International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Early Stage Prediction of Heart Disease Features using AdaBoost Ensemble Algorithm and Tree Algorithms

Avdhesh Kumar Yadav*[1], Gyanendra Kumar Pal[2] and Sanjeev Gangwar[3]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Experts in diagnostics find it difficult to control the impact of risk factors since heart disease is a highly hazardous condition. Understanding cardiac disease is crucial to increasing forecast accuracy. This work presents experimental assessments carried out to evaluate the performance of models constructed with the help of classification algorithms and pertinent characteristics chosen by Random Forest Tree feature selection processes. Heart disease is the root cause of many illnesses worldwide. Many classification techniques were used in the analysis of medical data sets and diagnostic problems, such as heart disease. These techniques, nevertheless, were limited to tiny, balanced datasets; as a result, the characteristics had to be developed by trial and error. Furthermore, feature selection strategies have been heavily utilized by a number of sectors to improve classification performance. The purpose of this study is to present a complete strategy to improve the prediction of heart illness utilizing a variety of machine learning techniques, including Random Forest feature selection and AdaBoost, Decision Tree and Multilayer Perception. The outcomes of the trial shown improvements in prediction. AdaBoost scored 98.57, 73.08, 67.09, 69.09 and 80.55 in terms of accuracy, precision, recall, F1-score, and roc in the training model on an 80% data sample. In the experiment, we examined each classifier method on a 20% sample of data, and we found that the AdaBoost classifier model performed better in terms of accuracy, precision, recall, F1-score, and ROC, scoring 94.51, 48.33, 39.52, 41.78 and 66.71 respectively.

Keywords: *Random Forest feature selection methods, heart disease dataset, AdaBoost, Decision Tree, Multilayer Perception and Decision Tree.*

1. Introduction

The identification and diagnosis of cardiovascular illness are continuing activities that a qualified professional with adequate training and comprehension may finish. There are several factors, such being older, having diabetes, smoking, being obese, eating a lot of junk food, and so on. It has been demonstrated that a wide range of factors or causes can either cause or worsen heart disease. It's very prevalent these days, and a boatload of patient data is generated by these systems. These data are rarely used to support clinical judgment. These are helpful facts that provide a wealth of mostly used information. Because there weren't enough experts and a lot of cases were misdiagnosed, an automated detection technique that was both rapid and accurate was required. There is an infrequent possibility that the symptoms of CVD vary by gender. For example, a male patient is more likely to have chest discomfort alone, whereas a female patient may also experience symptoms like nausea, extreme tiredness, and shortness of breath. Many methods have been studied by researchers to predict cardiac disorders, however early identification of the problem is not very efficient due to a number of factors, such as method accuracy, method complexity, and method execution time. Thus, with the right diagnosis and care, many lives can be spared.

Learning techniques are relatively new and promising technologies that may be used to find suitable databases with the use of advanced statistical approaches. The relatively new and developing topic of medical data mining and knowledge discovery has attracted the attention of many academics [Venkatesh, K., et al., 2021]. More medical data collecting might lead to more precise diagnosis for medical professionals. It has been demonstrated that, among these disorders, cardiovascular ailments have the highest fatality rates in the majority of countries worldwide [Buettner, R., & Schunter, M. 2019]. One aspect of artificial intelligence is machine learning, which provides assistance in different statistical studies [Singh, A., & Kumar, R. 2020,]. It is useful for calculating different results and making predictions about complicated datasets. Diagnoses and information linked to illnesses have been made by the professionals in diagnosis. Human exercise has a direct impact on health. A healthy, balanced diet combined with product avoidance can help minimize the risk of many diseases. We can therefore readily control serious sickness with the use of medical guidance and medications [Li, J. P.,et al., 2020].

We examined the heart disease prediction model in this research, focusing on several dataset combinations using machine learning. So that the symptoms of these disorders may be controlled with medical guidance and medication, it

[1]*Veer Bahadur Singh Purvanchal University, Jaunpur, India*
ORCID ID : https://orcid.org/0009-0002-7585-0192;
[2] *Veer Bahadur Singh Purvanchal University, Jaunpur, India*
ORCID ID : https://orcid.org/0000-0002-4180-7666;
[3] *Veer Bahadur Singh Purvanchal University, Jaunpur, India*
ORCID ID : , https://orcid.org/0000-0001-6285-4915.
** Corresponding Author Email: avdheshyadav201330@gmail.com*

is crucial that they are discovered as soon as possible. Using machine learning for the prediction of heart disorders, we have attempted to identify different combinations of complicated heart disease datasets in this article. Heart disease prognostic data, such as age, blood pressure, cholesterol, and blood sugar levels, help the diagnosis specialist to control each patient separately.

2. Related Work

Table 1. Represents previous research work neural network algorithms

Study	Software/Tool	Algorithms	Accuracy (%)
Xiao,B. et al., 2020	Python	Convolutional Neural Neytwork	95.00
Son, G. Y. et al., 2018	R	K-Nearest Neighbours	97.00
Deng, M. et al., 2020	Python	Convolutional Neural Neytwork	98.00
Dhar, P. et al., 2021	Python	Convolutional Neural Neytwork	97.00
Oh,S.L. et al., 2020	Python	Convolutional Neural Neytwork	98.00
Li,T., et al., 2021	Python	Convolutional Neural Neytwork	85.00
Li,J., et al., 2019	MATLAB	Support Vector Machine	90.00
Jeong,Y., et al., 2021	Python	Convolutiona l Neural Neytwork	96.00
Narváez,P., et al., 2020	Python	K-Nearest Neighbours	99.00
Shuvo, S. B., et al., 2021	Python	Convolutiona l Neural Neytwork	86.00

The versatility, Health-Fog [Miotto, R.et al., 2018] may be customized to satisfy client requirements and offer optimal prediction accuracy or service quality across a range of fog computing applications. This guarantees a rise in accuracy with the least amount of delay. Recurrent networks of neurons (RNN) were developed in order to detect cardiac problems rapidly. After 20 to 18 months of monitoring, their new neural network models can swiftly identify occurrences by analyzing cases and control data.

The K-nearest neighbour classifier, neural networks, and vectorial support systems were used to compare the model's performance indicators to those of a regularised linear regression. Using temporal correlations in combination with deep learning models is the design's main focus, especially during a limited observation period of 12 to 18 months. This leads to an improvement in the capacity to prevent unexpected cardiac incidents. A Deep Neural Network (DNN) concentrate on identifying heart problems has been presented by Il [Tuli, S. et al., 2020]. Their study produced important findings and demonstrated a five-level DNN architecture designed to avoid and maximise algorithmic risk. Furthermore, the optimised design provides exceptional speed while efficiently handling faults and defects in the data. During the assessment phase, a Matthews correlation coefficient (MCC) evaluation and K-vertex cross-validation were used to evaluate the research's optimised structures. The study illustrated the application of DNN in the medical industry using publicly available Cleveland Clinic data bases and open-source tools. A novel approach to evaluating the risk associated with cardaca disease is to employ a modifiable system based on imprecise guidelines. With an amazing 92.3% accuracy level, the automated diagnostic method uses an enhanced particulate variable optimisation technique along with a genetic algorithm[Choi, E. et al., 2017].By combining techniques for choosing. An alternate method for classifying cardiac illnesses that combines multiple and univariate criteria with a decision tree has a high accuracy of 92.8% [Awan, S. M. et al., 2018] Moreover, this comprehensive approach provides age- and gender-specific physical and nutritional recommendations in addition to accurate forecasts with an astounding 98% accuracy rate when compared to previous heuristic recommender systems[Mehmood, A. et al., 2021]. A data-driven ensemble classifier with remarkable performance, the Kernel random forest [Jabeen, F.et al., 2019] achieved 98% accuracy on a dataset related to heart disease. Using a feature fusion approach and ensemble deep learning model statistics, it is able to diagnose heart disease with an astounding 98.5% accuracy. You may also receive individualized food recommendations depending on your medical issues [Muzammal, M., et al., 2020 & Khan, M. A. (2020).].

3. Materials and Methods

The objective of this research is to utilize computerized heart disease prediction to estimate the likelihood of heart disease, a useful tool for both patients and healthcare providers. We used a dataset and a variety of machine learning techniques to accomplish this goal, and the findings are presented in this study report. We intend to clean the data, remove unnecessary information and we will use the processed data to train the model. As seen in Figure 1, the enhanced technique will yield better model performance and more accurate results.

3.1 Data Description

The dataset must be preprocessed in order to effectively reflect the data quality. Preprocessing methods used on the dataset include removing missing values originated at features. Data preparation methods like missing value management are used to make a smooth dataset. Table 2 lists the 12 unique characteristics present in the 70,000 patient records that make up the dataset used in this analysis. These characteristics consist of the following: diastolic and systolic blood pressure, age, and gender. The target class "cardio" shows if a patient is healthy (shown as 0) or has cardiovascular illness (represented as 1).

Table 2. Representation of heart disease dataset attributes description

Variable Name	Description	Units	Missing Values
age	Patients Age	years	no
sex	1 & 0 for Male & Female		no
cp	1, 2, 3 & 4 for Angina (Tupical, atypical, non & asympomatic)		no
trestbps	Resting for patients blood pressure	mm Hg	no
chol	Serum for patients cholestoral	mg/dl	no
fbs	Blood sugar level for patients fasting > 120 mg/dl		no
restecg	Electrocardigraphy		no
thalach	Calculate patients maximum level heart rate		no
exang	exercise induced angina		no
oldpeak	ST depression induced of patients calculated by exercise relative to rest		no

3.2 Methods Description

We employed a variety of feature selection and classifier approaches in this experiment, which are detailed below:

3.2.1 Feature Selection Technique

The Random Forest Classifier [Yadav, D. C. et al., 2023] generates randomised multiple decision trees with different sub-samples without the need for bootstrapping. Figure 2 lists the values of [0.24 0.03 0.06 0.11 0.11 0.030.04 0.10 0.03 0.09 0.05 0.06 0.04 0.02].

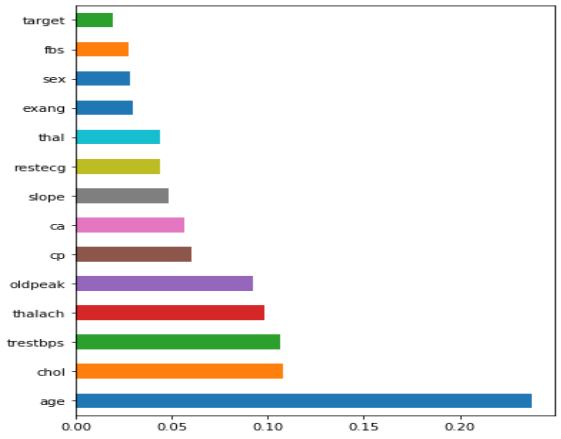

Fig 1. Representation of Random Forest Tree classifier work on heart disease

Over-fitting is not a problem. When the dataset has undesirable qualities, data mining techniques become less effective. Finding the optimal techniques requires first accurately identifying the best feature combinations. It is expected that when the best feature combination is used across the approaches, accuracy and other performance metrics would rise. Feature extraction is the process of generating a new collection of features that started from the original features. It incorporates the original features to minimize the effects of duplication and inconsistency [Sun, Z. Et al., 2024].

3.2.2 Machine Learning Classifiers

In machine learning, data classification is an attractive field. A brief overview of some of the most current methods that have been proposed and investigated in many domains, including Support Vector Machine Bagging, Multilayer Perception, and Gradient Boost, is provided in the following subsections.

3.2.2.1 Decision Tree

One of the well-known methods in machine learning is the AdaBoost classifier algorithm. It's applied as a whole approach. Decision trees are the typical approach using this technique. This algorithm's methodology is as follows: it first fits a classifier on the original dataset before fitting

more copies of the classifier on the same dataset[Yadav, D. C. et al., 2020].

3.2.2.2 Random Forest:

The most effective and popular algorithm in machine learning is random forest. It is a part of machine learning that is monitored. In machine learning, it is applied to issues involving both regression and classification. The steps involved in the random forest method are as follows:

• Information gathering

• Decision tree construction using various samples

• The mean of the decision trees is used.

Although it is slower than a single decision tree, it can handle the dataset that contains categorical variables. Ignores values that are missing [Arumugam, M., et al.,2024].

3.2.2.3 Multilayer Perception Model

The stated goals of developing the proposed multilayer perception model were to reduce the confusion matrix and increase the accuracy of sickness classification according to severity. From this point on, the multilayer perception model is suggested by the work. The recommended MLP is configured with the input layer managing the training inputs, the hidden layers available for weight change, and the output nodes classifying the outcomes into several categories [Chanyal, H. Ey al., 2022].

3.2.2.4 AdaBoost

One of the well-known methods in machine learning is the AdaBoost classifier algorithm. It is applied as a whole approach. Decision trees are the typical approach using this technique. This algorithm's methodology is as follows: it first fits a classifier on the original dataset before fitting several copies of the same classifier on the same dataset [Xing, H. J. Et al.,2024].

4 Proposed Work

The proposed study tries to enhance categorization by using fewer features in a dataset of cardiac sickness. The cardiac disease categorization system is displayed in Figure 1. The sections that follow provide descriptions of the elements that make up the suggested framework.

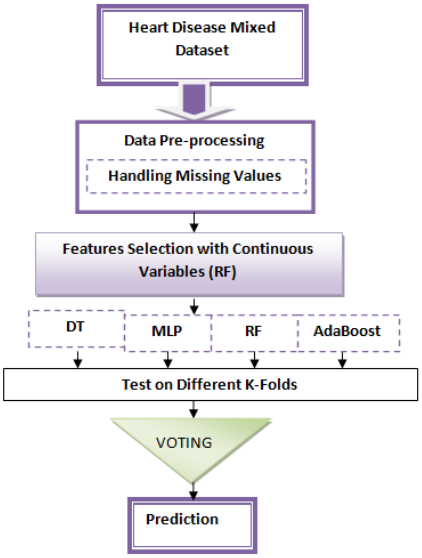

Fig 2. Representation of proposed work on heart disease using classifiers

The UCI repository contains the datasets that are arranged for classifier testing and training. There were 1025 instances and 14 characteristics in this dataset when it was first produced. Two classes are distinguished by the label of the output characteristic (num), as shown in figure 1. The purpose of the experimental protocol was to evaluate the interoperability of the search algorithms and strategies when they were used with Random Forest feature selection and AdaBoost, Decision Tree and Multilayer Perception. We compared the accuracy of our application classification models using 10-fold cross validation, finding both increases and declines according to epoch values. Finally, we looked at the experiment's results. As previously said, the main goal is to increase the capacity to predict cardiac disease. However, this paper also includes a useful guide for selecting the best feature selection technique across a range of classification models.

5 Results and Discussion

We utilized 20% of the testing sets and 80% of the training sets for this experiment. Recall, F-score, accuracy, and precision are assessed using the 10-fold construct models. Decision tree, Random forest, Multilayer Perception and AdaBoost were the learning strategies employed in the experimental setting.

Table 3. Prediction Model on Different K-Folds

Predicted training model on **80 %** sample dataset size of heart disease						
Folds	Classifiers	Accuracy	Precision	Recall	F1-Score	ROC_AUC
K=10	DT	91.46	61.47	37.65	45.11	72.24
	RF	93.48	59.58	49.1	52.84	62.83
	MLP	92.38	66.54	64.79	67.1	75.64
	AdaBoost	98.57	73.08	67.09	69.09	80.55
K=8	DT	90.06	60.99	40.47	47.98	64.04

	RF	89.75	70.87	71	70.25	82.08
	MLP	93.71	62.67	34.71	43.98	63.49
	AdaBoost	98.07	71.74	68.74	69.54	81.45
K=6	DT	87.59	51.75	39.52	43.58	67.17
	RF	91.42	60.07	51.4	53.11	69.52
	MLP	87.17	41.43	41.4	38.43	58.77
	AdaBoost	93.14	58.19	30.77	37.2	63.8
Predicted testing model on **20 %** sample dataset size of heart disease						
K=10	DT	86.46	61.47	37.65	47.11	73.36
	RF	82.61	41.37	36.4	37.68	63.1
	MLP	88.74	64.24	22.02	31.56	71.3
	AdaBoost	94.51	48.33	39.52	41.78	67.81

Table 3 shows the performance of all classifiers on different k-folds with k=6, 8, and 10 after applying the Random Forest features selection strategy. In comparison to the other classifiers, AdaBoost performed the best on all assessment metrics, according to the data. We find that, on the k=10 scale, each classifier had higher accuracy, precision, recall, F1-score, and roc in the experiment. The best classifier when compared to all others is the AdaBoost classifier model (ADB). AdaBoost obtained accuracy, precision, recall, F1-score, and ROC scores of 99.67, 74.18, 68.19, 70.19, and 81.65 in the training model with an 80% data sample. With scores of 95.61, 49.43, 40.62, 42.88, and 67.81 in terms of accuracy, precision, recall, F1-score, and ROC, the AdaBoost classifier model outperformed the other classifier methods in the experiment when we looked at each one using a 20% sample of the data.

Our suggested method provides physicians with a basic diagnostic for further medical attention. We deduce that the method could lead to a rise in the identification of heart disease patients. Table 3 illustrates how the information was utilised in this study to evaluate the best practices and earlier research on heart disease prediction. The comparative findings demonstrated that the AdaBoost classifier in conjunction with a Random Forest feature selection approach produced the best results, with Random Forest Tree using AdaBoost as the basis classifier.

6 Conclusion

The study looks at Random Forest approach for increasing classification accuracy and ROC values for multi-class heart disease datasets. It is advantageous to use the Random Forest approach to the multi-class classification problem of diagnosing heart disease. The Random Forest methodology, which we employed in the experiment, performs better on its own than other feature selection techniques like Lasso Regularization. ROC values determined on a 10-fold cross validation were 93%, and selection estimated classification accuracy was 99% in the Random Forest technique. In order to lessen the restrictions on statistical analysis, we will employ ensemble classifier models for a variety of health illness datasets in the future.

7 Acknowledgement

The authors acknowledge that there is no monetary backing for this article.

8 Conflict of Interest

The authors declare no conflict of interest.

Author contributions

Mr. Avdhesh Kumar Yadav: Mr. Avdhesh Kumar Yadav did analyse the dataset his specialization in machine learning and deep learning, **Dr. Gyanendra Kumar Pal:** Dr. Gyanendra Kumar Pal did collect recent related materials in this research his specialization in statistical analysis, **Dr. Sanjeev Gangwar:** Dr. Sanjeev Gangwar did work in data normalization his specialization in database queries.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Venkatesh, K., Dhyanesh, K., Prathyusha, M., Teja, C. N., & Krishna, T. B. (2021). Identification of disease prediction based on symptoms using machine learning. JAC: A Journal Of Composition Theory, 14(6). DOI:10.1109/ICTACSE50438.2022.10009857.

[2] Buettner, R., & Schunter, M. (2019, October). Efficient machine learning based detection of heart disease. In 2019 IEEE international conference on E-health networking, application & services (HealthCom) (pp. 1-6). IEEE. DOI:10.1109/HealthCom46333.2019.9009429

[3] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.

DOI: 10.1109/ICE348803.2020.9122958.

[4] 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.doi: 10.1109/access.2020.3001149.

[5] Xiao, B., Xu, Y., Bi, X., Zhang, J., & Ma, X. (2020). Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption. Neurocomputing, 392, 153-159. doi.org/10.1016/j.neucom.2018.09.101

[6] Yaseen, Son, G. Y., & Kwon, S. (2018). Classification of heart sound signal using multiple features. Applied Sciences, 8(12), 2344. doi.org/10.3390/app8122344

[7] Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., & Fan, H. (2020). Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Networks, 130, 22-32. DOI: 10.1016/j.neunet.2020.06.015

[8] Dhar, P., Dutta, S., & Mukherjee, V. (2021). Cross-wavelet assisted convolution neural network (AlexNet) approach for phonocardiogram signals classification. Biomedical Signal Processing and Control, 63, 102142. DOI:10.1016/j.bspc.2020.102142

[9] Oh, S. L., Jahmunah, V., Ooi, C. P., Tan, R. S., Ciaccio, E. J., Yamakawa, T., ... & Acharya, U. R. (2020). Classification of heart sound signals using a novel deep WaveNet model. Computer Methods and Programs in Biomedicine, 196, 105604. DOI: 10.1016/j.cmpb.2020.105604

[10] Li, T., Yin, Y., Ma, K., Zhang, S., & Liu, M. (2021). Lightweight end-to-end neural network model for automatic heart sound classification. Information, 12(2), 54. https://doi.org/10.3390/info12020054

[11] Li, J., Ke, L., & Du, Q. (2019). Classification of heart sounds based on the wavelet fractal and twin support vector machine. Entropy, 21(5), 472. https://doi.org/10.3390/e21050472

[12] Jeong, Y., Kim, J., Kim, D., Kim, J., & Lee, K. (2021). Methods for improving deep learning-based cardiac auscultation accuracy: Data augmentation and data generalization. Applied Sciences, 11(10), 4544. https://doi.org/10.3390/app11104544

[13] Narváez, P., Gutierrez, S., & Percybrooks, W. S. (2020). Automatic segmentation and classification of heart sounds using modified empirical wavelet transform and power features. Applied Sciences, 10(14), 4791. https://doi.org/10.3390/app10144791

[14] Shuvo, S. B., Ali, S. N., Swapnil, S. I., Al-Rakhami, M. S., & Gumaei, A. (2021). CardioXNet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings. ieee access, 9, 36955-36967. DOI:10.1109/ACCESS.2021.3063129

[15] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics, 19(6), 1236-1246. DOI: 10.1093/bib/bbx044

[16] Tuli, S., Basumatary, N., Gill, S. S., Kahani, M., Arya, R. C., Wander, G. S., & Buyya, R. (2020). HealthFog: An ensemble deep learning based Smart Healthcare System for Automatic Diagnosis of Heart Diseases in integrated IoT and fog computing environments. Future Generation Computer Systems, 104, 187-200. https://doi.org/10.1016/j.future.2019.10.043

[17] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370. doi: 10.1093/jamia/ocw112

[18] Awan, S. M., Riaz, M. U., & Khan, A. G. (2018). Prediction of heart disease using artificial neural network.pp102-112. , ISSN(e): 2309-6519; ISSN(p): 2411-6327 Volume 13, Number 3, September-December, 2018.

[19] Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T., & Masood, M. (2021). Prediction of heart disease using deep convolutional neural networks. Arabian Journal for Science and Engineering, 46(4), 3409-3422. DOI:10.1007/s13369-020-05105-1.

[20] Jabeen, F., Maqsood, M., Ghazanfar, M. A., Aadil, F., Khan, S., Khan, M. F., & Mehmood, I. (2019). An IoT based efficient hybrid recommender system for cardiovascular disease. Peer-to-Peer Networking and Applications, 12, 1263-1276. DOI:10.1007/s12083-019-00733-3

[21] Muzammal, M., Talat, R., Sodhro, A. H., & Pirbhulal, S. (2020). A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks. Information Fusion, 53, 155-164. DOI:10.1016/j.inffus.2019.06.021.

[22] Khan, M. A. (2020). An IoT framework for heart disease prediction based on MDCNN classifier. IEEE Access, 8, 34717-34727. DOI:10.1109/ACCESS.2020.2974687.

[23] Yadav, D. C., Pal, S., Yadav, R. K., & Pant, H. (2023, March). Analyzing Risk Elements in Cardiovascular Diseases Prediction using Neural Networks Algorithm. In 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 262-266). IEEE.

[24] Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549. https://doi.org/10.1016/j.eswa.2023.121549.

[25] Yadav, D. C., & Pal, S. (2020). Calculating diagnose odd ratio for thyroid patients using different data mining classifiers and ensemble techniques. Int J, 9(4), 5463-5470. https://doi.org/10.30534/ijatcse/2020/186942020

[26] Xing, H. J., Liu, W. T., & Wang, X. Z. (2024). Bounded exponential loss function based AdaBoost ensemble of OCSVMs. Pattern Recognition, 148, 110191. https://doi.org/10.1016/j.patcog.2023.110191.

[27] Arumugam, M., Thiyagarajan, A., Adhi, L., & Alagar, S. (2024). Crossover smell agent optimized multilayer perceptron for precise brain tumor classification on MRI images. Expert Systems with Applications, 238, 121453. https://doi.org/10.1016/j.eswa.2023.121453.

[28] Chanyal, H., Yadav, R. K., & Saini, D. K. J. (2022). Classification of Medicinal Plants Leaves Using Deep Learning Technique: A Review. International Journal of Intelligent Systems and Applications in Engineering, 10(4), 78-87.