109/ICCMC56507.2023.10084156.

[11] Adam S. Vaughan, published, Widespread recent increases in county-level heart disease mortality across age groups”. on 27 dec 2017.

[12] Ramal Moone singhe, Muin j Khoury ,published on July 2019 Prevalence and Cardiovascular Health Impact of Family History of Premature Heart Disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007–2014.

[13] Mohammad Monirujjaman Khan ,Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction, published on 10 dec 2021.

[14] K. K. Baseer, Dr M Jahir Pasha, Telkapalli Murali Krishna, Jeribanda Mohan Kumar, Silpa C, “COVID-19 Patient Count Prediction using Classification Algorithm”, International Journal of Early Childhood Special Education (INT-JECSE), DOI:10.9756/INTJECSE/V14I7.7 ISSN: 1308-5581 Vol 14, Issue 07, 2022.

[15] Vijeta Sharma; ShrinkhalaYadav; Manjari Gupta, “Heart Disease Prediction using Machine learning Techniques“, published on 01 march 2021 https://ieeexplore.ieee.org/document/9362842.

[16] vineet Sharma; Akhtar Rasool; Gaurav Hajela; “Prediction of Heart disease using DNN”, published on 01-september-2020. https://ieeexplore.ieee.org/document/9182991.

[17] Manmohan Singh et.al, “A Drug-Target Interaction Prediction Based on Supervised Probabilistic Classification”, Journal of Computer Science, 19(10), 1203-1211.Sep-2023, https://doi.org/10.3844/jcssp.2023.1203.1211.

[18] Abdul Saboor; Muhammad Usman; Sikandar Ali, Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. published on march 2022.

[19] Silpa C, Dr. S Srinivasa Chakravarthi, Jagadeesh kumar G, Dr. K.K. Baseer, E. Sandhya, “Health Monitoring System Using IoT Sensors”, Journal of Algebraic Statistics, Volume 13, No. 3, June, 2022, p. 3051-3056, ISSN: 1309-3452.