EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

Analyzing Machine Learning Classifiers for the Diagnosis of Heart Disease

Saravanan Thangavel[1]*, Saravanakumar Selvaraj[2] Ganesh Karthikeyan V[3], and K Keerthika[4]

[1]Dept. of CSE, GITAM School of Technology, GITAM (Deemed to be University), Bengaluru, India
[2]Dept. of CSE, Faculty of Engineering and Technology, Jain (Deemed to be University), Bengaluru, India
[3]School of Computing, SASTRA Deemed University, Thanjavur, India
[4]Dept. of Computer Science, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Mysuru, India

Abstract

INTRODUCTION: Preventable deaths from cardiovascular diseases outnumber all others combined. Detecting it at an early stage is crucial. Human lives will be saved as a result.
OBJECTIVES: Improved cardiac disease prediction using machine learning classifiers is the focus of this article.
METHODS: We have used many different classifiers, such as the support vector machine, naive bayes, random forest, and k-nearest neighbours, to achieve this goal, even though we can't predict high accuracy in this classifier. So, we have proposed Hyper parameter adjustment was applied to the classifiers, which increased their precision. It was possible to compare the classifiers.
RESULTS: In comparison to other machine learning classifiers, Logistic Regression achieves higher prediction accuracy, at 95.5%.
CONCLUSION: To help people find the nearest cardiac care facilities, Google Maps has been integrated into a responsive web application that has been built for forecasting heart illness.

Keywords: Breast Heart disease, Edge detection Classification, Human intelligence, Segmentation

Received on 04 December 2023, accepted on 20 February 2024, published on 29 February 2024

doi: 10.4108/eetpht.10.5244

1. Introduction

Machine learning is "a technique for automatically acquiring meaningful knowledge from data by analyzing and learning from examples presented to it" [1]. Machine learning is an expansive field that is continually expanding. Supervised learning, unsupervised learning, and ensemble learning are all forms of machine learning that can improve the precision of your predictions and analyses. We can put this knowledge to good use in our HDPS research now.
Cardiovascular diseases encompass a wide spectrum of illnesses that might have an impact on the heart yet are all too common in modern society. Globally, CVDs are responsible for an estimated 17.9 million deaths in 2015 [2]. The majority of adult deaths can be attributed to this. By analyzing a person's medical records, our approach can identify those who are at high risk of being diagnosed with cardiovascular disease [3]. It helps doctors pinpoint patients having cardiac symptoms like chest pain or high blood pressure so they may administer a more precise diagnosis and more targeted treatment. Regression analysis, k-nearest neighbour, and random forest classifier are three data mining methods that play a significant role in this study. Our research has an accuracy of 87.5%, which is an improvement over systems that just used a single data mining technique. As a result, the HDPS's precision and productivity went up after we included more data mining procedures. The field of study known as logistic regression

*Corresponding author. Email: tsaravcse@gmail.com is classified as supervised learning. In order to do logistic regression, only discrete values are allowed.

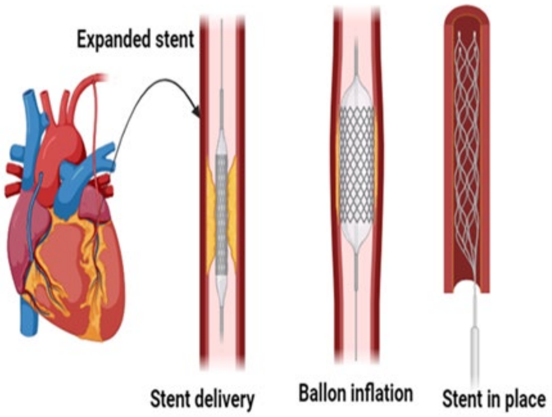

Figure 1. Heart disease structural model

The purpose of this study is to analyze demographic and clinical factors to predict whether or not a patient would acquire cardiovascular disease as shown in Fig. 1. A dataset containing medical records and patient information is sought out from the UCI repository. Using this data, we may determine the likelihood that a given patient suffers from cardiovascular disease. We use a patient's 14 medical characteristics to forecast whether or not he will develop heart disease. As a matter of fact, you have three options here: Regression analysis, k-nearest neighbours, and the random forest classifier are used to hone these clinical characteristics. With an accuracy of 88.52 percent, KNN is the most accurate algorithm to date. We can also classify individuals according to their potential for developing cardiovascular disease in a time- and labour-saving manner.

2. Related Works

The doctors relied heavily on auscultation to discern between regular and aberrant heart sounds [4]. Doctors listened to the heart sounds with stethoscopes and were able to diagnose all forms of heart illness [5]. There are limitations to the auscultation technique used by medical professionals to identify cardiac problems. Long, detailed examinations equip doctors with the knowledge and experience necessary to identify and categorize cardiac sounds [6].

Multiple machine learning strategies have been offered as alternatives to manual CVD detection. Daliri et al. [7] did a study to identify the best characteristics for predicting cardiovascular disease. NB, K - nearest neighbours, LR, DT, NN, SVM, and Vote are just some of the seven methods utilized for data classification. The Cleveland machine learning datasets were retrieved from the UCI collection. This database contains 303 records and 76 attributes. In order to train and test models, we use 10-fold cross-validation. Rather than using a train-test split, which would be more appropriate for a dataset with more training examples than test instances, we utilized 1-fold cross-validation to assess the model's accuracy. The ten-fold validation method guarantees that 90% of the data is available to your model. As of right now, the Vote Classifier is 87.4% accurate. Dwivedi et al. [8] used a KNN classifier with few parameters to successfully predict the existence of cardiac disease. KNN is computationally expensive due to its use of 90% of the input for training, and its performance falls as the number of parameters grows. The research was done by Polat et al. [9] of the scientific community. Predictions of cardiac issues are made using a backpropagation-trained perceptron neural network. The Cleveland dataset, available in the UCI machine learning repository, is used for both training and testing models. It contains data on 303 occurrences and 76 attributes. Six records were processed to remove missing values and 14 heart disease characteristics were used after a rough preprocessing stage. The experimental results showed that MLN-NN improved accuracy by 93.39 percent while only taking 3.86 seconds to complete. Liu et al. [10] did an extensive analysis with some of the most common machine learning classifiers to predict cardiac disease. Just 14 out of the 303 features available in the Cleveland (UCI) datasets are used during training and testing. A dataset with 296 records was produced after some data preprocessing work was done.

The accuracy of SVM classifiers' output increased to 90.00%. In order to foresee cardiovascular disease, Anouncia al. [11] used a mixed approach to data mining classifiers. UCI's machine learning dataset repository provided the 303 records and 76 attributes used in this study. Model training and evaluation involved 14 different qualities. Preprocessing the +e data brought the total number of features down from 14 to 12. The classification algorithms KNN, NN, SVM, GA, J48, RF, and NB are method for evaluating the accuracy of heart illness prediction by recall and exactitude. When it came to predicting cardiac abnormalities, SVM and NB were more reliable (89.2% accuracy). The purpose of the research of Anitha and colleagues is to develop a method for predicting cardiovascular disease. [12], which used learning vector quantization algorithms. This method has an accuracy of 85.5%. Sets of data consist of 303 records and 76 attributes, and were taken from the UCI artificial intelligence toolkit. Due to missing values, the +e dataset was pre-processed, resulting in a 302-record sample from which 14 characteristics were employed to detect cardiac illness. The dataset is split into a training portion (70%) and a testing portion (30%). Muthukaruppan et al. [13] developed another machine learning study for cardiac illness prediction. In both development and testing, we employ LR, NB, and SVM as classification methods of choice. With the help of the UCI machine learning repository, the Cleveland datasets were divided into a 75%-25% training and testing division.

·The data accuracy increased to 66.2% because to SVM's abilities to remove outliers and substitute missing values. The study's lack of sensitivity to early risk variables in human heart disease patients was a significant limitation.

Ganji et al. [14] developed a method that combines optimizing particle swarms and radial basis function algorithms for classification to provide accurate cardiac disease predictions. The VA Long Beach machine learning dataset from the UCI repository was utilized for training and testing the models, and it has 270 records and 14 attributes. However, only 7 of these features were employed to predict heart disease. When PSO is used in conjunction with NB, NB's performance accuracy rises to 87.91 percent. An increase in precision of 8.7 percent relative to NB precision has been demonstrated. Separately, Ozcift et al. [15] developed a study that also predicted instances of heart disease using machine learning techniques.

The dataset's 303 records and 14 attributes originated from UC Irvine's machine learning repository (UCI). The back propagation bridge method is used in both the training and the testing phases.

The DT algorithm has a 93.19 percent success rate in predicting cardiac problems. In their article, "artificial intelligence enhanced electrocardiogram for detection of heart disease in High-Risk Populations," Zou et al. [16] review the value of AI-enhanced ECG for healthcare decisions in patients with cardiovascular ailment, evaluate its current and future condition, and examine its potential downsides. To aid those suffering from heart disease, Wang et al. [17] introduced a novel health information system. Their preliminary results suggest that doctors have trouble deciding how much exercise to recommend.

3. The Proposed Methodology

The Fig. 2 illustrates the proposed work done as follows

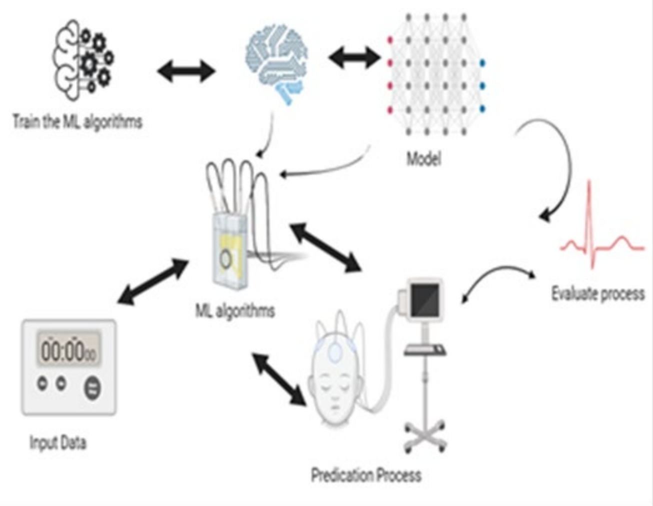

Figure 2. Heart disease proposed diagram

3.1. Support Vector Machine

The support vector machine is a common supervised learning technique used to address classification and regression issues. It is widely used in the field of machine learning, especially when attempting to resolve categorization issues. The SVM technique may quickly identify upcoming data points by finding the best line or decision boundary for splitting an n-dimensional space into classes. One of the most optimal decision boundaries is a hyper-plane. Maximum hyper plane-constructing vectors are selected with support vector machines. This technique, known as a Support Vector Machine, is best illustrated by the use of support vectors.

3.2. Working of SVM

It is possible to use an example to show how the SVM algorithm functions. Let's pretend we have a data set with two groups and two characteristics: a blue group representing depressed individuals and a black group representing healthy individuals who are not depressed (x1 and x2). In order to tell the difference between black and blue locations, we need a classifier [Fig. 3]. Since this is a two-dimensional set, we can easily demarcate the two categories by drawing a line down the middle. To divide these groups, one can take a variety of different approaches, as shown in Fig. 3. Therefore, the best line or decision boundary, also called a hyper plane, can be identified using the SVM method. The Svm classifier involves looking for a point where the lines (representing the two classes) meet. It is possible to zero in on these regions with the help of support vector machines. The vectors' margin is the space they use beyond the hyper plane. The reason for implementing SVM is to maximise this profit margin. As can be seen in Fig. 3, the hyper plane with the biggest margin is the optimal choice. [18][19].

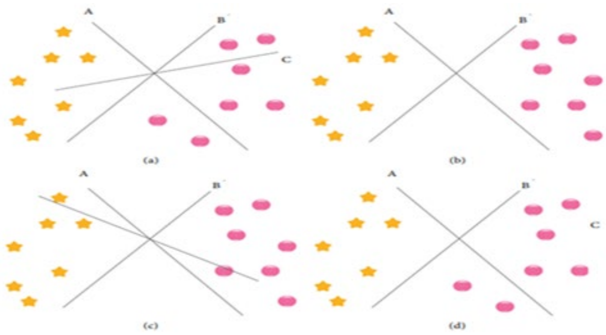

Figure 3. Heart disease proposed diagram

The HD pseudo code, algorithm 2.

```
HDD, the Heart Disease Dataset, is the input.
Step 1: Determine the K-nearest neighbor samples (ksi) of each sample sid in the HDD minority class.
Step 2: A new sample is built using the formula nsd = sid + (sbidsid) + ;
```

Step 3: Then, for each sample hi in HDD, if hi's class is less common than the class of HDD's k-nearest neighbours, remove hi;
Step 4: finally, append the resulting sample ns to HDD.
Step 5: Output: Balanced dataset HDD

3.3. Random Forest

For regression and classification problems, random forest is a supervised machine learning technique. It employs averaging for regression and majority vote for classification to produce decision trees from multiple data. The random forest algorithm's versatility in dealing with large-scale categorical and continuous datasets, such as those encountered in classification and regression problems, is one of its most valuable qualities. When compared to similar products, it provides superior performance in terms of categorization accuracy [20]. Ensemble learning, on which random forest is built, is a method for improving the performance of a model by merging many classifiers to tackle difficult issues [21].

3.4. Working of Random Forest

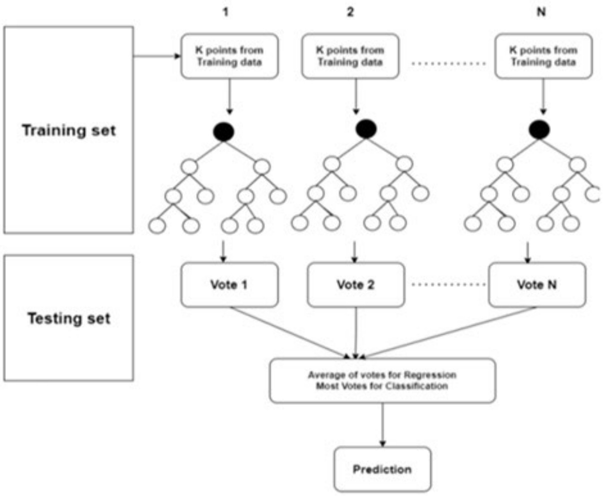
1
2
N
Training set
K points from Training data
K points from Training data
K points from Training data
Testing set
Vote 1
Vote 2
Vote N
Average of votes for Regression
Most Votes for Classification
Prediction

Figure 4. Process of a Random Forest, as depicted

This Fig. 4 elaborates the training set, from the training data which then processed, classified and then predicted.

3.5. Bayes Naives

We use the naive bayes technique of pattern recognition for categorization jobs. Bayes' rule gives support for this theory. Despite its apparent simplicity, this technique of machine learning has broad use across a variety of industries. The naive bayes classifier excels at learning large data sets with hundreds of variables and millions of data points [22]. This type of method is extremely quick, especially in comparison to its forerunners.

3.6. Naive Bayes' Theorem in Action

It is commonly used for large-scale training dataset-required applications like grouping of texts for analysis. When it comes to building predictive machine learning models, the Classification Algorithm is a fast and effective classification algorithm. Since it is based on probabilities, it predicts the likelihood of different states for objects. Probability theory is a theoretical technique for testing assertions given existing evidence is known as the Bayes theorem, Bayes' rule, or Bayes' law. In this case, we use conditional probability to make a call. [14] [15]. The Bayes theorem can be expressed in the form of a formula, as shown below.

PT (ba/bX) = PT (ba/ya) * PT (ba) / P(ba) (1)

Replace ya with the class variable you'd like to use.
ba = Feature space that relies on a reliant attribute (ba= a1 ,a2, a3... a4)
Given a range over all possible values of the class variable ba, we need only pick the output with the highest sampling frequency.

ba = argmaxiyP(ba) Πni=1P (abi | ba) (2)

The initial phase of the 3 parts of the naive bayes classifier requires integrating the information into tables and charts. The probability of the characteristics that have been given is then evaluated, yielding a Possibility list. Bayesian' method is used to calculate the prediction error as a last step [23].

4. Experiments

Many academics have used Kaggle 102 H&E to conduct studies of a similar nature. This data set includes both healthy and cancerous examples of the human body. Careful partitioning of the dataset resulted in the generation of validation and testing datasets with identical distributions, which accurately reflected the model's generalizability. To optimize the model's output, hyperparameters like the learning rate and decay are tweaked in response to training and validation data, respectively. Training data is essential for learning indicators like weights and biases. The completion of a model is the result of careful analysis of the test results. As a whole, the image needs to be normalized so that all of the pixels are in the same ballpark and any semblance of bias has been removed. From 102 HDSI mounted scanned samples, approximately 36 x 36 pixels RGB digital image patches were created. Table 1 enumerates the values of ML Classifiers using Proposed Work. Figure 5 illustrates the Heart disease Machine Learning Classifiers and Figure 6 exposes the SVM-Naive Bayes –KNN-RF.

· Table 1. ML Classifiers using Proposed Work

ML Classifiers	Precision	Recall	F1-score	Accuracy
SVM-Navies Bayes	82.15	73.03	85.67	87.07
SVM-Navies Bayes -KNN	86.50	79.34	86.5	90.2
SVM-Navies Bayes –KNN-RF	90.05	89.56	88.7	93.34

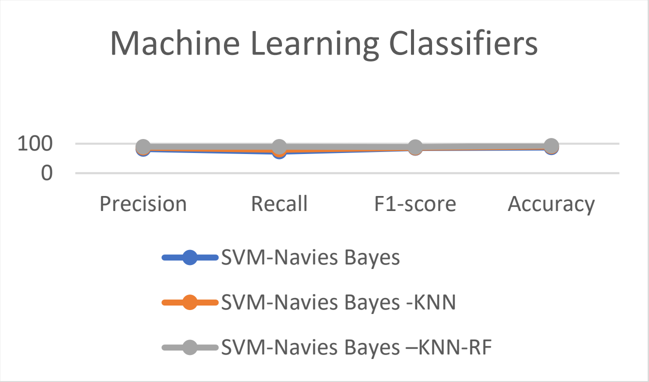

Figure 5. Heart disease Machine Learning Classifiers

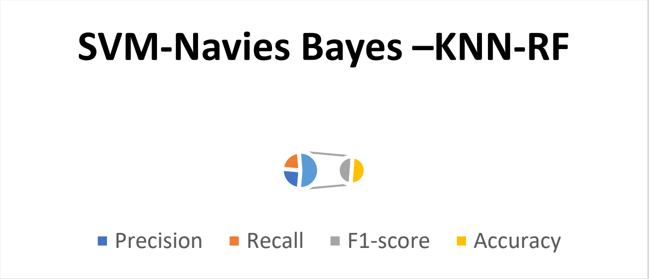

Figure 6. SVM-Navies Bayes –KNN-RF

5. Conclusion

Three machine learning (ML) classification modelling strategies were used to create a cardiovascular disease detection model. Patients' medical records, including those for chest discomfort, blood sugar, blood pressure, and so on, are included in the dataset from which this study draws its predictions about people with cardiovascular disease. If a patient has a history of cardiac illness, this approach can help determine if further testing is necessary. Navies bayes, Random Forest Classifier, and SVM-Naive Bayes –KNN-RF are the methods utilized to construct this particular model. Our model has a reliability of 93.34%. Increasing the amount of training data increases the likelihood that the model will correctly identify whether or not a given individual has heart disease. These computer-aided methods enable us to make more accurate, timely, and affordable patient predictions. Due to the superiority of Machine Learning techniques over human prediction, we may apply them to a variety of medical datasets that benefit both patients and medical professionals. Since this project involves cleansing the dataset and applying logistic regression and KNN, we were able to improve upon the accuracy of earlier models (90%) and provide more accurate predictions for patients who have been diagnosed with heart problems. In addition, we find that SVM-Naive Bayes –KNN-RF has the highest accuracy (93.34%) of the three algorithms we tested. In the dataset, heart disease affects 7% of the patients, as seen in Figure 6.

Acknowledgements.

I thank the coauthors for the research work done and there was no funding received for this research.

References

[1] Rehan Ahmed, Maria Bibi, Sibtain Syed. Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms. International Journal of Computations, Information and Manufacturing. 2023; Vol. 3.1, pp 49-54.

[2] Jayachitra S, Aruchamy, P, Lebbe A:. An efficient clinical support system for heart disease prediction using TANFIS classifier. Computational Intelligence. 2022; Vol. 38:pp.610-640.

[3] Balasubramaniam S, Kumar, K, Kavitha, V: Feature Selection and Dwarf Mongoose Optimization Enabled Deep Learning for Heart Disease Detection.Computational intelligence and neuroscience. 2022; Vol. 2022:pp.1-18.

[4] Kavitha, M, Roobini, S, Systematic View and Impact of Artificial Intelligence in Smart Healthcare Systems, Principles, Challenges and Applications, Machine Learning and Artificial Intelligence in Healthcare Systems. 2023; 25-56.

[5] Ghulab Nabi Ahamad, Shafiullah, Hira Fatima, Imdadullah, S. M. Zakariya, Mohamed Abbas, Mohammed S. Alqahtani and Mohammed Usman.: Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease. Processes. 2023; Vol. 11(3), pp. 734.

[6] Khondokar Oliullah, Alistair Barros, Md. Whaiduzzaman. Analyzing the Effectiveness of Several Machine Learning Methods for Heart Attack Prediction. Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering. 2023; Vol. 618, pp. 225-236.

[7] Mohammad Reza Daliri. Automatic diagnosis of neuro-degenerative diseases using gait dynamics. Measurement. 2012; Vol. 45, pp. 1729–1734.

[8] Karnika Dwivedi, Hari Om Sharan, Vinod Vishwakarma. Analysis of decision tree for diabetes prediction. International Journal of Engineering and Technical Research. 2019; Vol. 9, pp. 3-6.

[9] Kemal Polat, Salih Güneş. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing. 2007; Vol. 17, pp. 702–710.

[10] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, Kevin Murphy. Progressive neural architecture search. In

Proceedings of the European conference on computer vision. 2018; pp. 19–34.

[11] Margret Anouncia S, Clara Madonna L. J, Jeevitha P, Nandhini R. T. Design of a diabetic diagnosis system using rough sets. Cybernetics and Information Technologies. 2013; Vol. 13(3), pp. 124–139.

[12] Peter J. Valdez, Vincent J. Tocco, Phillip E. Savage. A general kinetic model for the hydrothermal liquefaction of microalgae. Bioresource Technology. 2014; Vol. 163, pp. 123–127.

[13] S. Muthukaruppan, M.J. Er. A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications. 2012; Vol. 39, pp. 11657-11665.

[14] Mostafa Fathi Ganji, Mohammad Saniee Abadeh.: A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications. 2011; Vol. 38(12), pp.14650-14659.

[15] Akin Ozcift, Arif Gulten. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Computer Methods and Programs in Biomedicine. 2011; Vol. 104, no. 3, pp. 443–451.

[16] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics. 2018; Vol. 9(515).

[17] Wenbo Wang, Meng Tong, Min Yu. Blood glucose prediction with VMD and LSTM optimized by improved particle swarm optimization. IEEE Access. 2020; Vol. 8, pp. 217908–217916.

[18] Md. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain, Mahmudul Hasan. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access. 2020; Vol. 8, pp. 76516-76531.

[19] Shaksham Kapoor, K Priya. Optimizing hyper parameters for improved diabetes prediction. International Research Journal of Engineering and Technology. 2018; Vol. 5(05).

[20] Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, Hemant Darbari. Prediction of diabetes using artificial neural network approach. In Engineering Vibration, Communication and Information Processing, 2020; pp. 679-687.

[21] T. Santhanam, M.S. Padmavathib. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science. 2015; Vol. 47, pp. 76-83.

[22] Nongyao Nai-arun, Rungruttikarn Moungmai. Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science. 2015; Vol. 69, pp. 132-142.

[23] Aishwarya Mujumdar, V Vaidehi. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019; Vol. 165, pp. 292-299.

[24] Vandana Roy, Prashant Kumar Shukla, Amit Kumar Gupta, Vikas Goel, Piyush Kumar Shukla, Shailja Shukla. Taxonomy on EEG artifacts removal methods, issues, and healthcare applications. Journal of Organizational and End User Computing. 2021; Vol. 33, pp. 19–46.

[25] Geetanjli Khambra, Prashant Shukla. Novel machine learning applications on fly ash based concrete: an overview. Materials Today Proceedings. 2023; Vol. 80(3), pp. 3411-3417.

[26] Prashant Kumar Shukla, Jasminder Kaur Sandhu, Anamika Ahirwar, Deepika Ghai, Priti Maheshwary, Piyush Kumar Shukla. Multiobjective genetic algorithm and convolutional neural network based COVID19 identification in chest X-ray images. Mathematical Problems in Engineering. 2021; Vol. 2021, pp. 1-9.

[27] Neeraj Kumar Rathore, Neelesh Kumar Jain, Prashant Kumar Shukla, UmaShankar Rawat, Rachana Dubey. Image forgery detection using singular value decomposition with some attacks. National Academy Science Letters. 2021; Vol. 44, pp. 331–338.

[28] Manish Agrawal, Asif Ullah Khan, Piyush Kumar Shukla. Stock price prediction using technical indicators: a predictive model using optimal deep learning. International Journal of Recent Technology and Engineering. 2019; Vol. 8(2), pp. 2297–2305.

[29] Vandana Roy, Shailja Shukla, Piyush Kumar Shukla, Paresh Rawat. Gaussian elimination-based novel canonical correlation analysis method for EEG motion artifact removal. Journal of Healthcare Engineering. 2017; Vol. 2017.

[30] Rajendra Gupta, Piyush Kumar Shukla. Performance analysis of antiphishing tools and study of classification data mining algorithms for a novel anti-phishing system. International Journal of Computer Network and Information Security. 2015; Vol. 7, pp. 70–77.

[31] Manish Kumar Ahirwar, Piyush Kumar Shukla, Rakesh Singhai. CBO-IE.: A Data Mining Approach for Healthcare IoT Dataset Using Chaotic Biogeography-Based Optimization and Information Entropy. Scientific Programming. 2021; Vol. 2021, pp.1-4.