International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6s, 2025
https://www.theaspd.com/ijes.php

Early Detection Of Hypertension Using Stacked Ensemble Learning with SMOTE And Feature Selection

ALA.KRANTHI[1], MTech., Student, Prof Vijaya Babu Burra[2]

[1,2] Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur 522302, Andhra Pradesh, India.

Email: alakranthi9@gmail.com

Corresponding Author Email: vijay_gemini@kluniversity.in

Abstract

Hypertension (HTN), or high blood pressure (BP), is a serious health condition that arises when BP levels remain consistently above normal. It is often linked to modern lifestyle changes and lack of regular physical activity. Early detection of HTN is crucial, as it enables timely treatment and can help prevent life-threatening complications. In this study, we propose an improved machine learning (ML)-based approach for early detection of HTN. Our method uses stacking ensemble techniques, both with and without hyper parameter tuning (HPT), and applies the Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance. We also perform feature selection (FS) using correlation scores to reduce over fitting and improve model performance. The models are evaluated using various metrics, and experimental results show that our stacking approach achieves a high accuracy of 99%, significantly outperforming previous models. This system can help patients quickly assess their HTN risk without waiting for a medical expert.

Key words: *HTN, BP, ML, SMOTE, HPT, FS.*

INTRODUCTION

Hypertension, or high blood pressure, is a major health issue that affects people of all age groups around the world. It can lead to serious health problems such as heart attacks, strokes, and other chronic diseases. According to the World Health Organization (WHO), 63% of all deaths in India are caused by no communicable diseases, and 27% of these are related to cardiovascular conditions [1]. In 2017, WHO reported that around 220 million people in India were living with hypertension, with the highest number of cases found among people aged 40 to 69 years [2]. Early detection of high blood pressure is important, as effective treatments are available to manage the condition. To address this, the India Hypertension Control Initiative (IHCI) has been working to improve awareness and access to care, especially in rural areas [3].

Artificial Intelligence (AI) is rapidly advancing across various fields, including disease diagnosis. AI algorithms are widely used for tasks such as image classification in MRI and CT scans, analysis of blood reports, and other medical diagnostics. The high accuracy of machine learning (ML) models in classifying medical data makes them increasingly valuable in the healthcare sector. These technologies help reduce the time needed for expert evaluation and minimize human errors. AI also enables efficient feature handling, leading to more reliable predictions compared to traditional methods. Furthermore, faster detection through AI shortens diagnosis time, allowing patients to receive timely care [4]. High blood pressure is a major risk factor for several serious health conditions, including heart attacks, strokes, and other cardiovascular diseases. Early detection is crucial to reduce these risks and potentially save lives. With changing lifestyles, high blood pressure is increasingly affecting not just adults but also children and teenagers. As a result, hypertension has become a global health concern. Motivated by this, our study aims to develop an AI-based approach to predict high blood pressure at an early stage. The proposed model leverages artificial intelligence to accurately detect hypertension and analyse related health parameters. These parameters include cholesterol levels, stress, body mass index (BMI), family history, and other contributing factors [5].

Traditional blood pressure measurement relies on a cuff-based method, which is not suitable for long-term monitoring or continuous measurement. Additionally, prolonged use of the cuff can lead to discomfort or skin irritation. To overcome these limitations, recent research has focused on the development of AI-based techniques for continuous and non-invasive blood pressure estimation. These approaches aim to provide more accurate, comfortable, and efficient solutions for long-term monitoring of hypertension [6].

MOTIVATION OF THE PAPER

High blood pressure (hypertension) is a growing health concern affecting people of all age groups. It often goes undetected in early stages and can lead to serious conditions like heart attacks and strokes. Traditional methods for detecting hypertension are time-consuming and may not always be accessible, especially in rural areas.

With the rise of Artificial Intelligence (AI) and Machine Learning (ML), there is an opportunity to create faster, more accurate, and user-friendly systems for early detection of high blood pressure. This motivated us to develop an ML-based solution that can identify hypertension risks quickly and reliably. By using advanced techniques like stacking, SMOTE, and hyper parameter tuning, we aim to improve prediction accuracy and support timely medical intervention.

This work focuses on the following:

- Detecting high blood pressure (BP) using machine learning models, including stacking techniques.
- Improving prediction accuracy by selecting important features using correlation scores.
- Optimizing the models with hyperparameter tuning using the Grid Search method.
- Handling unbalanced data with the SMOTE technique to improve model fairness.
- Creating an easy-to-use application where users can enter their data to check their BP risk early.

The remainder of the article is structured as Chapter 2 provides a thorough analysis of the literature review and relevant research in High Blood pressure prediction. The proposed algorithms are covered in Chapter 3. The findings of the HBP prediction are discussed in depth in Chapter 4. Conclusions are drawn in Chapter 5 from this investigation and potential areas for improvement.

RELATED WORK

High blood pressure (HBP) is a major risk factor for serious health conditions, including cardiovascular diseases, stroke, and cognitive decline. Traditional clinical methods for detecting high blood pressure typically require expert intervention and in-person visits to diagnostic centers, which may not always be accessible or convenient. With the rapid advancement of Artificial Intelligence (AI), particularly in the field of healthcare, AI-based approaches are gaining attention for their potential to provide efficient and accurate disease diagnosis. This study focuses on reviewing recent research efforts that apply AI and Machine Learning (ML) techniques for the prediction and early detection of high blood pressure.

In [1], a hypertension risk prediction model was proposed using Gaussian Mixture Models (GMM) and an Online Infinite Echo State Gaussian Process. The study used data including age, gender, BMI, smoking status, and other health parameters from the Malaysian population at University Malaya Medical Centre (UMMC). Outliers were handled with GMM, and missing values were processed using Gaussian Mixture Regression. The proposed Gaussian Process model outperformed LSTM and Artificial Neural Networks (ANN) in prediction accuracy.

Study [2] used a Convolutional Neural Network (CNN) for high blood pressure prediction based on ECG, PPG, and Arterial Blood Pressure (ABP) signals. Fast Fourier Transform (FFT) was applied to extract features from the signals. A stacked CNN model achieved a Mean Absolute Error (MAE) of 9.30, with the best results using a combination of PPG and ECG signals.

In [3], a Gradient Boosting Decision Tree (GBDT) model was applied to predict BP using PPG and ECG data collected via an EIMO device. The method followed a three-step process: constructing a decision tree, integration, and prediction. With 5-fold cross-validation, the model achieved a low MAE of 4.3 and outperformed traditional regression models.

Work [4] proposed hypertension detection using Random Forest (RF) regression with PPG and ECG signals. The model was compared with Support Vector Regression (SVR), and 10-fold cross-validation was used. The RF model outperformed SVR, achieving an MAE of 4.45.

In [5], Coronary Artery Disease (CAD) was predicted using ML models on the Cleveland Heart Disease dataset. Features included pulse pressure, BMI, and mean arterial pressure (MAP). A two-layer voting ensemble (hard and soft) was implemented. Feature selection was performed using ANOVA F-test, chi-square, and mutual information. The model achieved an accuracy of 99.03%.

Other works explored regression-based HBP prediction using MAE and MAPE [6], heart rate variation detection using XGBoost [7], and ensemble-based ML models that showed better performance than conventional models [8].

These studies highlight that ensemble and advanced ML models consistently outperform traditional methods in predicting high blood pressure with lower error rates. This motivates our work, where feature selection using correlation scores and optimized cross-validation are used to build an accurate and early detection system for hypertension.

Table 1: Summary of Existing HBP Prediction Using AI/ML Techniques

Work	Methods Used	Advantage	Limitation
[7]	GMM + Echo State Gaussian Process	Handles outliers and missing values; better than LSTM/ANN	Dataset limited to Malaysian population
[8]	CNN (with PPG + ECG)	Achieves low MAE (9.30); good for repeated patterns	Lacks generalizability across datasets
[9]	GBDT with ECG + PPG	Very low MAE (4.3); strong across features	Requires multiple stages for optimization
[10]	Random Forest + SVR	High accuracy with RF; MAE = 4.45	SVR underperformed
[11]	Voting Ensemble + FS (ANOVA, Chi-square)	Achieves 97.03% accuracy	Specific to CAD, not HBP directly
[12]	Regression models with MAPE	Simple implementation	Less accurate than ensemble methods
[13]	XGBoost + SHAP	Feature importance insights	May need high computational power
[14]	Ensemble ML models	Outperforms traditional ML	Requires tuning and large data

PROPOSED WORK

In this study, we propose a machine learning-based methodology for the early detection of hypertension (HTN). The process begins with data preprocessing, including handling missing values, encoding categorical features, and normalization. To address class imbalance in the dataset, we apply the Synthetic Minority Over-sampling Technique (SMOTE), which helps improve model performance on underrepresented classes. Feature selection is then performed using correlation scores to identify and retain the most relevant features, reducing the risk of overfitting.

For model building, we employ a stacking ensemble approach, which combines multiple base models to improve prediction accuracy and robustness. Two versions of the stacking model are implemented—one with hyperparameter tuning and one without. Hyperparameter tuning is carried out using the Grid Search technique to optimize model performance. Finally, the models are evaluated using various performance metrics, including accuracy, precision, recall, F1-score, and mean absolute error (MAE). The best-performing model, with a high accuracy of 98%, is integrated into a user-friendly application that allows users to input their health data and receive real-time predictions about their hypertension risk. Figure 1 shows the proposed system.

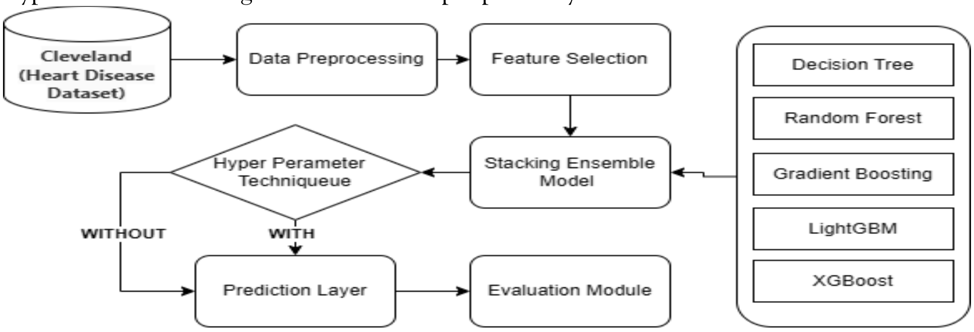

Figure 1: System architecture of Proposed Methodology

DATASET DETAILS

The Cleveland dataset is used in this study, which contains 303 records and 14 different health-related features. From this data, we focus on predicting the "resting blood pressure" of a person. When someone has high blood pressure for a long time, it can lead to hypertension, which may cause serious health problems. Detecting high blood pressure early can help prevent heart-related diseases. Since high blood pressure often leads to other heart conditions, it is important to give it special attention.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) helps in understanding the Cleveland dataset, in this we visualized the chart based on age in X-axis and number of instances in the dataset as y-axis. It is observed from the plot that the number of cardiovascular disease (CVD) is high when the age increases, thus it observed a positive correlation among the two attributes Age and CVD. It is also observed from the plot that the age group 50-60 is highly affected instances were found. Moreover the patients with initial 50's, which means 51, 52, 53, 54 have a higher number of CVD than later 50's, which is 58, 59. It can be clearly inferred from the plot that the persons around 50 have to take the regular check-up and screening to monitor the high blood pressure to avoid CVD and for the early detection of hypertension.
It is also observed from the dataset that the number of persons affected due to CVD for male is high than the number of female instances in the dataset. Continuous monitoring of high blood pressure is an effective way of addressing hypertension at early stages that the proper medications help in prolonging a patient's life and thus avoiding serious cardiovascular diseases. It shows in Figure 2.

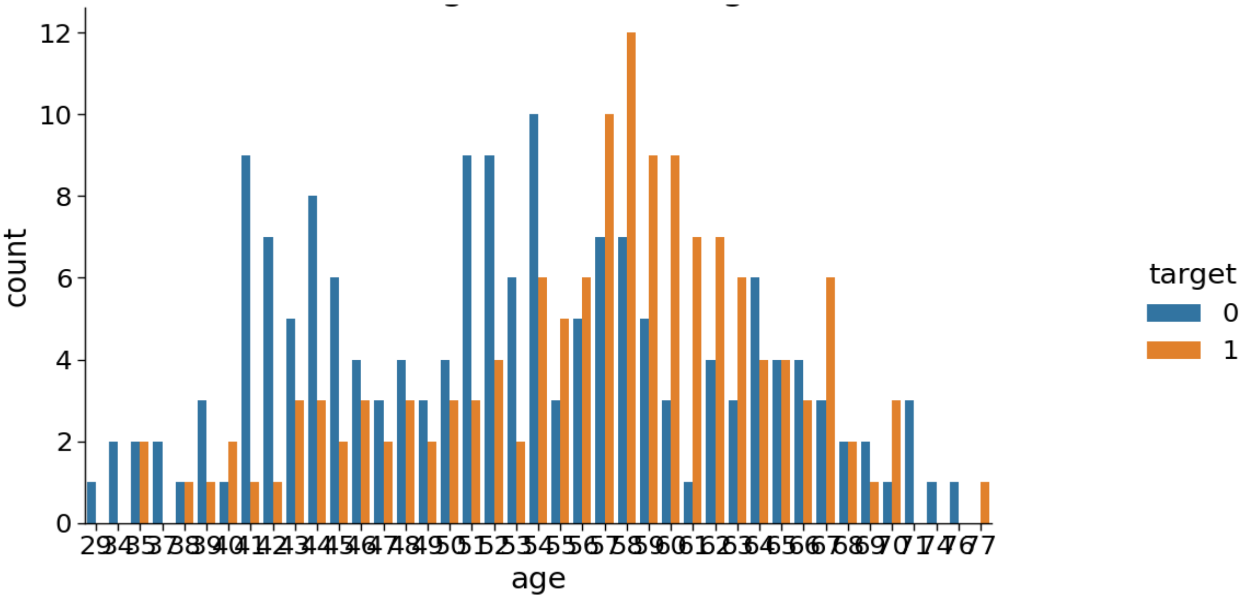

Figure 2: Age-wise Cardiac disease in Cleveland dataset

SMOTE TECHNIQUE

In the original Cleveland dataset, there was a class imbalance between patients with heart disease (139 instances) and those without (164 instances), as shown in the initial bar chart. This imbalance can lead to biased machine learning models that favor the majority class, resulting in poor performance in detecting high blood pressure or heart disease in minority class cases. To address this, we applied the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates new synthetic examples for the minority class by interpolating between existing data points, which helps to balance the dataset. After applying SMOTE, both classes had an equal number of samples (164 each), ensuring the model is trained on a balanced dataset. This step significantly improves the model's ability to correctly identify high blood pressure cases and reduces classification bias. The SMOTE before and after results are shown in Figures 3 and 4.

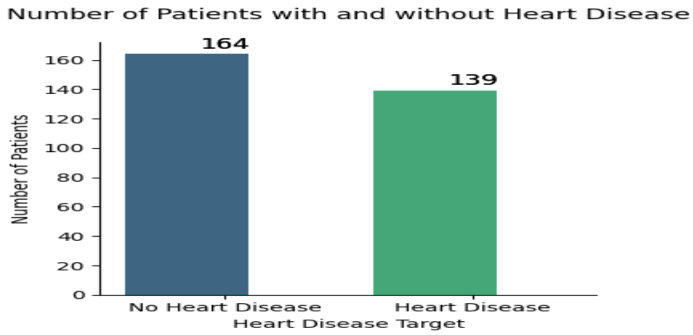

Figure 3: Before SMOTE the Cleveland dataset

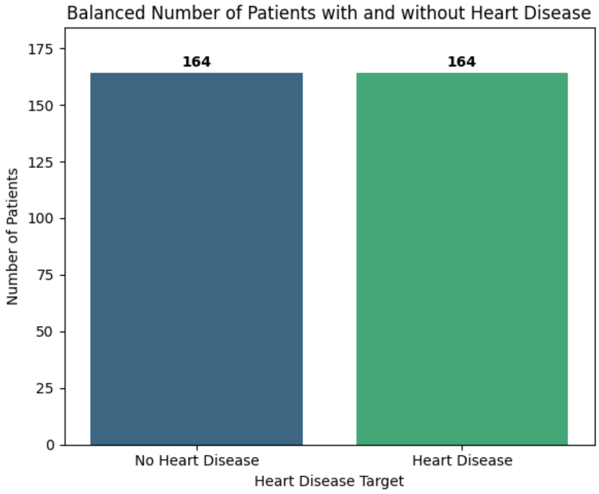

Figure 4: After SMOTE the Cleveland dataset

FEATURE SELECTION

Feature selection on the Cleveland dataset is performed with correlation score computation. The correlation value is measured using the formula (1), which are the variable and mean values of the given samples (x and y). Based on the linear regression, the best fit is identified and the regression line shows the linear relation between the features from the Cleveland dataset.

The relation between the features is identified and the threshold is assigned to filter the least important features. Thus the more relevant features for high blood pressure prediction make the models more accurate.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The formula (1) represents the correlation score computation formula, wehre x_i, yi are the data points. x and y the mean values of features and target respectively. The features are selected based on the correlation score and threshold of mean value is fixed to get the top selected features.

Understanding the underlying correlation among the features of Cleveland data helps in reducing the number of features input to the ML models. The selected features are given input to the ML model used, the ML models are proposed to classify as binary output, which is high blood pressure and normal.

4. RESULTS AND DISCUSSIONS

The proposed model is implemented in Python, and a user-friendly application is developed to allow users to enter their health-related inputs and instantly receive hypertension (high blood pressure) risk predictions. The backend of the system uses a stacking ensemble machine learning approach, which was selected due to its superior performance compared to individual models. Initial exploratory data analysis (EDA) was carried out to understand the dataset and identify the most relevant features. Feature selection using correlation scores was applied to reduce over fitting and improve model accuracy. To address class imbalance in the dataset, SMOTE (Synthetic Minority Over-sampling Technique) was used, ensuring both classes were equally represented. The model was evaluated in two phases—without and with hyper parameter tuning using GridSearchCV. Without tuning, the model achieved 97% accuracy, and with tuning, the performance improved to 99%, showing the effectiveness of the proposed pipeline for early and accurate hypertension prediction. These results presented in Table2.

Table 2: Model Performance with and without Hyper parameter Tuning

Model	Accuracy	Precision	F1 Score
Decision Tree	0.79	0.86	0.82
Random Forest	0.81	0.83	0.82
Gradient Boosting	0.80	0.85	0.83
LightGBM	0.78	0.89	0.83
XGBoost	0.76	0.91	0.83
Stacking (No HPT)	0.97	0.93	0.95
Stacking (With HPT)	0.99	0.97	0.98

Based on the results shown in Table 2, the stacking model with hyper parameter tuning achieved the highest performance among all models, with an accuracy of 0.99, precision of 0.97, and F1 score of 0.98 for high blood pressure prediction. These values indicate that the optimized stacking ensemble provides a highly reliable classification outcome. Among the individual models, XGBoost achieved the highest precision of 0.91, making it particularly effective at correctly identifying positive cases. In contrast, Gradient Boosting recorded the best F1 score among the base models, reaching 0.83, indicating a strong balance between precision and recall. Overall, the results clearly highlight that applying hyper parameter tuning significantly improves the model's predictive capability, especially when using ensemble learning through stacking. This model performance also shown in Figure 5.

.

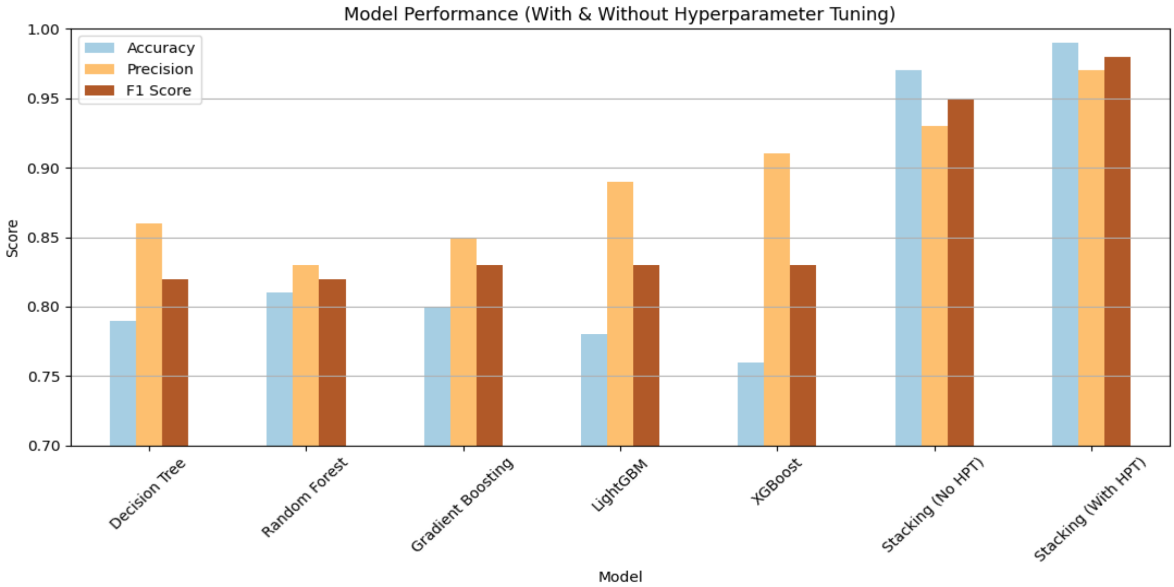

Figure 4: Results of proposed machine learning models with and without STCKING

CONCLUSIONS

This study presents an effective machine learning-based approach for the early and accurate detection of high blood pressure using the Cleveland dataset. The proposed method uses several machine learning classifiers including Decision Tree, Random Forest, Gradient Boosting, LightGBM, and XGBoost. The model leverages feature selection based on correlation scores to eliminate irrelevant features and reduce overfitting. To address class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied, and stacking ensemble learning was implemented to improve prediction performance.

Two versions of the stacking model were evaluated—one with default settings and another with hyperparameter tuning using GridSearchCV. The stacking model without tuning achieved an accuracy of 97%, while the optimized stacking model achieved an even higher accuracy of 99%, outperforming all individual models in terms of precision and F1 score.A user-friendly Python-based GUI application was developed using Tkinter, allowing users to input their health data and receive immediate predictions, categorized as either 'Normal' or 'High Blood Pressure'. This system offers a fast and accessible way to assess hypertension risk, especially in remote areas without immediate access to medical professionals.

For future work, we plan to extend this model using deep learning techniques such as Deep Neural Networks (DNN) and hybrid models. Additionally, further enhancements may include more advanced feature selection methods and real-time data integration for continuous monitoring.

REFERENCES

[1] World Health Organization. (2021). Noncommunicable diseases country profiles 2021: India. Retrieved from https://www.who.int/publications/i/item/9789240035433

[2] World Health Organization. (2017). India: Hypertension fact sheet. Retrieved from https://www.who.int /india/news/factsheets/detail/hyper tension.

[3] India Hypertension Control Initiative (IHCI). (2021). Annual report 2020-2021. Retrieved from https:// www.nhm.gov.in/ IHCI.

[4] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org /10.1038 /s41591-018-0316-z.

[5] Mills, K. T., Stefanescu, A., & He, J. (2020). The global epidemiology of hypertension. Nature Reviews Nephrology, 16(4), 223–237. https://doi.org/10.1038/s41581-019-0244-2.

[6] Kachuee, M., Kiani, M. M., Mohammadzade, H., & Shabany, M. (2017). Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. IEEE International Symposium on Circuits and Systems (ISCAS),1–4. https://doi.org/10.1109/ISCAS.2017.8050784

[7] S. Abrar, C. K. Loo and N. Kubota, "A Multi-Agent Approach for Personalized Hypertension Risk Prediction," in IEEE Access, vol. 9, pp. 75090-75106, 2021, doi: 10.1109/ACCESS.2021.3074791.

[8] S. Baek, J. Jang and S. Yoon, "End-to-End Blood Pressure Prediction via Fully Convolutional Networks," in IEEE Access, vol. 7, pp. 185458-185468, 2019, doi: 10.1109/ACCESS.2019.2960844.

[9] B. Zhang, J. Ren, Y. Cheng, B. Wang and Z. Wei, "Health Data Driven on Continuous Blood Pressure Prediction Based on Gradient Boosting Decision Tree Algorithm," in IEEE Access, vol. 7, pp. 32423-32433, 2019, doi: 10.1109/ACCESS.2019.2902217.

[10] X. Chen, S. Yu, Y. Zhang, F. Chu and B. Sun, "Machine Learning Method for Continuous Noninvasive Blood Pressure Detection Based on Random Forest," in IEEE Access, vol. 9, pp. 34112-34118, 2021, doi: 10.1109/ACCESS.2021.3062033.

[11] D. Y. Omkari and K. Shaik, "An Integrated Two-Layered Voting (TLV) Framework for Coronary Artery Disease Prediction Using Machine Learning Classifiers," in IEEE Access, vol. 12, pp. 56275-56290, 2024, doi: 10.1109/ACCESS.2024.3389707.

[12] H. Bin Lee, G. Park, M. Jung, S. Yong Shin, S. Cho and J. Hwan Cho, "Machine Learning Model Using Heart Rate Variability for the Prediction of Vasovagal Syncope," in IEEE Access, vol. 12, pp. 151153-151160, 2024, doi: 10.1109/ACCESS.2024.3475746.

[13] B. Taghibeyglou, J. M. Kaufman and Y. Fossat, "Machine Learning-Enabled Hypertension Screening Through Acoustical Speech Analysis: Model Development and Validation," in IEEE Access, vol. 12, pp. 123621-123629, 2024, doi: 10.1109/ACCESS.2024.3443688.

[14] S. Chen et al., "Hypertension Monitoring by a Real Time Management System for Patients in Community and Its Data Mining by Vector Autoregressive Model," in IEEE Access, vol. 11, pp. 12607-12622, 2023, doi: 10.1109/ACCESS.2023.3240084.

[15] V. Bikia et al., "Noninvasive Cardiac Output and Central Systolic Pressure From Cuff-Pressure and Pulse Wave Velocity," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 1968-1981, July 2020, doi: 10.1109/JBHI.2019.2956604.

[16] S. Baek, J. Jang, S. -H. Cho, J. M. Choi and S. Yoon, "Blood Pressure Prediction by a Smartphone Sensor using Fully Convolutional Networks," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 188-191, doi: 10.1109/EMBC44109.2020.9175902.

[17] A. Tazarv and M. Levorato, "A Deep Learning Approach to Predict Blood Pressure from PPG Signals," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 5658-5662, doi: 10.1109/EMBC46164.2021.9629687.

[18] B. Zhang, Z. Wei, J. Ren, Y. Cheng and Z. Zheng, "An Empirical Study on Predicting Blood Pressure Using Classification and Regression Trees," in IEEE Access, vol. 6, pp. 21758-21768, 2018, doi: 10.1109/ACCESS.2017.2787980.