Efficient Predictive Modelling for Classification of Coronary Artery Diseases Using Machine Learning Approach

Efficient Predictive Modelling for Classification of Coronary Artery Diseases Using Machine Learning Approach

Savita[1], Ganga Sharma[1], Geeta Rani[2], Vijaypal Singh Dhaka[2]

[1]Department of Computer Science, GD Goenka University, Gurugram, India
[2]Manipal University Jaipur, Jaipur-Ajmer Express Highway, Dehmi Kalan, Near GVK Toll Plaza, Jaipur, Rajasthan 303007 India

E-mail: geetachhikara@gmail.com

Abstract. Cardiovascular is one of the most critical diseases that affect persons very abominably. Coronary artery diseases (CAD) are one of the categories of cardiovascular diseases that cause a high death rate. So it is imperative to control these death rates by developing an advanced model of machine learning through which diseases can be detected at the premature stage. Due to the lack of enough facilities for tools like angiography it has become a substantial challenge for the health care organization to detect such diseases. If tools exist, then these are known for being expensive and also have numerous side effects. The main goal of this research is to enhance the accuracy of existing models using optimization techniques with machine learning techniques. Alizadeh-Sani CAD dataset has been used which consists of a total of 303 records with 56 attributes. The proposed model reported following values of precision (0.92), recall (0.92), and accuracy (0.93). This proves the efficacy of employing the optimization techniques with machine learning algorithms.

1. Introduction

Coronary artery diseases (CAD) are one of the categories of cardiovascular diseases. Being a critical disease, it is imperative to reduce the number of deaths to rescue people's life. Detection of CAD at a premature stage is only the approach to reduce this death rate. A waxy substance called plague is developed inside the coronary arteries. Coronary arteries perform the work of supplying blood and oxygen to the heart muscles. As time passes, this plague can rupture or harden. It takes several years for the development of the plague. A large number of blood clots sometimes completely block the blood flow in the coronary artery [1]. In this case, heart muscles begin to die and without quick treatment, this can cause a heart attack or even death. Various symptoms are arms pain, stomach pain, nausea feeling dizziness, fatigue, and sweating, etc [1]. Angina is one of the examples of CAD. World health organization (WHO) report says that (CAD is the most common type of cardiovascular disease [2]. Around 30 % of deaths take place due to CAD all around the world. In America, around 360,000 death occurs from heart attacks [3].

In this research, work attention has been paid to the prediction of CAD using optimization techniques over machine learning classification techniques which shows extraordinary results. Machine learning (ML) is an emerging field where a large amount of massive data exists which is also a big challenge. Through ML techniques this unstructured data can be converted into structured data. It is a prominent step to find a better quality of data to analyse the data efficiently. An automatic decision support system can be developed to improve the accuracy of the CAD prediction model. Numerous combinations of

algorithms can be seen in surveys where distinct CAD prediction models emphasize. The very first is the machine learning classification models. Second, feature extraction algorithms over machine learning (ML) algorithms. The third one is the combinations of ML algorithms. Fourth is the Hybridization of feature extraction and the last is the use of optimization techniques. There is a need to work with optimization techniques over ML techniques where limited work has been done [4].

In proposed model, PCA purpose is to reduce the dimensions which is very first step applied to the dataset, which improve learning accuracy and remove irrelevant data. In second step, Firefly Optimization (FA) is used to select the important characteristics of the data set. Best subset of characteristics selected which improves classification accuracy. Classification methods are applied in the third step to diagnose CAD and measure the classification accuracy to evaluate the performance [5][6].

The main objective of this research is the diagnosis of CAD using different efficient predictive classification algorithms decision tree optimized by firefly optimization in Anaconda tool is used with Pycharm tool to analyze data and perform predictions. In the proposed approach the decision tree is used because it's an efficient hierarchical structure which deals with the pruning process to reduce the unnecessary data and perform meaningful insight for the classification process. It works in the hierarchical structure which divides the data into various subsets which reduces the classification time also and performs depth classification. The main contributions of this paper are [7][8]:

- Extraction of classified accuracy useful for CAD prediction.
- Using the feature extraction algorithm to remove redundant and irrelevant features with the principal component analysis (PCA) method.
- Using optimizations with Firefly swarm optimization over machine learning algorithm.
- Comparison of among optimization & without optimization on the dataset.
- Perform efficient CAD predictive modelling.

The remainder of this paper is organized as follows. Recent work in this area is discussed in Section II. Section III describes the detailed description of the proposed methodology. Section IV explains the algorithm steps for the proposed approach. Section V explains in detail the experiments using the proposed machine learning model. Finally, Section VI presents conclusions and future research directions.

2. Related Work

The objective of the paper [9] is to reveal important features that can enhance the accuracy in the prediction of heart diseases with 88.7 % accuracy. Cleaveland dataset is used with the R-studio tool. Numerous combinations of parameters with machine learning techniques (HRFLM approach (Merge Random Forest (RF) and Linear Method (LM)) is used. In the future author can go ahead with feature selection which can help improve the accuracy of the model. The author [10] increases the accuracy of the weak classifier by using ensemble techniques. Around 7 % improvement can be noticed with the help of ensemble learning techniques. Cleveland dataset is used with the WEKA tool. Various techniques like Bayes Net, SVM, C4.5, Multilayer Perceptron, Naive Bayes, and PART with ensemble algorithms have been used. In the future, different techniques can be used with ensemble learning algorithms to improve accuracy. Classifying patient with the groups of 333 previous observations with the presence or absence of cardiovascular diseases by using Naïve Bayes, Random Forests, k-NN, Support Vector Machine Logistic regression, and Deep Learning techniques is also proposed by the authors [11] in their recent works. Cleveland dataset is used with the rapid miner tool. Logistic regression shows the best accuracy. In the future author can go ahead by adding more attributes. Frequent item sets achieved by authors [12] depend on the selected symptoms and minimal support value. Authors predicted the level of risk of the patients have cardiovascular disease over frequent item sets and keep away from unnecessary item set generation in each frequent item set. A simulated heart

disease dataset is used and implemented in java. Random forest algorithm by the author [13] is used with 98% accuracy which predicts cardiovascular with less number of attributes. Cleveland dataset is used with spark framework. Prediction of various cancers at early stages may be the direction of the future. The author [14] proposed a model by combining various techniques like Naïve Bayes, neural network, into one technique with Cleveland dataset which helps health care professionals in CAD detection at premature stage where Naïve Bayes and SVM shows the best accuracy around 89.2%. In the future author can go ahead in proposing a better remote cardiovascular disease detection model that detect cardiovascular disease depend on the patient data. The author [15] emphasis on reducing the features and feature selection enhancement with the help of the KNN algorithm and Imperialist competitive algorithm along with a meta-heuristic approach using the WEKA tool. The author designs an automatic heart disease system that categorizes the patients and suitable to use in clinics. Limited feature selection affects the approach calculation efficiency have a preference In future, author may apply the feature selection algorithm on missing data. The authors [16] detect heart disease using classification algorithms optimized by Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) approaches. Optimization of the algorithm improves accuracy level of prediction. In future, work may done on the knowledge base and tools used in this study. Weka tool used on heart diseases dataset. The authors [17] studied model of machine learning for predicting CAD was compared to classical techniques of machine learning like Decision Tree, PSO, Genetic Algorithm, SVM, Neural Networks Genetic, and Bayesian Net. The Z-Alizadeh-Sani dataset was used with the MATLAB tool. The N2Genetic-SVM algorithm shows the best accuracy of 93.08%. In paper [18] authors improve the performance of coronary heart disease detection after choosing prominent predictive features along with their rank order using Decision tree of Chi-squared automatic interaction detection (CHAID), Random trees C5 and SVM techniques. Various techniques show the accuracy like this SVM-80.90%, CHAID-82.30%, C5.0-83.00%, RTs- 96.70%. In future Fuzzy intelligent system with artificial intelligence, models can be developed to detect CAD with different datasets.

3. Proposed Model

The proposed model deals with the classification process which is used to detect the disease and normal category based on heart diseases. The proposed model can achieve a high accuracy rate with low classification or loss functions to achieve high true positive and negative rates.

The very first step is to create a virtual environment (VE). The VR will reduce the redundancies and dependencies among different libraries which can affect the simulation environment. The next step is to load the data and the scaling of the data which is one of the most important steps in the normalization of the data. If the data is not normalized then the variances or standard deviations among the data points can increases which can create problems during the testing process. After data normalization, the next step is the feature extraction process which is applied over the normalized data after scaling which is used to extract prominent features without losing much actual value. To improve the efficiency in the developing model, an optimization method i.e., Firefly optimization is used which is used to select the relevant instances among the redundant features as it is mostly used to solve complex problems and associated with nature inspiration. Maximum research is being done on PCA but the most efficacious so far is the use of optimization methods over ML techniques for the prediction of CAD. The final step is to train the model is the partitioning of the data into two parts training and testing to analyze how the model works while using ML performing ML classification. 10 fold cross-validation was applied to check the correctness of the proposed model. Amazing results proved that the CAD detection model can assist the health care professional to detect CAD at a premature stage. The proposed model is shown in figure 1.

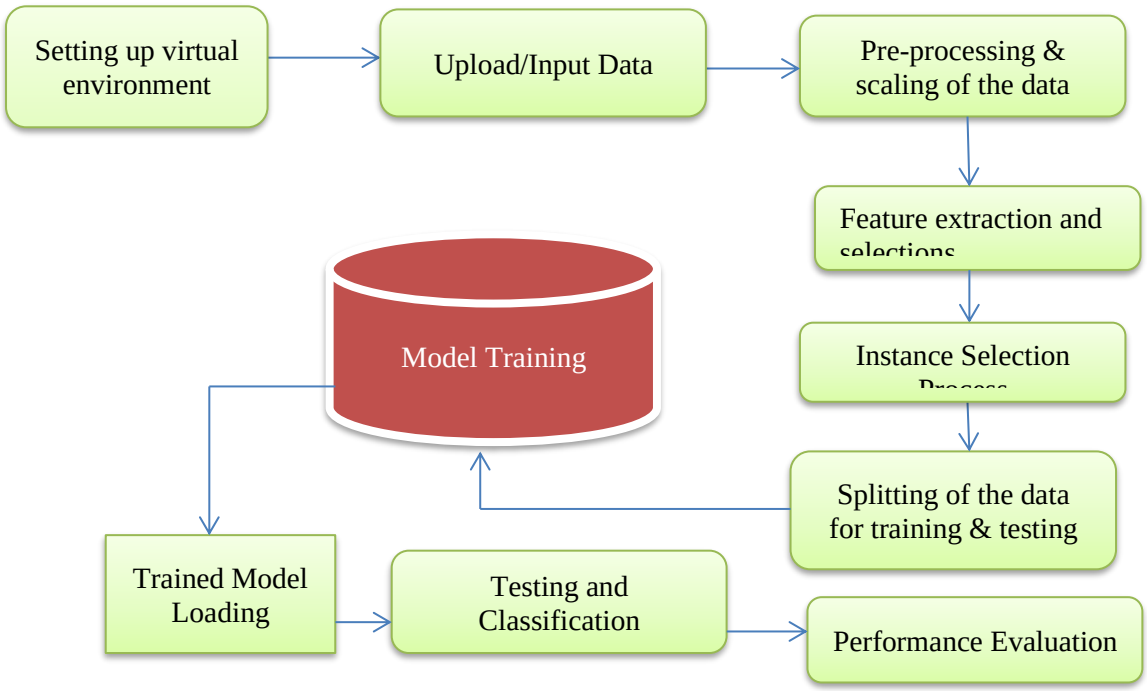

Figure 1. Proposed Model

4. Proposed Algorithm

The proposed algorithm steps are listed below:

Step 1: Initialize VE where VE is the virtual environment to remove the dependencies of the libraries & packages.
Step 2: Input Data such that D = D1, D2… DN as data frame columns.
Step 3: Normalize & scale the data such that
For i=1 to N
ND = Norm {DN} to reduce the variances among data points.
EndFor
Where N is the total number of data points
Step 4: Generate the feature extraction F(x) & transform to generate the covariance process such that
For i=1 to N
C(x) = COV (ND)
EndFor
Step 5: Extract the Eigenvalues and vectors to extract meaningful insights for the transformation T = X W for the data mapping (M) to generate a new space vector E(x).
Where E(v) where V = V1, V2… VN are the processing vector values.
Step 6: Optimize the feature extraction using instance selection for the dimensionality reduction process
Step 7: Splitting of the E(x) into train and test set.
Step 8: Upload Test Image in such a way that TS = TS1, TS2, TS3, TS4… TSN
Step 9: Load the trained model and perform classification on TSN.
Step 10: Evaluate the performance analysis in terms of Precision, Recall & Accuracy Rate.

5. Results & Discussions

The proposed work simulation is implemented using a virtual environment in the python language using the Pycharm editor tool. The Pycharm is a widely useful tool for the simulation of the real-time

environment by taking care of all dependencies and will not affect the simulation environments. The simulation obtained results are given below. For the classification of the CAD, the decision tree classifier is used. As the tree is the well-known data structure for the classification, the decision tree performs the bagging of the data to obtain high classification rates.

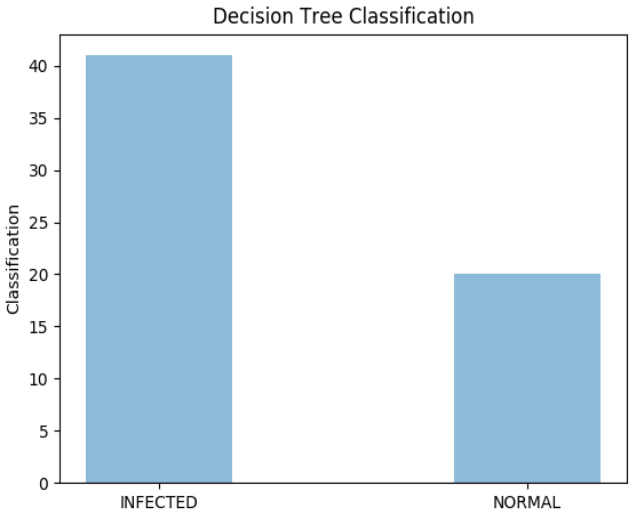

Figure 2. Classification Count

Figure 2 shows the classification count of the detections performed for the health monitoring of the patients. The fig shows that how many persons from the unknown data is classified as the Infected and Disinfected. It can be noticed from fig 2 that the proposed classifier can obtain infected persons from the training of the data with low losses which can help doctors in real-time to categorize based on certain characteristics.

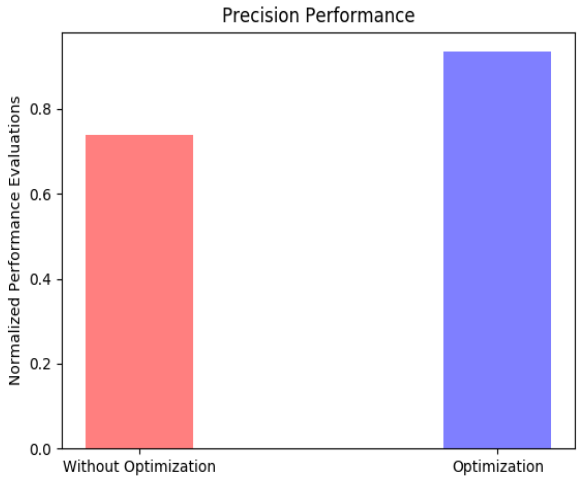

Figure 3. Precision Performance Evaluation

Figure 3 shows the precision performance evaluation without optimization process and with optimization. The optimization process means to say the instance selections. In the proposed approach the performance comparison needs to be done so that the model performance can be analyzed and evaluated. It can be seen that model is performing well in case of instance selection but not that much

which is performed using the feature optimization process with high precision as shown in fig. 3. This can be due to over fitting of the data which means that the model performs well in the training process but doesn't perform in the testing i.e. on the unknown data.

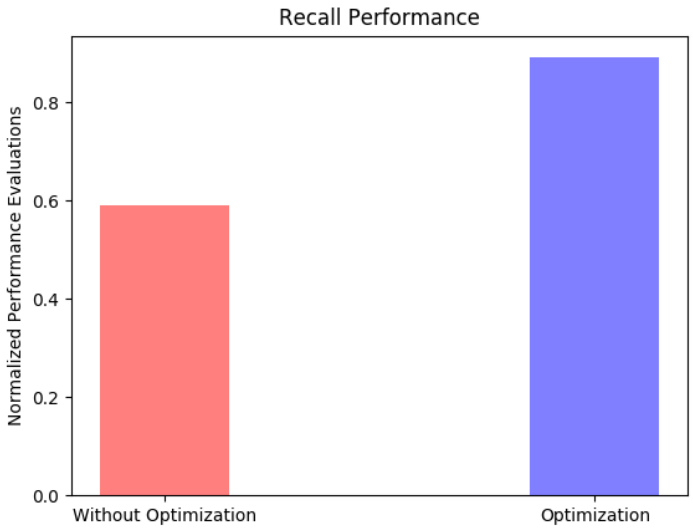

Figure 4. Recall Performance Evaluation

Figure 4 shows the performance evaluation using the instance selection process based on true predictions using recall parameter. It can be noticed that the proposed approach can achieve high recall i.e. 91.84% for the true predictions in case of optimization process which is not in case of without applying any optimization. Also, the precision and recalls are achieved which show the relevant data predictions based on the efficient model training for the classification of CAD i.e. Cardio Vascular Disease. The precision and recall must be high which shows that our proposed machine learning model is well suited to perform high predictive modelling for the true classifications.

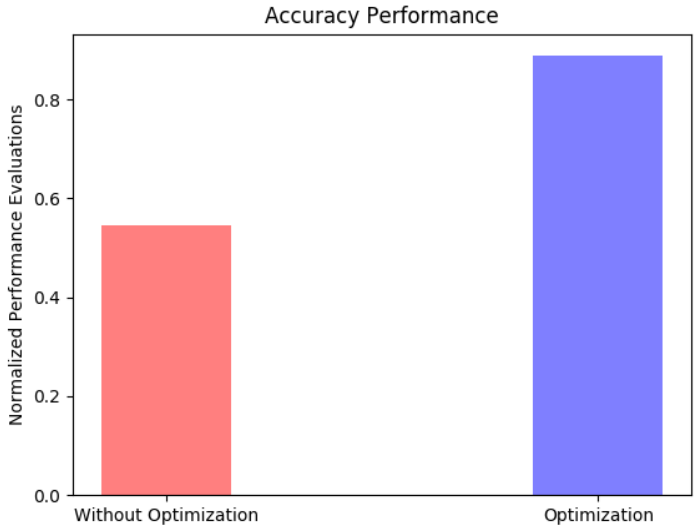

Figure 5. Accuracy Performance Evaluation

Figure 5 shows accuracy performance among the prediction using optimization process and without instance selection process and it can be seen that the true positive and negative rates are achieving high performance in terms of true predictions among the CAD process. It can be seen that the proposed approach is attaining high accuracy which reduce the redundancies among the data to achieve high performance of the system.

Table 1. Performance Evaluation (Without Optimization)

Test Sequence	Recall	Precision	Accuracy
1	0.59	0.72	0.62
2	0.64	0.73	0.64
3	0.62	0.70	0.70
4	0.64	0.74	0.64
5	0.58	0.71	0.71
6	0.65	0.79	0.65
7	0.68	0.77	0.67
8	0.71	0.78	0.62
9	0.74	0.80	0.60
10	0.69	0.74	0.70

Table 2. Performance Evaluation (With Optimization)

Test Sequence	Recall	Precision	Accuracy
1	0.89	0.91	0.92
2	0.90	0.89	0.91
3	0.91	0.92	0.90
4	0.92	0.93	0.93
5	0.88	0.90	0.89
6	0.93	0.92	0.90
7	0.91	0.92	0.93
8	0.88	0.90	0.91
9	0.90	0.91	092
10	0.92	0.92	0.91

Table 1 and 2 shows the performance evaluation of the proposed system which is evaluated in two phases. The very first phase is the extraction and classifications and the second phase is the extraction, instance selections, and classifications and it can be seen that the proposed extraction with instance selections which is well known by the feature optimization is performing well and achieve 93% of classification accuracy with low classification error rate to perform true predictions and low false positive and negative rates.

6. Conclusion and Future Scope

Heart diseases are responsible for high death rates these days. In this research, the authors proposed the efficient and optimized machine learning model for automatic diagnosis of the heart disease. The model is hybrid of PCA, decision tree, and firefly optimization techniques. The PCA algorithm is employed for the feature extraction, nature-inspired firefly optimization technique is used for optimizing the feature selection and decision tree algorithms for classifications. They trained their model using the Alizadeh-Sani dataset. The model is effective in classifying the CAD patients. To prove the importance of optimization techniques, the authors compared the results obtained without employing optimization and on employing the optimization techniques. The comparison shows that the use of firefly optimization technique improves the accuracy of classification and lowers the error rates. Thus, the machine learning models with firefly optimization technique may prove effective in early diagnosis of cardiovascular artery disease. This may become a life-saving tool.

7. References

[1]. Hazra A., Mandal S. K., Gupta A., Mukherjee A., "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", *Advances in Computational Sciences and Technology* ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2137-2159.

[2]. Wah, T.Y.; Gopal Raj, R.; Iqbal, U. Automated diagnosis of coronary artery disease: A review and workflow. *Cardiol. Res. Pract.* 2018, 2018, 2016282.

[3]. Wang, Y.; Kung, L.; Gupta, S.; Ozdemir, S. Leveraging big data analytics to improve quality of care in healthcare organizations: A configurational perspective. *Br. J. Manag*. 2019, 30, 362–388.

[4]. Khourdifi, Y., & Bahaj, M. (2019), Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *Int. J. Intell. Eng. Syst.*, 12(1), 242-252.

[5]. Alizadehsani, R., Khosravi, A., Roshanzamir, M., Abdar, M., Sarrafzadegan, N., Shafie, D & Bishara, A. (2020). Coronary Artery Disease Detection Using Artificial Intelligence Techniques: A Survey of Trends, Geographical Differences, and Diagnostic Features 1991-2020. *Computers in Biology and Medicine*, 104095.

[6]. Setiawan, N. A., Venkatachalam, P. A., & Hani, A. F. M. (2020), Diagnosis of coronary artery disease using the artificial intelligence-based decision support system. *arXiv preprint arXiv*:2007.02854.

[7]. Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020), Decision tree-based diagnosis of coronary artery disease: CART model. *Computer Methods and Programs in Biomedicine*, 192, 105400.

[8]. Kolukisa, B., Yavuz, L., Soran, A., Bakir-Gungor, B., Tuncer, D., Onen, A., & Gungor, V. C. (2020), Coronary Artery Disease Diagnosis Using Optimized Adaptive Ensemble Machine Learning Algorithm. *International Journal of Bioscience, Biochemistry, and Bioinformatics*, 10(1).

[9]. Mohan, S., Thirumalai, C., & Srivastava, G. (2019), Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, 81542-81554.

[10]. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, 100203.

[11]. Khateeb, N., & Usman, M. (2017, December), Efficient heart disease prediction system using K-nearest neighbor classification technique. *In Proceedings of the International Conference on Big Data and Internet of Thing* (pp. 21-26).

[12]. Pereira, Tania, et al. "Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation." *Physiological measurement* 40.12 (2019): 125002.

[13]. Saboji, R. G. (2017, August), A scalable solution for heart disease prediction using classification mining technique. *In 2017 International Conference on Energy, Communication, Data Analytics, and Soft Computing* (ICECDS) (pp. 1780-1785). IEEE.

[14]. Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia computer science*, 120, 588-593.

[15]. Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M. R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y., & Tayebi, R. M. (2019). A new hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health and Technology*, 1-12.
[16]. Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *Int. J. Intell. Eng. Syst.*, 12(1), 242-252.
[17]. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, 179, 104992
[18]. Joloudari, Javad Hassannataj, et al. "Coronary artery disease diagnosis; ranking the significant features using a random trees model." *International journal of environmental research and public health* 17.3 (2020): 731.