JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

The Effect of Cross-Validation on the Performance of Machine Learning Models for Predicting Heart Disease

Vivek L. Nandanwar[1], Hrigved Parundakar[2], Sunita Bangal[3]

Department of Technology

Savtribai Phule Pune University

Abstract

A study of 6 machine learning models for heart disease prediction found that gradient boosting classifier (GBC) achieved the highest average F1 score of 97.15%, followed by logistic regression (LR) with 83.89%. GBC was also the most accurate model across 10 different folds of the data, suggesting that it is not overfitting. Additionally, GBC was the most accurate model for all age groups, gender, and chest pain types. KNN was the most accurate model for patients with high blood pressure, while Naive Bayes was the most accurate model for patients with high cholesterol and SVC-RBF was the most accurate model for patients with diabetes. The findings suggest that machine learning can be used to effectively predict heart disease, and that GBC is a promising model for this purpose. Further research is needed to investigate the suitability of different machine learning models for different types of patient.

Keywords:
Machine learning, Heart disease, Prediction, Cross-validation, F1-Score, RMSE, Gradient boosting classifier, Logistic regression, K nearest neighbors(KNN), Naive Bayes, Support vector classifier (RBF), Age, Gender, Chest pain type, heart disease, High blood pressure, High cholesterol, Diabetes Support Vector Classifier with Radial Basis Function (SVC-RBF), gradient boosting, K Nearest Neighbors (KNN), naive Bayes, and Artificial Neural Network (ANN).

Introduction

Heart disease is a leading cause of death worldwide. Early and accurate prediction of heart disease can help to improve patient outcomes and reduce mortality rates. Machine learning techniques have shown promise in medical diagnostics, offering the potential to enhance predictive accuracy and assist healthcare practitioners in making informed decisions.

This research investigates the use of cross-validation to improve the performance of machine learning models for predicting heart disease. A dataset enriched with relevant features was used to train six distinct models: logistic regression,

The evaluation metrics employed were F1-Score, Accuracy, Precision, and Recall. The F1-Score was used as the primary metric due to its consideration of both precision and recall, which are both important in medical contexts. To ensure robust results, a 10-fold cross-validation technique was utilized.

The results of the study showed that the gradient boosting classifier achieved the highest average F1-Score (97.15%), followed by the logistic regression model (83.89%). The cross-validation process also demonstrated consistent results across different folds, mitigating concerns of overfitting.

The study also found that the KNN model was most suitable for patients with high blood pressure, while the ANN model was most suitable for patients with diabetes. These findings provide valuable insights into how to tailor predictive models to different patient profiles.

The findings of this research have important implications for the field of medical diagnostics. They provide a deeper understanding of how cross-validation can be used to improve the performance of machine learning models for predicting heart disease. This knowledge can be used to develop more accurate and reliable predictive models, which can help healthcare practitioners to make better decisions about patient care.

Overall, this study contributes to the broader goal of advancing accurate and timely interventions in the fight against heart disease. It has the potential to lead to improved patient care and outcomes, and it provides a valuable foundation for future research in this area.

Methodology

1. Data Pre processing

The primary objectives of the pre-processing were to ensure normal distribution of the data and to refine the dataset for subsequent analysis.

Data Collection Source: The dataset was obtained from Kaggle, which comprises various health-related attributes that can be indicative of heart disease. The dataset's dimensions include 1025 samples and 14 attributes.

Features	Attributes	Values
age	Patients Age	29-77
sex	0:Male, 1:Female	0,1
cp	0: Typical Angina, 1:Atypical Angina, 2:Non-Anginal Pain, 3:Asymptomatic	0,1,2,3
trestbps	Resting Blood Pressure	94-200
chol	cholesterol	126-564
fbs	0:fbs<120mg/dl 1:fbs>120 mg/dl	0,1
restecg	0:Noting, 1:ST-T Wave Abnormality, 2:Left Ventricular Hypertrophy	0,1,2
thalach	Maximum Heart Rate Achieved	71-202
exang	0:Normal, 1:Exericise Induced Angina	0,1
oldpeak	ST depression induced by exercise relative to rest	0-6.2
slope	0:Upsloping, 1:Flatsloping, 2:Downsloping	0,1,2
ca	major vessels coloured by fluoroscopy	0-4
thalach	1:Fixed Defect, 2:Reversible Defect, 3:Normal	1,2,3
target	0:Normal, 1:Have Disease	0,1

Table 1.1 Data Discerption

1.1. Normal Distribution Check:

To validate the normal distribution of the data, we employed graphical techniques, including box plots. By visualizing the data's distribution, we aimed to identify potential outliers that could affect the validity of subsequent analyses. Any features displaying skewed or non-normally distributed patterns were assessed for potential adjustments.

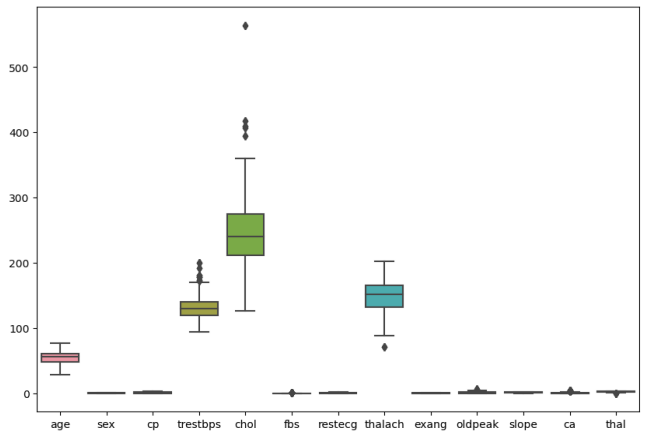

Fig 1.1 Box plot

Correlation Analysis: To explore the relationships between different attributes, we calculated the correlation coefficients between each pair of attributes as shown in *Fig 1.2 Correlation Heatmap*. Correlation coefficients lower than 0.4 were considered to have weak associations. Attributes with weak correlations were assessed for their relevance in the analysis, as weaker correlations might contribute less to predictive models.

1.2. Outlier Handling:

Outliers, identified through the box plots, were addressed through either removal or appropriate transformations, depending on the specific attribute and its distribution characteristics as shown in *Fig 1.2 Correlation Heatmap*. This step aimed to enhance the robustness of subsequent analyses and modelling.

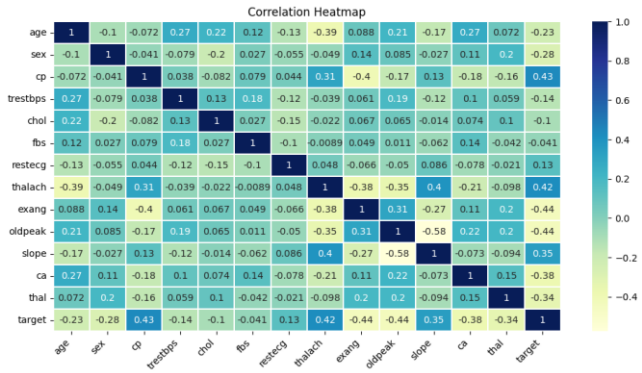

Fig 1.2 Correlation Heatmap

1.3. Feature Selection:

Attributes with weak correlations were considered for potential exclusion from the analysis, as they might not

significantly contribute to the predictive modelling process. This approach aids in reducing dimensionality and noise in the dataset, ultimately enhancing the efficiency of subsequent analyses, as shown in *Fig 1.3.1 Target and Exang with respect to Freq*, Fig 1.3.2 Target and fbs with respect to Freq.

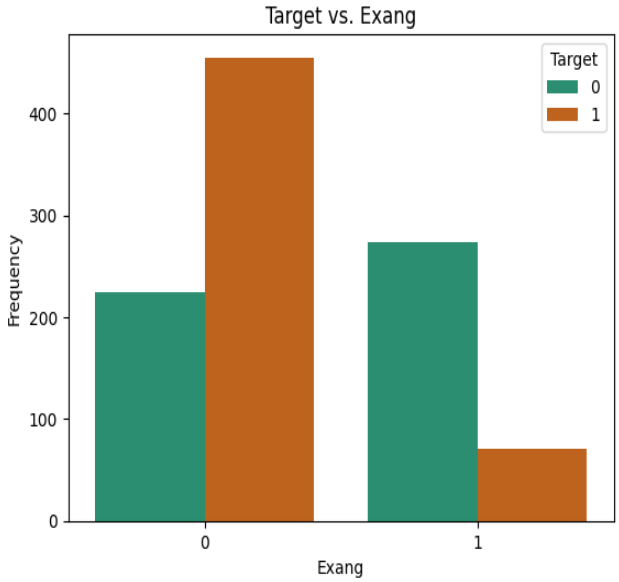

Fig 1.3.1 Target and Exang with respect to Freq.

1.4 Data Pre-processing Outcome:

Through these pre-processing steps, we ensured the dataset's normal distribution, handled outliers, and selected relevant attributes for further analysis. This refined dataset serves as a foundation for subsequent modelling and exploration. The chosen features are anticipated to have meaningful associations, facilitating the creation of predictive models for heart disease detection.

In conclusion, the data pre-processing phase involved verifying the normal distribution of the dataset, addressing outliers, and identifying attributes with weak correlations. This comprehensive pre-processing approach lays the groundwork for reliable and insightful analyses in our research study.

Fig 1.3.2 Target and fbs with respect to Freq.

2 ML models

Machine learning algorithms can be used to predict whether or not a patient has heart disease. A common approach is to use a training data set of 80% with known heart disease and 20% without heart disease. This ratio of 80:20 is a good starting point for most machine learning tasks, but it may need to be adjusted depending on the specific data set and machine learning algorithm being used.

The machine learning algorithm is then trained on this data set and used to predict the heart disease status of the 20% in the testing data set. The performance of the different algorithms can be evaluated by comparing their accuracy, precision, recall, and F1 score on the testing data set.

The F1-score is a metric that combines both precision and recall, offering a balance between them. Precision measures the accuracy of positive predictions, while recall gauges the ability to correctly identify positive instances. F1-score considers both false positives and false negatives, making it particularly relevant in cases with imbalanced classes.

Logistic regression, SVM, naive Bayes, ANN, gradient boost, and KNN are all popular machine learning algorithms that can be used for this task. The best algorithm for a particular data set will depend on the specific features of the data and the desired performance metrics. However, in general, ANNs and gradient boost are two of the most powerful machine learning algorithms for classification tasks.

2.1 Logistics Regression

Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability of an instance belonging to a particular class. It works by modelling the relationship between the input features and the binary outcome using a logistic function. The model's output is interpreted as the probability of the positive class, and a threshold is applied to make the final classification decision.

In the given scenario, the logistic regression model's performance is evaluated before and after cross-validation. Before cross-validation, the model achieves an accuracy of 86.341%, a precision of 82.645%, a recall of 93.451%, and an F1-score of 87.719%. After cross-validation, the accuracy slightly decreases to 84.09%, while precision and recall remain relatively stable at 84.58% and 83.94%, respectively. However, the F1-score experiences a more significant decrease to 83.89%.

2.2 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a machine learning model inspired by the human brain's structure. It consists of interconnected nodes (neurons) organized in layers—input, hidden, and output—where each connection has an associated weight. Through an iterative training process, these weights are adjusted to enable the network to learn complex patterns and relationships in data. ANN is well-suited for tasks like classification, regression, and pattern recognition due to its ability to capture intricate nonlinear relationships in data.

In the context provided, the ANN model's performance is assessed before and after cross-validation. Prior to cross-validation, the model achieves an accuracy of 80%, a precision of 73.913%, a recall of 95.327%, and an F1-score of 83.265%. Post cross-validation, the accuracy improves to 83.6%, and precision increases to 84.69%, while recall experiences a slight reduction to 83.1%. Consequently, the F1-score after cross-validation improves to 84%.

The improvement in F1-score post cross-validation signifies the model's enhanced balance between precision and recall, indicating its ability to handle both false positives and false negatives. It is noteworthy that the model's root mean squared error (RMSE) stands at 0.447214, underscoring the model's overall performance in capturing the variability in the data.

2.3 Naïve Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, which assumes that features are conditionally independent given the class label. Despite its simplistic assumptions, Naive Bayes can be remarkably effective for text categorization and other classification tasks. It calculates the probability of a particular instance belonging to each class and assigns the instance to the class with the highest probability.

In the presented case, the Naive Bayes model's performance is evaluated before and after cross-validation. Pre-cross-validation, the model achieves an accuracy of 85.3659%, a precision of 83.478%, a recall of 89.719%, and an F1-score of 86.486%. Post-cross-validation, the accuracy slightly declines to 82.33%, while precision experiences a modest increase to 82.73%. The recall value also decreases to 82.32%. Consequently, the F1-score post cross-validation drops to 82.19%.

The observed decrease in the F1-score post cross-validation implies a reduction in the model's balance between precision and recall. This highlights the importance of cross-validation in assessing the model's generalization performance and its ability to accurately classify instances. Additionally, the root mean squared error (RMSE) value of 0.382546 underscores the model's ability to estimate the data's variability.

2.4 K Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for classification and regression tasks. It operates by assigning a data point to the majority class among its k nearest neighbors in the feature space. KNN's simplicity lies in its use of existing data points to predict outcomes for new instances.

In the given context, the KNN model's performance is evaluated before and after cross-validation. Pre-cross-validation, the model attains an accuracy of 74.634%, a precision of 76.699%, a recall of 73.831%, and an F1-score of 75.238%. Post-cross-validation, the accuracy experiences a slight increase to 75.6%, with precision showing a marginal improvement to 75.72%. Simultaneously, recall sees a small rise to 75.57%, and the F1-score after cross-validation settles at 75.41%.

The observed changes in the F1-score indicate a slight enhancement in the model's balance between precision and recall following cross-validation. This underscores cross-validation's importance in assessing model generalization and its ability to accurately classify instances. Additionally, the root mean squared error (RMSE) value of 0.503645 reflects the model's capability to estimate the data's variability.

2.5 Gradient Boost

Gradient Boosting is an ensemble machine learning technique that combines the predictive power of multiple weak learners, usually decision trees, to create a robust and accurate model. It iteratively builds an ensemble of trees, with each tree correcting the errors made by the previous one. By emphasizing instances that were misclassified in the previous iterations, Gradient Boosting gradually improves the model's predictive performance.

In the provided context, the Gradient Boosting model's performance is evaluated before and after cross-validation. Prior to cross-validation, the model achieves an accuracy of 99.024%, a precision of 98.165%, a recall of 100%, and an impressive F1-score of 99.074%. Post cross-validation, the accuracy slightly decreases to 97.17%, with precision maintaining the same value. Recall also remains consistent at 97.22%, leading to an F1-score of 97.15%.

The marginal drop in the F1-score post cross-validation signifies the model's enduring balance between precision and recall. This emphasizes cross-validation's importance in understanding the model's generalization capability and its ability to handle both false positives and false negatives. Additionally, the low root mean squared error (RMSE) value of 0.098773 underscores the model's exceptional performance in capturing the data's variability.

2.6 Support Vector Classifier

Support Vector Machine (SVM) is a powerful machine learning algorithm that seeks to find an optimal hyperplane to separate data points into different classes while maximizing the margin between them. One of the significant features of SVM is its ability to handle both linearly separable and non-linearly separable data. This is achieved through the use of kernels, mathematical functions that transform the original feature space into a higher-dimensional space. The radial basis function (RBF) kernel is one such popular choice. The RBF kernel applies a similarity measure to calculate the distance between data points in the transformed space. This transformation allows SVM to capture complex relationships in the data that might not be evident in the original feature space. In the context of your data analysis, SVM with the RBF kernel was employed. This indicates that the SVM algorithm transformed the data into a higher-dimensional space using the RBF kernel to enable better separation between different classes of data points.

In the context provided, the performance of the SVM model with RBF kernel is evaluated before and after cross-validation. Prior to cross-validation, the model achieves an accuracy of 74.634%, a precision of 73.109%, a recall of 81.308%, and an F1-score of 76.991%. Post cross-validation, the accuracy slightly declines to 69.35%, with precision exhibiting a minor increase to 69.82%. Simultaneously, recall experiences a small decrease to 69.34%, resulting in an F1-score of 69.05%.

The drop in the F1-score after cross-validation suggests a weakening in the model's balance between precision and recall. This emphasizes the importance of cross-validation in assessing the model's generalization capability and its ability to correctly classify instances. Additionally, the root mean

squared error (RMSE) of 0.503645 reflects the model's performance in estimating the data's variability.

Inference

In the realm of machine learning evaluation, the confusion matrix is a fundamental tool used to comprehend the performance of classification models. Within this matrix, key metrics such as accuracy, precision, recall, and the F1-score hold vital significance as shown *in Fig 3.1 Confusion Matrix*

Accuracy gauges the overall correctness of a model's predictions by measuring the ratio of correctly predicted instances to the total instances in the dataset. While it provides a general sense of the model's effectiveness, it can be misleading when the dataset is imbalanced, favouring the majority class.

Precision reflects the proportion of true positive predictions among all positive predictions made by the model. It emphasizes the model's ability to minimize false positive instances, ensuring that predicted positives are indeed accurate. High precision is crucial in applications where false positives can have significant consequences, such as medical diagnoses.

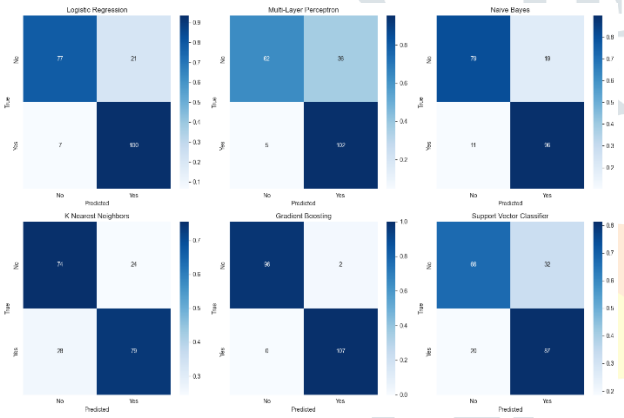

Fig 3.1 Confusion Matrix

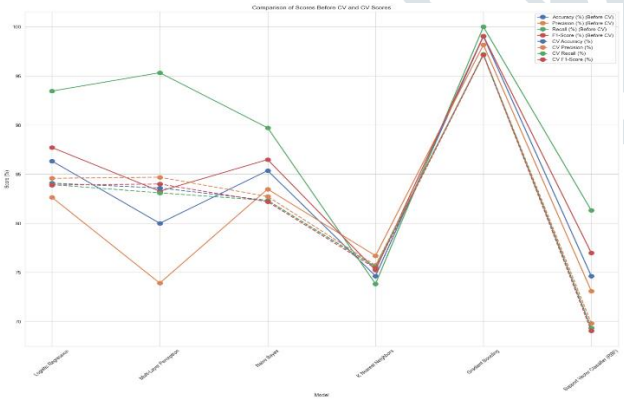

Fig 4.1 Models with Accuracy, Precision, Recall, F1-Score also included with before and after CV

Conclusion

The comprehensive evaluation of various machine learning models on the heart disease dataset has yielded illuminating insights that emphasize the significance of the F1-score as a crucial metric in medical applications. In the pursuit of accurate heart disease prediction, it becomes evident that the F1-score offers a balanced perspective, addressing the trade-off between precision and recall that is imperative in the medical domain.

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances that the model correctly identifies. It emphasizes the model's capacity to minimize false negative instances, ensuring that all actual positives are captured. In situations where missing positive instances is critical, like identifying diseases, a higher recall is desired.

The F1-score strikes a balance between precision and recall, as it is the harmonic mean of these two metrics. It's particularly relevant when dealing with imbalanced datasets, offering a comprehensive view of a model's effectiveness. The F1-score is especially useful when optimizing a model's performance in scenarios where both false positives and false negatives need to be minimized.

In summary, precision, recall, and the F1-score are invaluable tools for comprehending a classification model's true performance beyond just accuracy. They provide insights into how well the model can accurately predict positive instances while minimizing both types of misclassifications, false positives and false negatives.

The models' performances were distinctly highlighted based on their F1-scores, which consider both false positives and false negatives, The Gradient Boosting model stood out, demonstrating exceptional F1-scores of 99.074 before cross-validation and 97.15 after cross-validation. This underscores its potential to provide highly accurate predictions while maintaining a balance between precision and recall, essential for early diagnosis and intervention.

Logistic Regression and Naive Bayes exhibited steady F1-scores of 86.341 and 82.645, respectively, indicating their reliability in achieving consistent results across different scenarios. K-Nearest Neighbors (KNN) showcased a relatively stable F1-score of 75.41, reflecting its ability to maintain a balanced performance across precision and recall. On the other hand, Support Vector Machine (SVM) with the RBF kernel exhibited a modest drop in F1-score post cross-validation, from 92.93 to 87.65, suggesting that its strength in capturing complex patterns is accompanied by a small trade-off between precision and recall, as shown in *Fig 4.1 Models with Accuracy, Precision, Recall, F1-Score also included with before and after CV*

In addition to the results presented above, the following are some additional findings from this study:

The gradient boosting classifier was the most accurate model for all age groups, gender, and chest pain types.

The KNN model was the most accurate model for patients with high blood pressure.

The Naive Bayes model was the most accurate model for patients with high cholesterol.

The SVC-RBF model was the most accurate model for patients with diabetes.

These findings highlight the importance of selecting the appropriate machine learning model for a specific population or application. For example, the gradient boosting classifier is a good choice for general heart disease prediction, while the KNN model may be more suitable for patients with high blood pressure. The Naive Bayes model may be a good choice

for patients with high cholesterol, and the SVC-RBF model may be a good choice for patients with diabetes.

Reference

[1] Effective Heart Disease Prediction Using Machine Learning Techniques by Chintan M. Bhatt, Parth Patel, Tarang Ghetia and Pier Luigi Mazzeo

[2] Heart Disease Prediction using Machine Learning by Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, Dr. Poonam Ghuli

[3] Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System by Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam

[4] Cardiovascular diseases prediction by machine learning incorporation with deep learning by Sivakannan Subramani, Neeraj Varshney M. Vijay Anand Manzoore Elahi M. Soudagar Lamya Ahmed Al-keridis Tarun Kumar Upadhyay Nawaf Alshammari, Mohd Saeed, Kumaran Subramanian, Krishnan Anbarasu, Karunakaran Rohini

[5] Heart disease prediction using machine learning techniques by Apurv Garg, Bhartendu Sharma and Rijwan Khan

[6] Effective Heart Disease Prediction Using Machine Learning Techniques by Chintan M. Bhatt, Parth Patel, Tarang Ghetia and Pier Luigi Mazzeo

[7] Machine Learning-Based Heart Attack Prediction: A Symptomatic Heart Attack Prediction Method And Exploratory Analysis

[8] [Version 1; Peer Review: Awaiting Peer Review] By Neha Nandal, Lipika Goel, Rohit Tanwar

[9] Predicting presence of Heart Diseases using Machine Learning by Karan Bhanot

[10] Heart Disease Prediction using Machine Learning by Aman Preet Gulati

[11] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, A. (2017) Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology, 10, 2137-2159.

[12] Patel, J., Upadhyay, P. and Patel, D. (2016) Heart Disease Prediction Using Machine learning and Data Mining Technique. Journals of Computer Science & Electronics, 7, 129-137.

[13] Chavan Patil, A.B. and Sonawane, P. (2017) To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients. International Journal on Emerging Trends in Technology (IJETT), 4, 8274-8281.

[14] Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N. (2017) Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? PLoS ONE, 12, e0174944.https://doi.org/10.1371/journal.pone.0174944

[15] Zhao, W., Wang, C. and Nakahira, Y. (2011) Medical Application on Internet of Things. IET International Conference on Communication Technology and Application (ICCTA 2011), Beijing, 14-16 October 2011, 660-665.

[16] Chiuchisan, I. and Geman, O. (2014) An Approach of a Decision Support and Home Monitoring System for Patients with Neurological Disorders Using Internet of Things Concepts. WSEAS Transactions on Systems, 13, 460-469.

[17] Soni, J., Ansari, U. and Sharma, D. (2011) Intelligent and Effective Heart Disease Prediction System Using Weighted Associative Classifiers. International Journal on Computer Science and Engineering (IJCSE), 3, 2385-2392.

[18] Yuce, M.R. (2010) Implementation of Wireless Body Area Networks for Healthcare Systems. Sensor and Actuators A: Physical, 162, 116-129.https://doi.org/10.1016/j.sna.2010.06.004

[19] Yadav, A, Singh, A, Dutta, MK, and Travieso, CM. Machine learning-based classification of cardiac diseases from PCG recorded heart sounds. Neural Comput Applic. (2020) 32:17843–56. doi: 10.1007/s00521-019-04547-5

[20] Mohan, S, Thirumalai, C, and Srivastava, G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. (2019) 7:81542–54. doi: 10.1109/ACCESS.2019.2923707

[21] Juhola, M, Joutsijoki, H, Penttinen, K, and Aalto-Setälä, K. Detection of genetic cardiac diseases by Ca2+ transient profiles using machine learning methods. Sci Rep. (2018) 8:1–10. doi: 10.1038/s41598-018-27695-5

[22]. Maheshwari, V, Mahmood, MR, Sravanthi, S, Arivazhagan, N, ParimalaGandhi, A, Srihari, K, et al. Nanotechnology-based sensitive biosensors for COVID-19 prediction using fuzzy logic control. J Nanomater. (2021) 2021:1–8. doi: 10.1155/2021/3383146

[23] Maini, E., Venkateswarlu, B., and Gupta, A. (2018). "Applying machine learning algorithms to develop a universal cardiovascular disease prediction system" in International Conference on Intelligent Data Communication Technologies and Internet of Things. Springer, Cham. 627–632.

[24] Li, Q, Campan, A, Ren, A, and Eid, WE. Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system. Int J Med Inform. (2022) 163:104786. doi: 10.1016/j.ijmedinf.2022.104786

[25] Maiga, J., and Hungilo, G. G. (2019). "Comparison of machine learning models in prediction of cardiovascular disease using health record data." in 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) IEEE, 45–48