JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Ensemble Approach to Predict the Presence of Cardio Vascular Disease using Machine Learning and Deep Learning.

Prasad Vadamodula
GMR Institute of Technology, Rajam, Srikakulam

B Sai Satwika
GMR Institute of Technology, Rajam, Srikakulam

A Swamy
GMR Institute of Technology, Rajam, Srikakulam

D Pranitha
GMR Institute of Technology, Rajam, Srikakulam

G Pavan Kalyan
GMR Institute of Technology, Rajam, Srikakulam

C Mohan Rao
GMR Institute of Technology, Rajam, Srikakulam

***Abstract*— The heart muscle is injured when blood flow to the coronary artery is reduced or disrupted, resulting in coronary artery disease, often known as a heart attack. Without human assistance, hidden patterns can be found using machine learning. The suggested methodology intends to create an intelligent agent to detect any coronary heart disease well in advance of any unfavourable event. A dataset including around 70000 records consisting of 11 features is used. The model applies different feature selection techniques such as Pearson correlation & Information Gain Attribute Evaluator. This model is aimed to construct an ensemble method with classification algorithms- Naïve Bayes , Random Forest and Gradient Boosting. The presence or absence of the cardio vascular illness from the characteristics is also determined using the K Nearest Neighbor (KNN) and Support Vector Machine algorithms. AUC is a representation of separability's level or measurement. It demonstrates how well the model can distinguish between classes. The model's accuracy will be evaluated and enhanced.**

Keywords— Heart attacks, Cardio vascular diseases, Ensemble model, Classification, Pearson correlation.

I. INTRODUCTION

The most serious health problem is heart disease, which has affected many people all around the world. Shortness of breath, tightness and pressure in the chest, discomfort (angina), pain in the chest, and numbness, weakness, or coldness in the blood arteries in the limbs or legs being narrowed are just a few symptoms of cardiovascular disease. Additional causes of cardiovascular disorders include damage to the coronary arteries, the entire heart or just a section of it, or inadequate nutrient and oxygen delivery to the heart. When cardiovascular illness is still in its early stages, a machine learning model can help doctors choose the best treatment. This could cause death if it is not detected in time. A crucial first step in primary prevention is risk assessment.

Heart conditions can in a variety of forms, including:

Coronary Artery Disease: This particular type of heart disease is common and well-known. This disorder causes damage to or narrowing of the coronary arteries. These arteries provide the heart with the essential oxygen and nutrients, but they are unable to do so entirely because blockage with cholesterol has built up at the arteries.

Heart failure: When this issue arises, the heart fails to efficiently pump blood to various body parts. It's possible that you have a severe case of coronary artery disease, which stops your heart from pumping blood.

Congenital Cardiac Disease: This particular heart disease or abnormality usually shows up at birth. For instance, the two parts of the heart may have holes or septal anomalies. Cyanotic heart disease is characterised by an oxygen shortage in the body, and blockage problems, signal that the blood flow via different parts of the heart is entirely or partially blocked.

Cardiomyopathy: Cardiomyopathy weakens or changes the structure of the heart's muscles as the heart's capacity to pump blood declines. This may lead to heart failure.

II. RELATED WORK

The objective is to create an intelligent agent that can predict cardiovascular illness and research the necessary actions before any unfavourable incidence occurs. The ensemble approach has been suggested by Uddin, M. N et al. The model uses various feature selection strategies, including Information Gain Attribute Evaluator and Pearson Correlation. With the use of the classification methods Random Forest, Naive Bayes, and Gradient Boosting, this model aims to build an ensemble technique. [1]. Nawaz et al. proposed a study that seeks a more accurate and efficient diagnosis of heart issues. The dataset used for the analysis in the research was one on heart disease from the UCI Machine Repository. Applied machine learning algorithms include K-Nearest Neighbor, Naive Bayes, Artificial

Neural Network, Support Vector Machine, Gradient Descent Optimization and Random Forest. When compared to other classification techniques, the most effective optimization method is gradient descent. 1000 instances from a small dataset were used to build the models. The same models can be applied to compare performance on massive datasets [2]. The goal of study by Ghosh et al. is to successfully predict cardiovascular disorders. Combined datasets including Switzerland, Hungary, Cleveland and Long Beach VA. The Relief, Least Absolute Shrinkage and Selection Operator (LASSO) algorithm is used to do feature extraction. To develop novel hybrid classifiers, the standard classifiers are improved and bagged. K-Nearest Neighbors Bagging Method, the Decision Tree Bagging Method, Random Forest Bagging Method, Gradient Boosting Boosting Method and AdaBoost Boosting Method are a few of the novel hybrid classifiers. [3].

Employing data mining and machine learning techniques, three different kinds of heart blockages were predicted in this study by Rasel et al.: the first-degree A-V block (RBBB), the right bundle branch block (RBBB), and the left bundle branch block (LBBB). The cardiology division of Chittagong Medical College Hospital (CMCH) in Bangladesh is where the data samples are gathered. The dataset has 32 distinct kinds of numerical and categorical characteristics that reflect the patient's daily activities, food habits, and ECG report. Random Forest, K Nearest Neighbour, Decision Tree and Support Vector Machine are the algorithms implemented. The Random forest and Decision tree algorithms are the most efficient of the other methods. [4]. Velusamy et al. research's goal is to develop a machine learning technique for detecting coronary artery disease. The heterogeneous ensemble approach uses three fundamental classifiers: KNN, Random Forest, and support vector machines. The medical dataset collected by Z-Alizadeh Sani contains the clinical records of 303 patients with 56 feature attributes. Many ensemble classification procedures use voting techniques including majority voting, average voting and weighted average voting. The WAVEn algorithm performs better [5]. The hybrid decision support system described in Rani et al. study can help in the timely identification of cardiovascular diseases. The multivariate imputation approach employing chained equations has been utilised to handle the missing data, depending on the clinical characteristics of the patient. A hybridised feature selection technique that combines recursive feature reduction and the Genetic Algorithm (GA) has been used. Data preparation methods include SMOTE (Synthetic Minority Oversampling Technique) and conventional scalar procedures. The data has been classified using Support Vector Machine, Random Forests, Naive Bayes, Logistic Regression and Adaboost classifier. The random forest approach yields the best outcomes [6].

Making a precise diagnosis of heart illness is the goal of Nawaz et al research. This is beneficial when studying cardiovascular disease. the UCI Machine Repository's dataset on cardiac diseases. An optimization algorithm can improve the sensitivity and accuracy of heart disease diagnosis. Support vector machine (SVM), K-Nearest neighbour (KNN), Naive bayes (NB), Artificial neural network (ANN), Random forest (RF), and Gradient descent optimization are the machine learning methods used. Gradient Descent Optimization (GDO) performs well. This study demonstrated considerable improvements in capability for the detection of cardiovascular disease [7]. Beyond the Pooled Cohort Equations (PCE), van der Toorn et al. research attempts to determine the sex-specific relevance of vascular calcification. Carotid artery calcification, coronary artery calcification, aortic arch calcification, and extracranial and intracranial (ECAC) calcification are a few of these (ICAC) [8]. In this study, the ASCVD risk calculator will be computed utilising data from electronic medical records (EMR) and various ML algorithms. Li, Q et al gather around 1,01,110 EMR records from patients who are still alive and apply features like CS (cross sectional) and LT (longitudinal) to the EMR dataset. The ASCVD risk calculator outperforms the PCE calculator. When LT and CS data are combined, ASCVD prediction is better than when CS features are used alone [9].

In this study, machine learning techniques and deep learning techniques are used by Swathy, M et al, to forecast cardiac diseases. The data mining, classification, machine learning, and deep learning techniques used to predict cardiovascular illnesses are compared and reported in this study. These three groups—machine learning models for Cardio vascular diseases, deep learning models for cardio vascular disease prediction and data mining approaches for CVD. The three-pronged strategy that has been outlined is limited to the precise types of automation that are employed, together with their algorithms and techniques, to accurately forecast CVD. According to this study, we may either link them and select a certain approach or employ a number of techniques to create a suitable model. [10]. Wander et al. study predicts cardiovascular illnesses using machine learning as well as deep learning methods. The authors of this study looked at the most significant cultural norms for predicting and evaluating CVD in people with or without DM. The recommended strategy could help doctors better examine and treat their diabetes patients' cardiovascular conditions. The suggested approach should assist doctors in enhancing cardiovascular assessment and treatment of their diabetic patients [11]. In this study by Chu, C. S et al., life-time risk models for non-fatal coronary heart diseases from both CVD and non-CVD sources will be developed. As a result, 92,915 individuals' data was gathered. who took part in a lifestyle modification programme with a community focus. All of the analyses in this research is done in terms of gender and age. We can make excellent and bad profiles by modifying three variables. These models are used in this study to assess the probability of the first CVD events for various risk factor profiles. The estimations are more precise since the projections are based on a specific diagnosis [12].

The goal of the study proposed by Ali, M. M et al. was to compare various algorithms and identify the top ML approaches. The dataset was obtained via Kaggle. ReplaceMissingValues, Interquartile Range (IQR) filters, Random Forest, Decision Tree (DT), AdaboostM1 (ABM1), Logistic Regression (LR), and Multilayer Perceptron (MLP) algorithms were used for preprocessing [13]. Kamalapurkar et al study presents a web-based method for predicting cardiac illness that uses machine learning (ML) techniques. The creation of an accurate system or model is the primary goal of proposed model so that it is ensured that patients won't receive the wrong diagnosis owing to faulty forecasts and that it tends to help the patient in an emergent situation. The approach of ensemble classification is utilised to forecast cardiac disease. Competence of machine learning algorithms when seen separately, K-Nearest Neighbor, Decision Tree, SVM, and RF demonstrate that Random Forest performs best among the four methods, followed by SVM [14]. The research by Guo, C et al uses machine learning approaches to identify the important characteristics of the prediction of cardiovascular illnesses. Numerous feature combinations and well-known classification methods are included in the prediction model. The characteristics of the linear model and random forest are combined using the RFRF-ILM approach that has been presented. In this study, decision trees are used to cluster the datasets. The Internet of Medical Things (IoMT) platform permitted the analysis of important variables for data analysis. [15].

III. METHODOLOGY

The major health concern is heart disease, which has affected many people throughout the world. The suggested methodology seeks to create a smart agent that can foretell the presence of any cardiovascular illness well before any untoward incident occurs.

3.1 Dataset

A kaggle dataset consisting of around 70000 entries. The dataset consists of different features such as age, gender, cholesterol, heart rate, exercise, alcohol, smoking etc. These all features predict whether a particular person has any cardio vascular disease or not. The target class of these datasets is a binary class. It depicts whether cardio vascular disease is present or not, i.e. 0 or 1.

3.2 Data preprocessing:

Data processing is the process of converting the raw data into feasible format for implementing a machine learning model on it. This is the first stage to develop a machine learning model. It's possible that the initial raw data has different odds. So it has to be ensured that the data that is fed to the machine learning model is clean enough. Data preprocessing is done to clean the raw data to acquire desirable result. Data preprocessing include the following steps: Importing the librariesBringing the dataset in, Find the missing information, categorical data are encoded, Feature scaling and dividing the dataset into train set and test set.

To process the data, there are many libraries required such as numpy, pandas, matplotlib etc. Generally the datasets are in the form of .csv files. CSV stands for Comma Separated Values. In a dataset, there are two types of variables. They are dependant and independent variables. The dependant variable is nothing but the target variable. There can be a chance that the data may have missing values. These missing values can lead to mis-predictions also. So it is very important to remove the missing data. There are two methods for dealing with missing data. By removing the specific row, and by figuring out the mean and filling in the missing value.

3.3 Feature extraction

It is today changing into quite common to be operating with datasets of many options. Feature Extraction aims to scale back the amount of options in a very dataset by making new options from the present ones. Feature selection aims instead to rank the importance of the present options within the dataset and discard shorter ones. Reducing the amount of options to use throughout a applied math analysis will presumably result in many advantages like Accuracy enhancements, Overfitting risk reduction etc. If a lot of options are superimposed than those that are strictly necessary, then our model performance can simply decrease. It's vital that optimum number of options to be used.

The following are the feature choice techniques used.

3.3.1 Information gain attribute

By dividing a dataset by a specified value for a variable, information gain quantifies the decrease in entropy or surprise. Information gain demonstrates how to utilise entropy to determine how a change to the dataset affects its purity. By measuring the information gain in relation to the class, it determines if an attribute is effective.

3.3.2 Pearson's correlation

The Pearson Product Moment Correlation is the full name of PPMC. The parametric statistic Pearson's correlation, commonly referred to as Pearson's R, is frequently used in regression analysis. It displays the two sets of data's linear relationship. Pearson's Correlation methodology is employed for locating the association between the continual options and also the category feature. The Pearson correlation is represented by the Greek letters rho (ρ) for a population and "r" for a sample.

3.4 Algorithms used

3.4.1 Ensemble Approach

A machine learning method called ensemble learning combines the predictions from various models to produce prognosticative performance.

In ensemble learning, there are three main categories:

- Bagging
- Stacking
- Boosting

Bagging is the process of averaging the results of several decision trees that have been fitted to various samples obtained from the same dataset. Stacking is the process of fitting many models to the same data and using a different model to figure out how to best combine the results. Boosting includes successively adding ensemble members that update predictions made by earlier models, producing a weighted average of the forecasts. In the projected model, stacking model has been used. The algorithms used are Random Forest, Naïve Bayes, Gradient Boost. Stacked ensemble technique has been used. On entirely separate samples, Random Forest constructs decision trees and uses their majority vote for categorization and average in cases of regression. The Random Forest algorithm is a supervised machine learning approach. Depending on the Bayes theorem, naive Bayes algorithm is one of the supervised learning approaches for handling classification issues. Being a probabilistic classifier, it relies its judgments on the probability of an item. The basic idea behind boosting techniques is that after designing a model on the training dataset, we build a second model to correct any flaws in the first model. The main idea behind this algorithm is to make models consecutively and cut back the errors from the previous model.

3.4.2 Support Vector Machine

Classification and regression problems are addressed with Support Vector Machine. The goal of the technique is to identify the best classification boundary or line for an n-dimensional space, allowing subsequent data points to be quickly assigned to the appropriate class. This optimal decision boundary is referred to as a hyperplane. In essence, the model is a multidimensional hyperplane representation of several categories. Because SVM will continually generate the hyperplane, the error will be decreased.

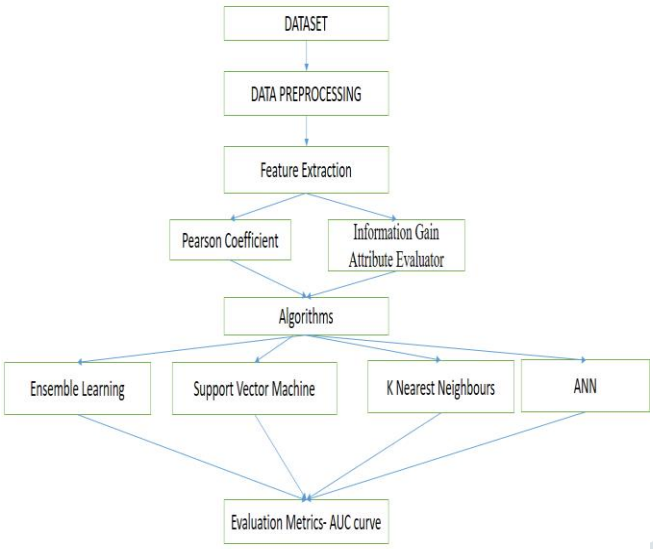
3.4.2 K Nearest Neighbours

KNN algorithm is a straightforward supervised machine learning (ML) technique for both Classification as well as Regression applications. It backed up the idea that the observations in a data collection that are most "similar" to a given piece of information are the observations, and that we should categorise unanticipated points based on the values of the nearby existing points. The range of closest neighbours to utilise is K.

3.4.3 Artificial Neural Network

The term "artificial neural network" originates from the networks of biological neurons that define the architecture of the physical brain of human beings. Dendrites from biological brain networks are used as inputs, organelles as nodes, clumps as weights, and nerve fibres as outputs in artificial neural networks. A computer's ability to see the world and make decisions in a manner that is shockingly human is enabled by an artificial neural network that makes an attempt to mimic the neural network seen in the human brain.

3.4.4 Model design

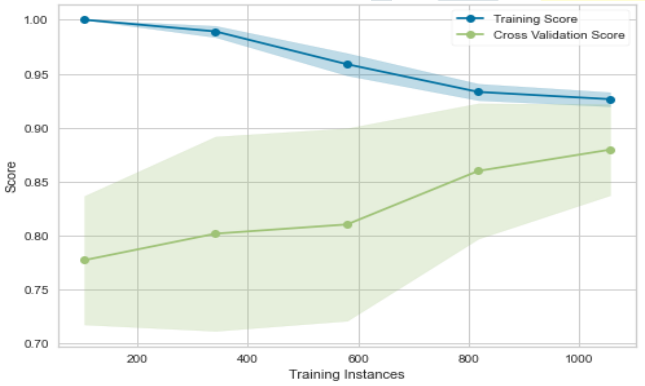


IV. RESULTS

The proposed ensemble model secures higher accuracy of 0.94 when compared to existing models. The dataset consists of 70000 entries. Upon the large dataset, the model shows higher accuracy. The deep learning model, Artificial neural networks shows 0.97 accuracy which is much higher than the existing models.

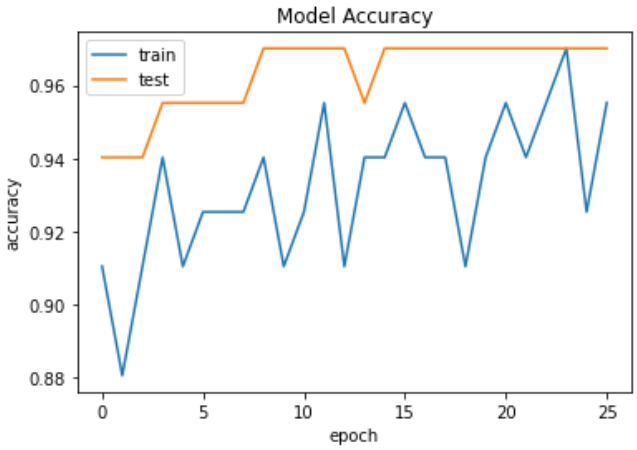
4.1 AUC ROC Curve

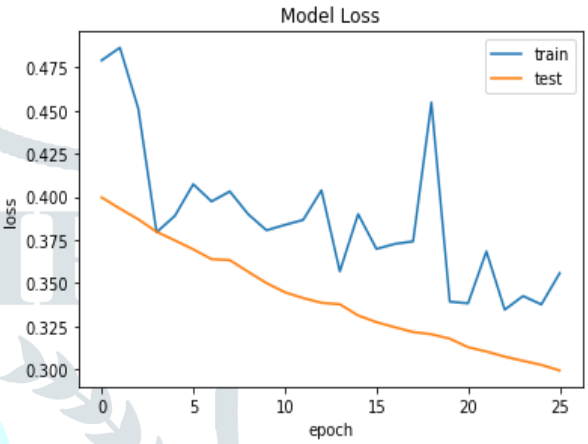
Performance for categorization problems at various threshold levels is measured by the AUC-ROC curve. Area Under The Curve is referred to as AUC. Receiver Operating Characteristics are referred to as ROC.



Ensemble learning accuracy: 0.94

The model accuracy curves are as follows





V. CONCLUSION

The early identification of heart illness is particularly helpful in predicting the occurrence of cardiovascular disease. When compared to earlier works, the ensemble learning model's accuracy has shown a noticeable improvement. The accuracy shown by Ensemble approach- 0.94 and Artificial Neural Networks- 0.97. It is ensured that based on the inputs given, the models can predict the presence of cardio vascular disease. The algorithm for artificial neural networks has been used. When compared to these models, earlier research has demonstrated less accuracy.

VI. FUTURE SCOPE

It is possible for the models' running times and accuracy to improve. To get better results, different hybrid classifiers might be utilised. Without the need for any special equipment, this model can be used when there is a suspicion of any cardiovascular condition. There can be different novel algorithms be used to obtain desirable results.

VII. REFERENCES

[1] Uddin, M. N., & Halder, R. K. (2021). An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach. Informatics in Medicine Unlocked, 24, 100584.

[2] Nawaz, M. S., Shoaib, B., & Ashraf, M. A. (2021). Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization. Heliyon, 7(5), e06948.

[3] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 19304-19326.

[4] Rasel, R. I., Sultana, N., Meesad, P., Chowdhury, A., & Hossain, M. (2020, July). 1st-degree Atrioventricular (AV-block) and Bundle Branch Block Prediction using Machine Learning. In Proceedings of the 11th International Conference on Advances in Information Technology (pp. 1-6).
[5] Velusamy, D., & Ramasamy, K. (2021). Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. Computer Methods and Programs in Biomedicine, 198, 105770.
[6] Rani, P., Kumar, R., Ahmed, N. M., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments, 7(3), 263-275.
[7] Nawaz, M. S., Shoaib, B., & Ashraf, M. A. (2021). Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization. Heliyon, 7(5), e06948.
[8] van der Toorn, J. E., Bos, D., Arshi, B., Leening, M. J., Vernooij, M. W., Ikram, M. A., ... & Kavousi, M. (2021). Arterial calcification at different sites and prediction of atherosclerotic cardiovascular disease among women and men. Atherosclerosis, 337, 27-34.
[9] Li, Q., Campan, A., Ren, A., & Eid, W. E. (2022). Automating and improving cardiovascular disease prediction using Machine learning and EMR data features from a regional healthcare system. International Journal of Medical Informatics, 163, 104786.
[10] Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. ICT Express, 8(1), 109-116.
[11] Wander, G. S., Bansal, M., & Kasliwal, R. R. (2020). Prediction and early detection of cardiovascular disease in South Asians with diabetes mellitus. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 385-393.
[12] Chu, C. S., Chan, H. C., Tsai, M. H., Stancel, N., Lee, H. C., Cheng, K. H., ... & Ke, L. Y. (2018). Range of L5 LDL levels in healthy adults and L5's predictive power in patients with hyperlipidemia or coronary artery disease. Scientific Reports, 8(1), 1-9.
[13] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, 104672.
[14] Kamalapurkar, S., & GH, S. G. (2020, October). Online portal for prediction of heart disease using machine learning ensemble method (PrHD-ML). In 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC) (pp. 1-6). IEEE.
[15] Guo, C., Zhang, J., Liu, Y., Xie, Y., Han, Z., & Yu, J. (2020). Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform. IEEE Access, 8, 59247-59256.
[16] Amarbayasgalan, T., Pham, V. H., Theera-Umpon, N., Piao, Y., & Ryu, K. H. (2021). An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets. IEEE Access, 9, 135210-135223.
[17] Ashri, S. E., El-Gayar, M. M., & El-Daydamony, E. M. (2021). HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm. IEEE Access, 9, 146797-146809.
[18] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. IEEE Access, 8, 107562-107582.
[19] Ahmad, G. N., Fatima, H., & Saidi, A. S. (2022). Efficient Medical Diagnosis of Human Heart Diseases using Machine Learning Techniques with and without GridSearchCV.
[20] KARIM, A., SHAMRAT, F. J. M., IGNATIOUS, E., SHULTANA, S., BEERAVOLU, A. R., & DE BOER, F. R. I. S. O. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques.
[21] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 26, 100655.
[22] An, Y., Huang, N., Chen, X., Wu, F., & Wang, J. (2019). High-risk prediction of cardiovascular diseases via attention-based deep neural networks. IEEE/ACM transactions on computational biology and bioinformatics, 18(3), 1093-1105.
[23] Song, S., Chen, T., & Antoniou, G. (2021, March). ANFIS Models for Heart Disease Prediction. In 2021 the 5th International Conference on Innovation in Artificial Intelligence (pp. 32-35).
[24] Zhang, X. (2021, April). Using Data Visualization to Analyze the Correlation of Heart Disease Triggers and Using Machine Learning to Predict Heart Disease. In 2021 3rd International Conference on Intelligent Medicine and Image Processing (pp. 127-132).
[25]Rajendran, P., Haw, S. C., & Naveen, P. (2021, September). Classification of Heart Disease Using Machine Learning Techniques. In 2021 5th International Conference on Digital Technology in Education (pp. 130-135).
[26]Rasool, A., Tao, R., Kashif, K., Khan, W., Agbedanu, P., & Choudhry, N. (2020, February). Statistic Solution for Machine Learning to Analyze Heart Disease Data. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing (pp. 134-139).
[27]Hassani, M. A., Tao, R., Kamyab, M., & Mohammadi, M. H. (2020, May). An approach of predicting heart disease using a hybrid neural network and decision tree. In Proceedings of the 2020 5th International Conference on Big Data and Computing (pp. 84-89).
[28]Lu, H. Y. (2020, August). Applying propensity score and support vector machine to construct a predictive model for heart disease. In Proceedings of the 4th International Conference on Medical and Health Informatics (pp. 18-21).
[29]Ghosh, P., Azam, S., Karim, A., Jonkman, M., & Hasan, M. Z. (2021, May). Use of efficient machine learning techniques in the identification of patients with heart diseases. In 2021 the 5th International Conference on Information System and Data Mining (pp. 14-20).
[30]Li, G., Yang, H., Guo, T., & Wang, W. (2022, January). Implementation and acceleration scheme of Heart sound classification Algorithm based on SOC-FPGA. In 2022 2nd International Conference on Bioinformatics and Intelligent Computing (pp. 258-266).