

Implementing Stacking with Cross - Validation for Heart Disease Prediction

Yogendra Narayan Prajapati, Dev Baloni

Department of Computer and Engineering, Quantum University, Roorkee, Uttarakhand, India.

Received: 18th May, 2024; Revised: 16th July, 2024; Accepted: 01st August, 2024; Available Online: 31st August, 2024

ABSTRACT

The objective of this work is to apply machine learning techniques for the prediction and early identification of cardiovascular disease, a major worldwide health problem. XGBoost, K-nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), and StackingCVClassifier were among the ensemble algorithms used to anticipate cardiac disease using a dataset that came from the UCI ML database. To improve the quality of the data, an exploratory data analysis was performed on the dataset using methods including outlier identification and missing value imputation. Stacking CV Classifier attained the best accuracy rate of 92.20%, according to a comparative examination of pre-processing and post-processing findings. When compared to previous approaches, the suggested strategies performed better in terms of accuracy, recall, and f1-score. Moreover, the flexibility of the model indicates its possible application to other illnesses with comparable features.

Keywords: Ensemble method, Extreme Gradient Boost, Cardiovascular disease, Stacking CV Classifier, Early detection.

International Journal of Pharmaceutical Quality Assurance (2024); DOI: 10.25258/ijpqa.15.3.62

How to cite this article: Prajapati YN, Balon D. Implementing Stacking with Cross - Validation for Heart Disease Prediction. International Journal of Pharmaceutical Quality Assurance. 2024;15(3):1493-1498.

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

Heart disease is still a major worldwide health risk that claims lives. International healthcare organizations have released statistics reports indicating that 17.9 million fatalities worldwide in 2019 were due to cardiovascular disorders, which make up an astounding 32% of all deaths.¹ Predictions suggest a concerning increase in the number of affected individuals 23 million by 2030, according to estimates. Heart disease and stroke account for a sizable majority of these cardiovascular disease-related deaths 85%. These conditions primarily affect low-income nations, where they cause 85% of disability and 80% of deaths.¹

To drastically reduce the number of premature fatalities, early detection and prediction of heart disease are essential. Heart-related disease risk and progression are influenced by a variety of factors, such as age, changes in lifestyle, dietary habits, and socioeconomic status, which can affect things like access to healthcare facilities.^{2,3} Other risk factors for heart-related issues include high blood pressure, abnormal blood lipid levels, increased glucose, obesity, and overweight disorders.⁴

Investigating artificial intelligence methods could improve the prognosis of heart-related conditions and allow for the early implementation of preventative measures. In particular, machine learning (ML) approaches offer ways to improve healthcare governance and resource efficiency. This allows

for better patient health services, which benefit a range of stakeholders, including patients, practitioners, telemedicine systems, hospital management, and providers.⁵

The objective of this work is to use Ensemble Learning (EL) techniques to create a sophisticated model for the prediction of heart disease. The goal is to improve the prediction model's accuracy and other performance indicators, given how important this application is. New contributions include feature engineering, thorough data preparation, and using boosting methods in an ensemble learning framework to improve model resilience and prediction accuracy.

This article's following sections are arranged as follows: A review of relevant work is provided in Section 2, the technique and dataset are outlined in Section 3, the experimental details and results are presented and analyzed in Section 4, and the conclusion and future directions are provided in Section 5.

Background Work

In a variety of fields, machine learning and ensemble learning approaches have become effective instruments that provide legitimate, dependable, and consistent solutions to real-world issues.^{4,5} A great deal of research has been done using these approaches in the field of illness prediction, namely in the area of cardiovascular disorders.⁶ To improve the field of identifying heart-related illnesses, researchers have investigated a variety of datasets, algorithms, and techniques.^{7,8}

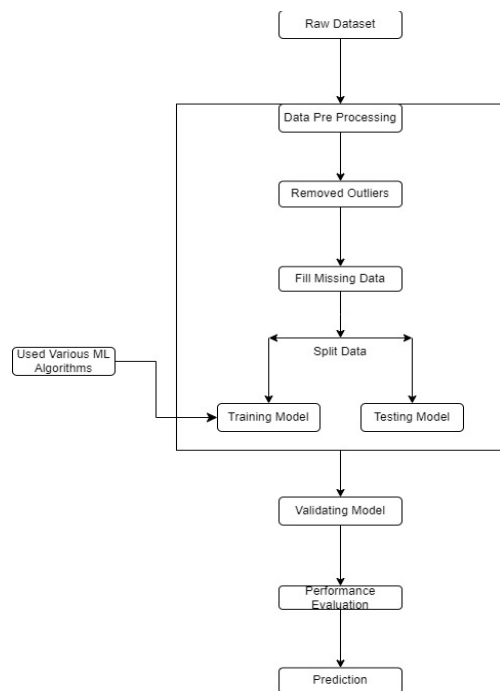


Figure 1: Our proposed implementation work

In order to improve illness risk prediction,⁹ investigated with ensemble approaches such as bagging, boosting, stacking, and majority voting using conventional classifier algorithms. A majority voting approach yielded the greatest accuracy.¹⁰ gradient boosting technique, which takes into account a number of medical indicators, is based on recursive feature reduction and is intended to detect cardiac disease.¹¹ investigated a number of ensemble approaches, including as AdaBoost, GBM, LGBM, XGBoost, and CatBoost; XGBoost produced the best results when it came to coronary disease prediction.

By using Bayesian optimization for hyper-parameter tweaking and one-hot encoding for categorical features,¹² improved the XGBoost classifier for accurate heart disease prediction. A comprehensive investigation of EL methods on a dataset comprising categorical and numerical variables was carried out,¹³ demonstrating the efficacy of combining SVM and AdaBoost for heart disease prediction.¹⁴ created a framework for predicting heart illness by contrasting EL approaches with traditional ML techniques; they found that SVM with boosting produced the best accuracy rate.¹⁵ used ensemble learning to predict heart disease, and bagged classifiers performed better than other classifiers.

Using ten machine learning algorithms,¹⁶ created a model to predict cardiac disease, with SVM producing the highest accuracy. Framework¹⁷ achieved great accuracy using the random forest method by combining the boosting method and neural network approaches. In a comparison of machine learning methods for heart disease prediction,¹⁸ found that a hybrid random forest and linear model produced encouraging findings. Even while ensemble learning has

made significant progress in heart disease prediction, many studies still lack the proper data preparation, normalization, and standardization all of which are essential for enhancing prediction performance.

Proposed Method

The approach used for this experimental work is shown in Figure 1, which also outlines the procedural steps necessary for early illness prediction using different ensemble learning algorithms. Using the flexibility and accessibility of a web-based Jupyter Notebook, a publicly available dataset on heart disease is imported. Python programming is used to install necessary library packages from Sklearn.

To forecast the illness, boosting classifiers are first applied without any data preparation. Further investigation into the data shows that pre-processing is essential for improving prediction accuracy. During the preparation stage, data imputation techniques are used to find and replace missing values, and the interquartile range approach is utilized to find and exchange dataset outliers. In addition, the required libraries are used to locate and fix any faulty data in the dataset.

The dataset is then split in half, with 80% going toward training the model and the remaining 20% going toward testing, or an 80:20 ratio. Results are validated using K-fold cross-validation (K = 10). In order to get the intended prediction results, the pre-processed data is finally reintroduced to the three boosting techniques that were studied: stacking and gradient boosting.

Methods Used

In order to forecast cardiac disease, ensemble learning approaches are investigated in great detail. Three ensemble-learning-based boosting techniques are examined in the study:

Gradient boosting is the sequential training of weak learners with an emphasis on reducing disparities between anticipated and actual values. Gradient boosting involves gradually adding estimators by adjusting weights.

XGBoost

A combination of many decision trees that computes similarity scores on their own. Regularization and gradient descent are adjusted to reduce overfitting.

KNN

Regression and classification using a non-parametric approach. It designates the bulk of its k-nearest neighbors' class.

Stacking CV Classifier

A stacking classifier that readies the input data for the second-level classifier using cross-validation. Through the use of a meta-classifier, it integrates many classification models.

Dataset

A well-known dataset on heart illness is used in the experiment, and it comes from the machine learning repository at the University of California, Irvine. This dataset is a favorite among researchers due to its extensive demographic coverage and wealth of clinical variables pertaining to heart disease.

Table 1: Representation of comprehensive details of a dataset as per column

Parameters	Overview	Assessment	Limits
Age	individual	Years	25–75
Sex	individual	1 = male, 0 = female	0 or 1
Cpericarditis	Level chest pain	Low, moderate, high, extremely high	0–3
RestingBP	Blood pressure of an individual while at rest	mm Hg	94–200
Cholesterol	Level of serum cholesterol	mg/dl	126–564
FastingBP	fasting	1 = true, 0 = false	0 or 1
RestingECG	Resting electrocardiographic results of an individual while inactive	0 = normal, 1 = having ST elevation, 2 = hypertrophy	0–2
MaximumHR	Highest heart disease rate	Beats per minute	71–202
ExerciseIA	angina disease	1 = True, 0 = False	0 or 1
Oldpeak	ST depression	Numeric value	Relative
Slope	When a person is exercising, the slope in the ST segment shows the previous peak value.	0 = downsloping, 1 = flat, 2 = upsloping	0–2
Ca	Total number of major blood vessels that fluoroscopy has colored.	Numeric	0–3
Thal	Presence of thalassemia	3 = normal, 6 = fixed defect, 7 = reversible defect	3–7
Outcome	Class attribute indicating presence of heart disease	0 = No, 1 = Yes	0 or 1

Attributes

The dataset is divided into dependent/target variables and predicate/independent variables, with 1329 occurrences and 14 characteristics total. Table 1 provides comprehensive parameters data, including assessment, limits, and overview.

Description of the dataset

Descriptive statistics are essential for determining the properties of the data. It condenses the information such that human comprehension of the data is made easier. For the clinical qualities, Table 2 offers statistical metrics such as the number of records, mean, standard deviation (Std), minimum (min) value, and maximum (max) value. The age characteristic, for instance, has a standard deviation of 8.97 and a mean value of 55.39, ranging from 25 to 75. The remaining qualities likewise have their statistical measures computed.

Equilibrium of classes

When the dataset used for machine/ensemble learning models is not balanced for the issue statement, the models perform poorly. When the target class is not uniformly distributed, a balanced dataset can be created using a variety of sampling approaches. Class 1 indicates heart illness with 691 cases in the dataset used for this experiment, while class 0 represents no heart disease with 637 occurrences. Figure 2 illustrates this good blend of classes.

Correlation coefficient analysis

A statistical method for figuring out the orientation and extent of the relationship between various factors in a dataset is correlation coefficient analysis (CCA). The links between the dataset’s properties are found and shown in this study by the

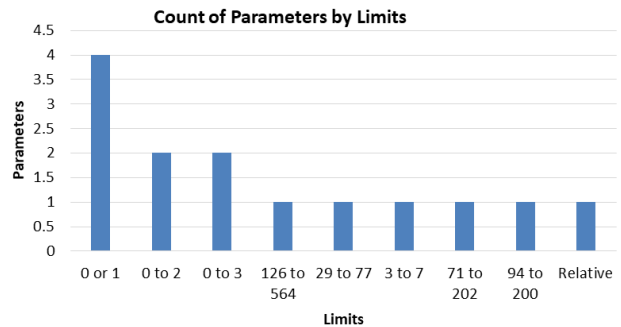


Figure 2: Bar chart representation between limits and parameters

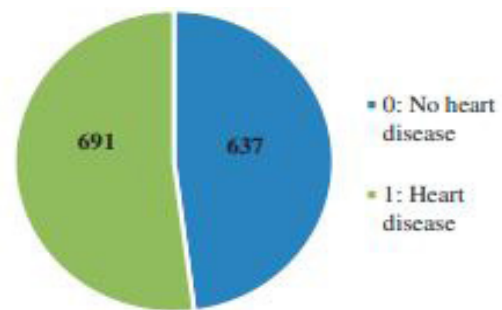


Figure 2: Class of heart disease division by pie chart

application of CCA.²⁰ A well-structured dataset is indicated by a significant correlation or link between independent and dependent features.

The correlation coefficient matrix for every feature used to predict illness is shown in Figure. 3. The associations span

Table 2: Lists the clinical statistical measurements along with their corresponding measures

Parameters	Count	Mean	Standard deviation	Minimum	Maximum
Age	1328	55.39	8.98	25	75
FastingBP	1328	0.20	0.45	0	1
Cpericarditis	1328	0.95	1.15	0	4
RestingECG	1328	0.61	0.52	0	2
Cholesterol	1328	235.11	50.99	127	565
Thal	1328	2.22	0.50	0	3
Sex	1328	0.71	0.51	0	1
MaximumHR	1328	150.02	23.01	72	203
Slope	1328	1.50	0.71	0	3
Ca	1328	0.82	1.15	0	4
Oldpeak	1328	1.12	1.28	0	1
RestingBP	1328	132.52	18.01	95	202
ExerciseIA	1328	0.44	0.50	0	1
Outcome	1328	0.61	0.50	0	1

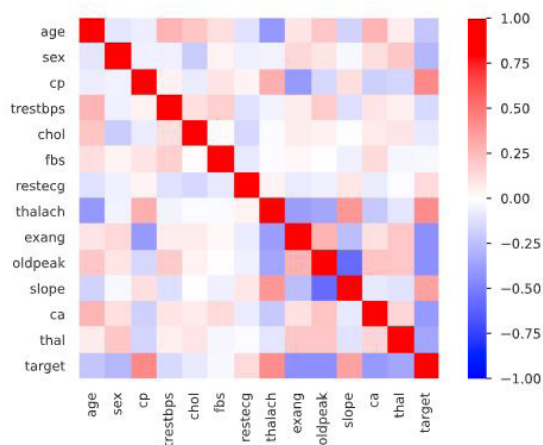


Figure 3: Correlation coefficient matrix

the x- and y-axes and vary from +1 to -1. The degree of link between the respective intersecting qualities is represented by each cell in the matrix. The correlation coefficient value of 0.12 indicates a slight positive link between age and resting blood pressure, for example.

RESULT AND EXPERIMENT DISCUSSION

This section explores the experimental design, results, and subsequent debate about the application of ensemble methods for the prediction of heart disease. The outcomes derived from applying the suggested framework are carefully examined and analyzed. The pre-processing and post-processing stages for illness prediction are the two sections of the results presentation. A thorough analysis is carried out, with a particular emphasis on the investigated boosting algorithms’ performance parameters, such as computational time, f1-score, precision, recall, and receiver operating curve.

Preparing the Data

In order to guarantee the stability and dependability of machine learning models, data pre-processing is essential.²¹ The data imputation approach was used in this study to find and replace missing variables. First, all missing values were found with the pandas library method. Then, mean and mode imputation techniques were implemented with Simple Imputer library method to fill in the missing values. The mean, median, and mode of the column were substituted for any missing values throughout this procedure. Employing the Interquartile range approach, which found and replaced outliers by using Z-score algorithms to center the mean around zero and modify the distribution of all data samples.

Classifier Accuracy

Figure 4 shows the testing accuracy of boosting methods. This study uses Extreme Gradient Boost (XGBoost), KNN, Decision Tree SVM and StackingCVClassifier algorithms. The accuracy of classifiers like XGBoost, KNN, DT, SVM and StackingCVClassifier was measured as 90, 89, 81.96, 88.52, and 92%, respectively, in the absence of pre-processing approaches. Following the use of pre-processing procedures, StackingCVClassifier outperformed the other ensemble method, attaining the greatest accuracy is 92.20%, ahead of XGB and KNN which achieved 90 and 89%, respectively.

Significance of Feature

When determining how much input feature (independent predicate variables) contributes to the prediction of the dependent variable, feature engineering is essential.²² It helps improve machine/ensemble learning models’ prediction outcomes. The feature importance score in this study indicates how frequently parameters is utilized for splitting throughout the training phase. A characteristic (like cholesterol) with a higher F-score is more significant in terms of prediction. Based

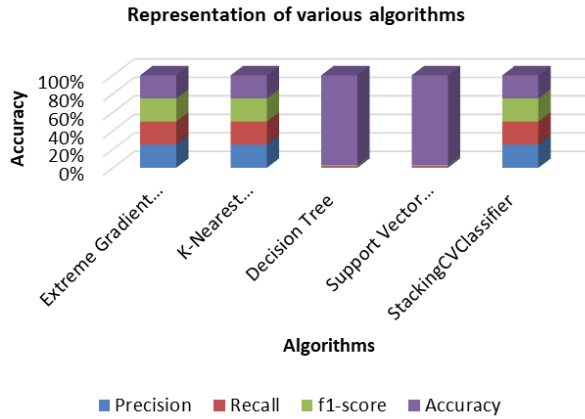


Figure 4: Shown the accuracy of used algorithm in our proposed model

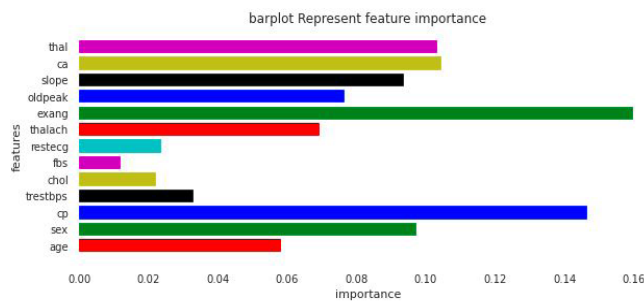


Figure 5: Shown feature importance

on each attribute's F-score, Figure 5 shows the importance of each attribute to prediction in descending order. For example, fasting blood pressure has the lowest value in prediction, but cholesterol has the highest significance.

ROC Curve

The predictive power of the examined ensemble methods at different thresholds is shown using the receiver operating characteristic (ROC) curve. It shows, on the x- and y-axes, respectively, the false-positive rate compared to the true-positive rate. Analysing how effectively the models separated with two classes 0 represented 'No' means there is no heart disease and 1 represented 'Yes' means there is heart disease made possible via the ROC curve. A higher ROC curve shows superior ability in predicting values between 0 and 1. Figure 6 displays the ROC curves for XGBoost, KNN, DT, SVM, logistic regression, Naïve bayes, random forest and StackingCVClassifier, and GB, in that order. Based on these numbers, the highest-performing option is StackingCVClassifier, followed by KNN and XGBoost.

Evaluation by Comparison

The suggested approach showed promising outcomes for heart disease prediction across a range of assessment parameters. Table 3 shows the results of a comparison of the suggested framework's accuracy, dataset, and approach performance with a number of pertinent research. The findings of our suggested framework were good, especially in terms of heart disease prediction accuracy. To produce better results than

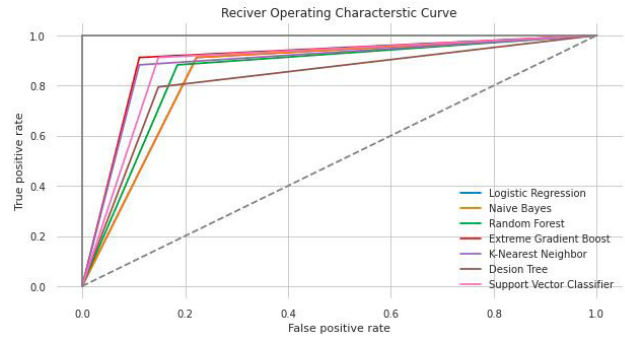


Figure 6: ROC curve for various used algorithms in this model

Table 3: Shows the results of a comparison with various research

Reference	Algorithms	Heart disease dataset	Accuracy (%)
11	XGB	Framingham	86.71
9	Voting	Cleveland	84.88
10	GB	--	88.68
12	Optimization applied on XGB	--	91.76
14	ADB	--	84.59
17	Random Forest	--	89.02
18	HREFLM	--	89.20
Our Method	XGBoost, KNN and StackingCVClassifier	--	92.20 for Stacking CV Classifier

earlier relevant research, strategies including data imputation for managing missing values and the Boxplot technique for identifying and replacing outliers were used.

CONCLUSION AND FUTURE WORK

In order to improve dataset quality and prediction accuracy, this study used a variety of pre-processing approaches in conjunction with boosting methods to successfully predict heart disease. Through the use of techniques, including imputation and data cleaning, the study sought to enhance the dataset's performance. Prior to pre-processing, boosting methods are StackingCVClassifier, XGBoost were put into practice. Gradient boosting was shown to be the most successful method based on evaluations used variation of statistical and ensemble methods. Their accuracy is 92.20% and other performance of methods attributes of recall, precision, and f1-score. Moreover, feature significance analysis clarified the noteworthy inputs of independent features to the ultimate prediction results.

It could be helpful to investigate other ensemble methods like voting and bagging in order to increase the effectiveness of this study. Furthermore, this methodology's usefulness and insights may be expanded by applying it to additional healthcare datasets with comparable feature characteristics. Additionally, using deep learning methods could provide new tools for better cardiovascular disease diagnosis and prediction.

REFERENCES

1. Badawy, Mohammed AbdElFattah Mohammed Darwesh, Lin Naing, Sofian Johar, Sokking Ong, Hanif Abdul Rahman, DayangkuSitiNurAshikinPengiran Tengah, Chean Lin Chong, and Nik Ani AfiqahTuah. "Evaluation of cardiovascular diseases risk calculators for CVDs prevention and management: scoping review." *BMC Public Health* 22, no. 1 (2022): 1742.
2. Ruan, Ye, YanfeiGuo, Yang Zheng, Zhezhou Huang, Shuangyuan Sun, Paul Kowal, Yan Shi, and Fan Wu. "Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and-middle-income countries: results from SAGE Wave 1." *BMC public health* 18 (2018): 1-13.
3. Crawford, Destiny L., and A. Patricia. "Nursing Care For Survivors Of Female Genital Mutilation." (2018).
4. Malik, Majid Bashir, Shahid Mohammad Ganie, and TasleemArif. "Machine learning techniques in healthcare informatics: Showcasing prediction of type 2 diabetes mellitus disease using lifestyle data." In *Predictive Modeling in Biomedical Data Mining and Analysis*, pp. 295-311. Academic Press, 2022.
5. Paul, Anand, AnandNayyar, Akshi Kumar, and JaffarAlzubi. "Preface—special issue "Energy Efficiency in Building using Intelligent computing for Smart Cities"." *Energy Systems* 13, no. 3 (2022): 563-566.
6. Malik, Majid Bashir, Mohd Ali, Sadiya Bashir, and Shahid Mohammad Ganie. "Performance Evaluation and Comparative Analysis of Machine Learning Techniques to Predict the Chronic Kidney Disease." In *International Conference on Artificial Intelligence on Textile and Apparel*, pp. 473-486. Singapore: Springer Nature Singapore, 2023.
7. Nissa, Najmu, Sanjay Jamwal, and Mehdi Neshat. "A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques." *Computation* 12, no. 1 (2024): 15.
8. Gaba, Charanjeet, SonamKhattar, and SheenamMiddha. "An Empirical Study of Machine Learning Methods for Analyzing Cardiovascular Disease." In *Proceedings of the 5th International Conference on Information Management & Machine Intelligence*, pp. 1-7. 2023.
9. Manur, Manohar, Alok Kumar Pani, and Pankaj Kumar. "A prediction technique for heart disease based on long Short term memory recurrent neural network." *International Journal of Intelligent Engineering and Systems* 13, no. 2 (2020): 31-39.
10. Theerthagiri, Prasannavenkatesan, and JyothiprakashVidya. "Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques." *Expert Systems* 39, no. 9 (2022): e13064.
11. Alizargar, Azadeh, Yang-Lang Chang, Tan-Hsu Tan, and Tsung-Yu Liu. "Comparative analysis of machine learning and ensemble approaches for hepatitis B prediction using data mining with synthetic minority oversampling technique." *Health and Technology* 14, no. 1 (2024): 109-118.
12. Mahadik, Avantika, Prashant Sharma, and VaibhavNarawade. "Proficient Prognostication through Hybrid Approach for Heart Disease." *Res Militaris* 13, no. 3 (2023): 3335-3349.
13. Sengupta, Sudip, Sanmay Kumar Patra, AritriLaha, RatneswarPoddar, Kallol Bhattacharyya, PradipDey, and Jajati Mandal. "Replacing conventional surface irrigation with micro-irrigation in vegetables can alleviate arsenic toxicity and improve water productivity." *Groundwater for Sustainable Development* 23 (2023): 101012.
14. Pouriye, Seyedamin, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, and Juan Gutierrez. "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease." In *2017 IEEE symposium on computers and communications (ISCC)*, pp. 204-207. IEEE, 2017.
15. Swati Sharma, Vijay Kumar Sharma, Rohit Aggarwal, Rashmi Gupta, "Covid-19 Pandemic Predictive System Using Machine Learning" Scopus index 3rd IEEE (ICAC3N-22).
16. Sharma, Vijay Kumar, Swati Sharma, MukeshRawat, and Ravi Prakash. "Adaptive Particle Swarm Optimization for Energy Minimization in Cloud: A Success History Based Approach." In *Towards the Integration of IoT, Cloud and Big Data: Services, Applications and Standards*, pp. 115-130. Singapore: Springer Nature Singapore, 2023.
17. Sharma, Swati, Vijay Kumar Sharma, Vimal Kumar, and Umang Arora. "Machine Learning Application: Sarcasm Detection Model." *Artificial Intelligence for a Sustainable Industry 4.0* (2021): 125-138.
18. Ganie, Shahid Mohammad, PijushKanti Dutta Pramanik, Majid Bashir Malik, AnandNayyar, and Kyung Sup Kwak. "An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms." *Comput. Syst. Sci. Eng.* 46, no. 3 (2023): 3993-4006.
19. Malik, Majid Bashir, Mohd Ali, Sadiya Bashir, and Shahid Mohammad Ganie. "Performance Evaluation and Comparative Analysis of Machine Learning Techniques to Predict the Chronic Kidney Disease." In *International Conference on Artificial Intelligence on Textile and Apparel*, pp. 473-486. Singapore: Springer Nature Singapore, 2023.
20. Kumar, Chaman, Priyanka Singh, Swati Hans, Laveena Sehgal, ShanuVerma, and Vijay Kumar Sharma. "Machine learning application: Detecting COVID-19 using X-Ray images."
21. Yang, Christopher C. "Explainable artificial intelligence for predictive modeling in healthcare." *Journal of healthcare informatics research* 6, no. 2 (2022): 228-239.
22. Datta, Debajit, Mansa Dey, Proshanta Kumar Ghosh, SohiniNeogy, and Asit Kumar Roy. "Coupling multi-sensory earth observation datasets, in-situ measurements, and machine learning algorithms for total blue C stock estimation of an estuarine mangrove forest." *Forest Ecology and Management* 546 (2023): 121345.