

# An Optimized Heterogeneous Stacking Ensemble with Hyperparameter Optimization for Multi Class Hypertension Risk Classification

Novi Yona Sidratul Munti<sup>1</sup>, Yuda Irawan<sup>2</sup>, Muhammad Habib Yuhandri<sup>3</sup>

<sup>1</sup>Department of Computer Science, Universitas Islam Sumatera Barat, Padang

<sup>2</sup>Department of Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru

<sup>3</sup>Department of Information System, Universitas Putra Indonesia "YPTK" Padang, Padang

## Article Info

### Article history:

Received June 09, 2026

Revised June 25, 2026

Accepted July 16, 2026

### Keywords:

Hypertension Classification  
Stacking Ensemble Learning  
CatBoost  
GridSearchCV  
SHAP

## ABSTRACT

Hypertension is a major global health concern and a leading risk factor for cardiovascular disease, stroke, kidney failure, and premature mortality. Accurate and interpretable prediction of hypertension risk is essential for supporting early intervention and preventive healthcare. This study proposes an optimized and explainable stacking ensemble framework for multi class hypertension classification by integrating LASSO feature selection, SMOTEENN data balancing, heterogeneous ensemble learning, CatBoost meta learning, GridSearchCV optimization, and SHAP based explainability. The proposed architecture combines five base learners, namely XGBoost, LightGBM, Random Forest, Extra Trees, and Support Vector Machine, whose probability outputs are transformed into meta features and processed by an optimized CatBoost meta learner. Experiments conducted on a dataset containing 3,000 hypertension related records demonstrated superior classification performance, achieving 97.50% accuracy, 97.46% precision, 97.50% recall, and 97.48% F1 score. 10 fold cross validation further confirmed the robustness of the framework with a mean accuracy of  $97.50\% \pm 0.0027$ . ROC analysis produced AUC values above 0.97 for all classes, indicating excellent discriminative capability. The results demonstrate that the proposed framework provides both high predictive accuracy and strong interpretability, making it a promising solution for intelligent hypertension risk assessment and clinical decision support.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Novi Yona Sidratul Munti,  
Department of Computer Science,  
Universitas Islam Sumatera Barat,  
Jl. By Pass No.KM.16, Koto Panjang Ikua Koto, Kec. Koto Tangah, Kota Padang, Sumatera Barat 25586,  
Indonesia  
Email: noviyona@uisb.ac.id

## 1. INTRODUCTION

Hypertension remains one of the most significant global health challenges and is widely recognized as a primary contributor to cardiovascular disease, stroke, renal dysfunction, and premature mortality [1]. According to the World Health Organization, more than 1.28 billion adults worldwide are affected by hypertension, while nearly 46% remain undiagnosed or inadequately managed, increasing the risk of severe cardiovascular complications [2]. The growing prevalence of hypertension is strongly associated with population aging, obesity, sedentary lifestyles, unhealthy dietary habits, psychological stress, and metabolic disorders [3]. Recent epidemiological studies have further demonstrated that hypertension contributes substantially to healthcare expenditures and long term socioeconomic burdens, particularly in developing countries where

early detection remains limited [4]. Consequently, developing intelligent and accurate hypertension prediction systems has become an important research priority for modern healthcare analytics [5].

Recent advances in artificial intelligence (AI), machine learning (ML), and healthcare informatics have significantly transformed disease prediction and clinical decision support systems [6]. Traditional statistical approaches often struggle to model complex nonlinear interactions among demographic, physiological, and lifestyle related variables. In contrast, machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost, and ensemble learning frameworks have demonstrated superior capability in extracting hidden patterns from high dimensional healthcare datasets [7], [8]. Furthermore, the emergence of Explainable Artificial Intelligence (XAI) techniques has improved model transparency, enabling clinicians to better understand prediction outcomes and enhancing trust in AI assisted healthcare systems [9].

Several recent studies have reported promising results in hypertension prediction using machine learning approaches. Researcher proposed a machine learning based hypertension risk prediction framework integrated with SHAP interpretation and reported improved model transparency [10]. Other researcher demonstrated that ensemble learning methods achieved higher predictive performance than single classifiers for hypertension classification [11]. Another researcher utilized a large scale healthcare dataset consisting of approximately 9.58 million records and showed that machine learning models could effectively identify personalized hypertension risk profiles [12]. Similarly, another researcher employed XGBoost and SHAP analysis to identify influential hypertension risk factors and achieved robust classification performance [13]. These studies collectively confirm the growing effectiveness of machine learning driven hypertension prediction systems.

Despite these advances, several critical limitations remain evident in the existing literature. First, many studies focus primarily on binary hypertension classification rather than multi class categorization involving Normal, Prehypertension, and Hypertension conditions [14]. Second, class imbalance remains a significant challenge because healthcare datasets often contain unequal class distributions, resulting in biased classification outcomes and reduced minority class recognition [15]. Third, several investigations rely on single machine learning models, limiting their ability to capture diverse learning patterns from heterogeneous clinical features [16]. Fourth, although ensemble learning frameworks have demonstrated superior performance, many studies employ homogeneous learners or lack systematic meta learner optimization, reducing predictive robustness [17]. Finally, explainability mechanisms are often treated as post hoc analysis tools rather than being integrated into the overall predictive architecture.

To address these limitations, this study proposes a comprehensive and optimized hypertension classification framework integrating LASSO based feature selection, SMOTEENN data balancing, heterogeneous stacking ensemble learning, CatBoost meta learning, GridSearchCV optimization, and SHAP explainability. The proposed framework combines XGBoost, LightGBM, Random Forest, Extra Trees, and SVM as heterogeneous base learners to capture complementary predictive information from clinical, physiological, and lifestyle related features. The probability outputs generated by these models are subsequently transformed into meta features and optimized using a CatBoost meta learner enhanced by GridSearchCV hyperparameter tuning.

The novelty of this study can be explained from four perspectives. Theoretically, this research integrates heterogeneous ensemble learning and explainable artificial intelligence within a unified multi class hypertension classification framework. Methodologically, it combines LASSO feature selection, SMOTEENN balancing, heterogeneous stacking, CatBoost based meta learning, and GridSearchCV optimization into a single predictive architecture. Technically, the framework exploits complementary learning mechanisms from boosting based, bagging based, and margin based classifiers to generate highly informative meta representations. Practically, the integration of SHAP analysis enables transparent clinical interpretation, supporting trustworthy hypertension risk assessment and facilitating potential real world deployment.

Therefore, the primary objective of this study is to develop an optimized and explainable machine learning framework capable of accurately classifying hypertension into three clinically relevant categories: Normal, Prehypertension, and Hypertension. Specifically, this research aims to

evaluate the effectiveness of heterogeneous stacking ensemble learning, investigate the contribution of CatBoost based meta learner optimization, and provide interpretable prediction outcomes through SHAP based explainability. The major contributions of this study are summarized as follows: (1) proposing a novel heterogeneous stacking framework integrating XGBoost, LightGBM, Random Forest, Extra Trees, and SVM; (2) implementing CatBoost with GridSearchCV optimization as an advanced meta learner; (3) incorporating LASSO feature selection and SMOTEENN balancing to improve data quality and class representation; and (4) providing transparent hypertension prediction through SHAP based explainable artificial intelligence.

## 2. RESEARCH METHOD

Accurate and interpretable multi-class hypertension classification remains a critical challenge in healthcare analytics due to the complexity of physiological and lifestyle-related risk factors. This study proposes an optimized stacking ensemble framework that integrates feature selection, data balancing, heterogeneous machine learning models, hyperparameter optimization, and explainable artificial intelligence to enhance predictive performance and model transparency. The overall research workflow is presented in Figure 1.

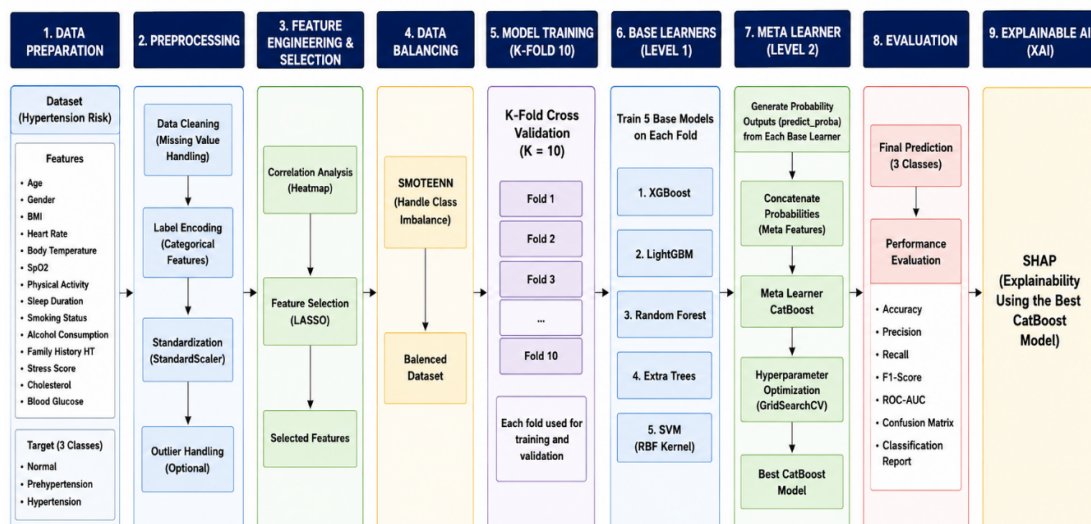


Figure 1. Proposed Methodology

The proposed methodology consists of several interconnected stages, including data preparation, preprocessing, feature analysis and selection, data balancing, model training, performance evaluation, and explainability analysis. Initially, the dataset is processed through data cleaning, label encoding, feature standardization, correlation analysis, and LASSO-based feature selection to improve data quality and identify the most relevant predictive attributes. Subsequently, class imbalance is addressed using the SMOTEENN technique, followed by the development of a two-level stacking ensemble architecture comprising five heterogeneous base learners (XGBoost, LightGBM, Random Forest, Extra Trees, and SVM) and a CatBoost meta learner optimized using GridSearchCV. Finally, the effectiveness of the proposed framework is evaluated using multiple classification metrics, while SHAP is employed to provide transparent and interpretable insights into the contribution of individual features to the model's prediction outcomes.

### 2.1. Data Preparation

The data preparation stage plays a crucial role in ensuring the quality, consistency, and reliability of the dataset before the development of the classification model. In this study, a hypertension dataset consisting of 3,000 records was originally obtained from a publicly available Kaggle repository during the initial stage of this research. The dataset contains demographic, physiological, lifestyle, and clinical attributes associated with hypertension risk, including Age,

Gender, BMI, Heart Rate, Body Temperature, SpO<sub>2</sub>, Physical Activity, Sleep Duration, Smoking Status, Alcohol Consumption, Family History of Hypertension, Stress Score, Cholesterol, and Blood Glucose. The target variable was categorized into three classes: Normal, Prehypertension, and Hypertension. Representative samples of the dataset are presented in Table 1.

Table 1. Representative samples from the hypertension dataset.

| Age | Gender | BMI   | Heart Rate | ... | Stress Score | Cholesterol | Blood Glucose | Label           |
|-----|--------|-------|------------|-----|--------------|-------------|---------------|-----------------|
| 56  | Male   | 21.01 | 76         | ... | 32           | 153         | 138           | Normal          |
| 69  | Female | 18.52 | 100        | ... | 35           | 192         | 121           | Prehypertension |
| 46  | Female | 22.27 | 83         | ... | 39           | 197         | 77            | Normal          |
| 32  | Male   | 23.02 | 78         | ... | 60           | 223         | 76            | Prehypertension |
| 60  | Female | 18.14 | 72         | ... | 86           | 165         | 133           | Normal          |

Table 1 presents several representative records from the hypertension dataset, illustrating the diversity of demographic, physiological, and clinical characteristics considered in this study. The dataset contains 14 predictor variables, namely Age, Gender, BMI, Heart Rate, Body Temperature, SpO<sub>2</sub>, Physical Activity, Sleep Duration, Smoking Status, Alcohol Consumption, Family History of Hypertension, Stress Score, Cholesterol, and Blood Glucose, together with a target variable representing three hypertension categories: Normal, Prehypertension, and Hypertension. The class labels were provided as part of the original dataset obtained from the Kaggle repository and were retained without modification throughout this study. The authors did not assign, relabel, or redefine the target classes; instead, the original annotations supplied by the dataset provider were used directly during model development. Although the original dataset documentation did not explicitly describe the label assignment procedure, the provided class annotations were adopted as the ground truth for model training and evaluation. Furthermore, variables directly associated with hypertension diagnosis, such as systolic and diastolic blood pressure measurements, were not included among the predictor variables used for model training. Consequently, the proposed model learns predictive relationships from demographic, physiological, lifestyle, and clinical characteristics rather than reproducing deterministic diagnostic rules, thereby minimizing the risk of data leakage.

## 2.2. Data Preprocessing

The data preprocessing stage was conducted to improve data quality and ensure that all variables were suitable for machine learning analysis. As illustrated in Figure 1, the preprocessing procedure included data cleaning to handle potential inconsistencies, followed by label encoding to transform categorical attributes into numerical representations that could be processed by the classification algorithms. Subsequently, feature standardization was performed using the StandardScaler technique to normalize the scale of the predictor variables and prevent features with larger numerical ranges from dominating the learning process. In addition, optional outlier handling was considered to reduce the impact of extreme observations, thereby enhancing the robustness and stability of the proposed stacking ensemble framework during model training and evaluation.

## 2.3. Feature Engineering and Selection

The feature engineering and selection stage was performed to identify the most relevant attributes contributing to hypertension classification while reducing redundant information. As illustrated in Figure 1, correlation analysis was first conducted using a heatmap to examine the relationships among predictor variables and assess the presence of multicollinearity within the dataset. Subsequently, the Least Absolute Shrinkage and Selection Operator (LASSO) technique was employed to evaluate feature relevance by assigning coefficient weights to each variable and retaining the most informative predictors[18][19]. This process improved the quality of the input features, reduced model complexity, and enhanced the predictive capability of the proposed stacking ensemble framework.

## 2.4. Data Balancing

Class imbalance is a common issue in medical datasets and can lead to biased model performance toward majority classes. To address this challenge, the Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbor (SMOTEENN) was employed to balance the class distribution before model training [20], [21]. SMOTE generates synthetic samples for minority classes by interpolating between existing minority instances and their nearest neighbors, while ENN removes noisy and ambiguous samples that may negatively affect the classification boundary. The generation of synthetic samples using SMOTE can be expressed as:

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \quad (1)$$

where  $x_i$  represents a minority class sample,  $x_{nn}$  denotes one of its nearest minority neighbors, and  $\lambda$  is a random value in the interval  $[0,1]$ . Subsequently, the ENN procedure eliminates samples whose class labels differ from the majority labels of their  $k$ -nearest neighbors, thereby improving class separability and reducing noise in the dataset. The combination of oversampling and data cleaning provided by SMOTEENN enables the construction of a more balanced and representative training dataset, ultimately enhancing the robustness and generalization capability of the proposed stacking ensemble model.

## 2.5. K-Fold Validation

To ensure the reliability and generalization capability of the proposed model, a 10-fold cross-validation strategy was employed during the training and evaluation process [22]. In this approach, the dataset was partitioned into ten equal subsets, where nine subsets were used for training and the remaining subset was used for validation. The process was repeated ten times until each subset had served as the validation set. To prevent information leakage and ensure a fair evaluation, all preprocessing procedures, including StandardScaler transformation, LASSO-based feature selection, and SMOTEENN resampling, were performed exclusively on the training subset within each fold. The transformation parameters and selected features obtained from the training data were subsequently applied to the corresponding validation subset. Furthermore, SMOTEENN was applied only to the training data and was never used on the validation fold. This procedure ensured that the validation data remained completely unseen during model development, thereby providing an unbiased assessment of the model's generalization capability. The final performance of the model was obtained by averaging the evaluation results across all folds, resulting in a more robust and reliable estimate of classification performance.

## 2.6. Stacking Model

The core contribution of this study lies in the development of an optimized two-level stacking ensemble framework for multi-class hypertension classification [21], [23]. Unlike conventional stacking approaches that typically employ a limited number of homogeneous classifiers, the proposed framework integrates five heterogeneous base learners, namely XGBoost, LightGBM, Random Forest, Extra Trees, and Support Vector Machine (SVM), to capture diverse learning patterns from the dataset. The prediction probabilities generated by these base learners are subsequently combined into a new feature space, referred to as meta-features, which are then utilized by a CatBoost meta-learner to produce the final classification output. Furthermore, the CatBoost meta-learner is optimized using GridSearchCV to identify the optimal hyperparameter configuration, thereby enhancing predictive performance and model generalization. At the first level, each base learner generates a probability prediction for each class. The probability output of the  $i$ -th base learner can be represented as:

$$P_i(x) = [p_{i1}(x), p_{i2}(x), \dots, p_{ic}(x)] \quad (2)$$

where  $P_i(x)$  denotes the probability vector produced by the  $i$ -th base learner, and  $c$  represents the number of classes. The meta-feature matrix is then constructed by concatenating the probability outputs from all base learners:

$$M = [P_1(x), P_2(x), \dots, P_n(x)] \quad (3)$$

where  $M$  denotes the meta-feature matrix and  $n$  represents the number of base learners. The CatBoost meta-learner utilizes the generated meta-features to produce the final prediction:

$$\hat{y} = f_{CatBoost}(M) \quad (4)$$

where  $\hat{y}$  denotes the final predicted class label and  $f_{CatBoost}$  represents the optimized CatBoost classifier. To identify the optimal hyperparameter combination, GridSearchCV systematically evaluates multiple parameter configurations using cross-validation. The optimal parameter set is selected according to:

$$\theta^* = \arg \max_{\theta \in \Theta} Score(\theta) \quad (5)$$

where  $\theta^*$  represents the optimal hyperparameter configuration,  $\Theta$  denotes the search space, and  $Score(\theta)$  corresponds to the cross-validation performance obtained for a given parameter set. The proposed stacking ensemble framework leverages the complementary strengths of boosting-based, bagging-based, and margin-based learning algorithms while incorporating meta-level optimization through CatBoost and GridSearchCV. This hybrid architecture enables the model to capture complex nonlinear relationships among hypertension-related risk factors and improve classification performance compared with individual machine learning models.

For each training fold, every base learner (Random Forest, XGBoost, LightGBM, SVM, and Extra Trees) was trained independently using the corresponding training subset. Instead of using the predicted class labels, the class probability estimates produced by the `predict_proba()` function were extracted for each validation instance. These out-of-fold probability outputs were concatenated to form a new feature matrix, referred to as the meta-feature matrix. Each row of this matrix represents the probability estimates generated by all base learners for a single sample, while each column corresponds to the predicted probability of a specific class from a particular base learner. The resulting meta-feature matrix was subsequently used as the input to the CatBoost meta-learner, enabling it to learn the optimal combination of the base learners' probabilistic predictions. By generating the meta-features exclusively from out-of-fold predictions, the proposed stacking framework effectively prevents information leakage and provides an unbiased estimate of model performance.

## 2.7. Evaluation

The performance of the proposed stacking ensemble model was evaluated using several widely adopted classification metrics, including accuracy, precision, recall, and F1 score, to provide a comprehensive assessment of predictive performance. Accuracy measures the overall proportion of correctly classified instances, while precision evaluates the proportion of correctly predicted positive samples among all predicted positive samples. Recall quantifies the model's ability to correctly identify actual positive instances, whereas the F1 score provides a balanced measure by considering both precision and recall simultaneously. For the multi-class classification task, the reported precision, recall, and F1-score represent the weighted average across the three hypertension classes. This averaging strategy was adopted to provide a single overall performance measure while accounting for the class distribution in the evaluation dataset. In addition, confusion matrix analysis and Receiver Operating Characteristic (ROC) curves were employed to further examine classification effectiveness and the discriminative capability of the proposed model across the three hypertension classes.

## 2.8. Explainable AI (XAI)

To enhance the transparency and interpretability of the proposed stacking ensemble framework, Explainable Artificial Intelligence (XAI) was incorporated using the SHAP (SHapley Additive exPlanations) method[24]. SHAP quantifies the contribution of each feature to an individual prediction by assigning an importance value based on cooperative game theory principles. The SHAP value for a feature can be calculated as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (6)$$

where  $\phi_i$  denotes the SHAP value of feature  $i$ ,  $S$  represents a subset of features excluding feature  $i$ ,  $N$  is the complete set of features, and  $f(\cdot)$  denotes the prediction function. By analyzing SHAP values, the proposed model provides interpretable insights into how individual attributes influence classification outcomes, thereby improving model transparency and supporting trustworthy decision-making in hypertension risk assessment.

### 3. RESULTS AND DISCUSSION

In this section, the experimental results of the proposed stacking ensemble framework are presented and comprehensively analyzed to evaluate its effectiveness in multi-class hypertension classification. The discussion covers feature analysis and selection, data balancing, classification performance, model optimization, and explainability analysis to demonstrate the robustness, predictive capability, and interpretability of the proposed approach.

#### 3.1. Data Preprocessing

The data preprocessing stage included an initial exploratory analysis to examine the relationships among predictor variables and assess the suitability of the dataset for subsequent machine learning modeling. Correlation analysis was performed using a Pearson correlation heatmap to identify potential multicollinearity issues and evaluate the statistical dependencies among the selected features. The results of the correlation analysis are presented in Figure 2.

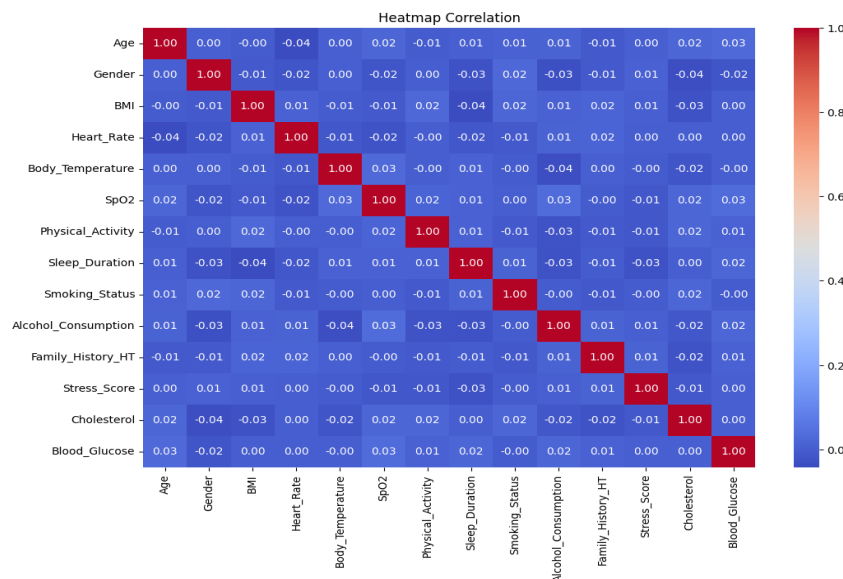


Figure 2. Correlation Heatmap of Features in the Hypertension Dataset

Figure 2 illustrates the pairwise correlation coefficients among the fourteen predictor variables used in this study. The results indicate that the majority of feature pairs exhibit very weak linear relationships, with correlation coefficients generally ranging between  $-0.04$  and  $0.03$ , suggesting a low degree of multicollinearity within the dataset. From a statistical perspective, none of the variables reached the commonly accepted multicollinearity threshold of  $|r| \geq 0.70$ , indicating that each feature contributes relatively independent information to the classification process. Furthermore, the absence of strong positive or negative correlations suggests that the dataset is well-suited for ensemble learning approaches, as the predictor variables capture diverse aspects of hypertension risk without introducing substantial redundancy. To further evaluate the contribution of individual features to the classification task, feature relevance analysis was performed using the LASSO technique, as presented in Figure 3.

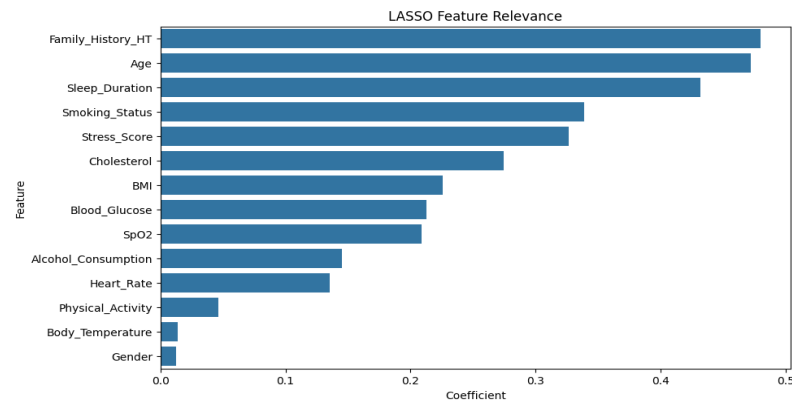


Figure 3. LASSO Based Feature Relevance Analysis for Hypertension Classification

Figure 3 illustrates the coefficient values assigned by LASSO to all 14 predictor variables included in the hypertension dataset. The results indicate that Family History of Hypertension and Age achieved the highest relevance scores, with coefficients approaching 0.48 and 0.47, respectively, suggesting that these variables contribute most substantially to the discrimination of hypertension risk levels. Furthermore, variables such as Sleep Duration, Smoking Status, Stress Score, and Cholesterol also exhibited relatively high coefficients ( $> 0.25$ ), indicating strong predictive significance, whereas Gender, Body Temperature, and Physical Activity obtained comparatively lower coefficients ( $< 0.05$ ). Despite the variation in coefficient magnitudes, all 14 features were retained for subsequent modeling because each variable maintained a non-zero coefficient, indicating that every predictor provided meaningful information and contributed to the overall predictive capability of the proposed stacking ensemble framework.

To ensure that all predictor variables contributed proportionally during model training, feature standardization was performed using the StandardScaler technique, and representative standardized samples are presented in Table 2.

Table 2. Sample of Standardized Feature Values After StandardScaler Transformation

| Age    | Gender | BMI    | Heart_Rate | ... | Family_History_HT | Stress_Score | Cholesterol | Blood_Glucose |
|--------|--------|--------|------------|-----|-------------------|--------------|-------------|---------------|
| 0.395  | 0.999  | -1.130 | -0.141     | ... | -0.818            | -1.043       | -1.500      | 1.545         |
| 1.121  | -1.001 | -1.699 | 1.812      | ... | -0.818            | -0.872       | -0.380      | 0.757         |
| -0.163 | -1.001 | -0.842 | 0.429      | ... | 1.222             | -0.644       | -0.237      | -1.282        |
| -0.944 | 0.999  | -0.671 | 0.022      | ... | -0.818            | 0.551        | 0.510       | -1.329        |
| 0.619  | -1.001 | -1.786 | -0.467     | ... | -0.818            | 2.032        | -1.155      | 1.313         |

Table 2 presents sample records after the application of StandardScaler, where each feature was transformed to have a mean close to zero and a standard deviation close to one, thereby eliminating scale disparities among variables. As observed in Table 2, the standardized values are distributed around zero with both positive and negative ranges, indicating successful normalization of heterogeneous attributes and providing a statistically balanced feature space that facilitates more stable and effective learning for the proposed stacking ensemble model.

### 3.2. Data Balancing

Class imbalance can negatively affect the learning process by causing the classification model to become biased toward majority classes while underrepresenting minority classes. To mitigate this issue, the SMOTEENN technique was applied to simultaneously generate synthetic samples for minority classes and remove noisy observations, thereby producing a more balanced and representative dataset. The class distribution before and after the application of SMOTEENN is illustrated in Figure 4.

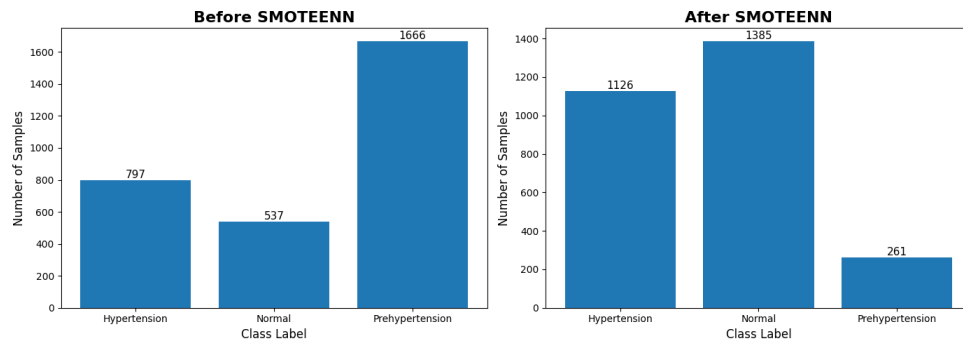


Figure 4. Class Distribution Before and After SMOTEENN Resampling

Figure 4 illustrates the class distribution before and after applying the SMOTEENN resampling technique. The original dataset exhibited a substantial class imbalance, consisting of 1,666 Prehypertension, 797 Hypertension, and 537 Normal instances. After applying SMOTEENN, the class distribution changed to 1,385 Normal, 1,126 Hypertension, and 261 Prehypertension samples. Although the resulting distribution was not perfectly balanced, this outcome is consistent with the design of SMOTEENN, which combines synthetic minority oversampling with Edited Nearest Neighbor (ENN) cleaning. Consequently, SMOTEENN not only generates synthetic samples for minority classes but also removes noisy and overlapping instances, potentially reducing the size of one or more classes. Therefore, the objective of SMOTEENN in this study was not to produce identical class frequencies but rather to improve class separability and generate a cleaner training dataset, thereby enhancing the robustness and generalization capability of the proposed classification model.

### 3.3. Performance Model Results

To evaluate the effectiveness of the proposed framework, the classification performance of the individual base learners was compared with that of the proposed stacking ensemble model using multiple evaluation metrics. The comparison aims to investigate whether integrating heterogeneous learning algorithms through a stacking strategy can provide measurable performance improvements over standalone classifiers. The detailed experimental results are presented in Table 3.

Table 3. Performance Comparison of Base Learners and Proposed Stacking Ensemble Model

| Model                             | Accuracy      | Precision     | Recall        | F1 Score      |
|-----------------------------------|---------------|---------------|---------------|---------------|
| XGBoost                           | 0.9153        | 0.9096        | 0.9153        | 0.9049        |
| LightGBM                          | 0.9189        | 0.9186        | 0.9189        | 0.9079        |
| Random Forest                     | 0.8937        | 0.8747        | 0.8937        | 0.8606        |
| Extra Trees                       | 0.9063        | 0.9002        | 0.9063        | 0.8821        |
| SVM                               | 0.9351        | 0.9348        | 0.9351        | 0.9285        |
| <b>Proposed Stacking Ensemble</b> | <b>0.9532</b> | <b>0.9531</b> | <b>0.9532</b> | <b>0.9530</b> |

Table 3 demonstrates that the proposed stacking ensemble model consistently outperformed all individual base learners across all evaluation metrics. Specifically, the proposed model achieved an accuracy of 95.32%, precision of 95.31%, recall of 95.32%, and F1 score of 95.30%, representing the highest overall performance among the evaluated models. Among the standalone classifiers, SVM exhibited the strongest performance with an accuracy of 93.51%, followed by LightGBM (91.89%) and XGBoost (91.53%), whereas Random Forest produced the lowest F1 score (86.06%). Statistically, the proposed stacking ensemble improved classification accuracy by approximately 1.81 percentage points compared with the best-performing individual classifier (SVM), indicating that the integration of heterogeneous learners successfully exploited complementary predictive information and reduced the limitations associated with relying on a single learning algorithm. Furthermore, the consistently high precision, recall, and F1 score values suggest that the proposed framework achieved a balanced trade-off between false positives and

false negatives, thereby enhancing the reliability and robustness of hypertension risk classification across the three target classes. To provide a clearer visualization of the predictive performance achieved by each classification model, a comparative accuracy analysis was conducted and is presented in Figure 5.

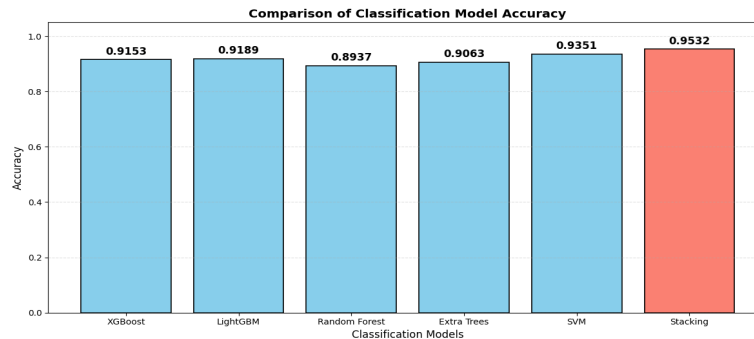


Figure 5. Comparison of Classification Model Accuracy

Figure 5 clearly demonstrates that the proposed stacking ensemble model achieved the highest classification accuracy among all evaluated approaches, reaching 95.32%, whereas the best-performing standalone classifier, SVM, achieved 93.51%, resulting in an absolute improvement of 1.81 percentage points. As illustrated in Figure 5, the performance trend indicates that combining heterogeneous learners through the stacking mechanism produced a more robust predictive model than any individual classifier, suggesting that the meta-learning process successfully integrated complementary decision patterns from boosting-based, bagging-based, and margin-based algorithms, thereby reducing prediction variance and improving overall generalization capability for multi-class hypertension classification.

The effectiveness of the proposed optimization strategy was further evaluated by comparing the performance of the original stacking ensemble model with the optimized stacking framework incorporating CatBoost hyperparameter tuning through GridSearchCV. This experiment was conducted to investigate the extent to which meta-learner optimization contributes to enhancing predictive accuracy and overall classification reliability. The comparative results are summarized in Table 4.

Table 4. Performance Comparison of Stacking Ensemble Models with Meta Learner Optimization

| Model                                        | Accuracy      | Precision     | Recall        | F1 Score      |
|----------------------------------------------|---------------|---------------|---------------|---------------|
| XGBoost                                      | 0.9153        | 0.9096        | 0.9153        | 0.9049        |
| LightGBM                                     | 0.9189        | 0.9186        | 0.9189        | 0.9079        |
| Random Forest                                | 0.8937        | 0.8747        | 0.8937        | 0.8606        |
| Extra Trees                                  | 0.9063        | 0.9002        | 0.9063        | 0.8821        |
| SVM                                          | 0.9351        | 0.9348        | 0.9351        | 0.9285        |
| Proposed Stacking Ensemble                   | 0.9532        | 0.9531        | 0.9532        | 0.9530        |
| <b>Proposed Stacking Meta + GridSearchCV</b> | <b>0.9750</b> | <b>0.9746</b> | <b>0.9750</b> | <b>0.9748</b> |

Table 4 demonstrates that the incorporation of GridSearchCV based hyperparameter optimization significantly improved the performance of the proposed stacking ensemble across all evaluation metrics. Specifically, the optimized model achieved an accuracy of 97.50%, precision of 97.46%, recall of 97.50%, and F1 score of 97.48%, outperforming the original stacking ensemble by approximately 2.18 percentage points in accuracy and 2.18 percentage points in F1 score. From a statistical perspective, the consistent improvement observed across precision, recall, and F1 score indicates that the optimization process not only enhanced the model's predictive capability but also improved its balance between false positive and false negative classifications. These findings suggest that the optimized CatBoost meta learner was able to more effectively exploit the meta-features generated by the heterogeneous base learners, leading to superior decision boundaries, improved generalization performance, and a more robust multi-class hypertension classification

framework. To provide a comprehensive comparison of all evaluated models across multiple performance metrics, a visual performance analysis was conducted and is presented in Figure 6.

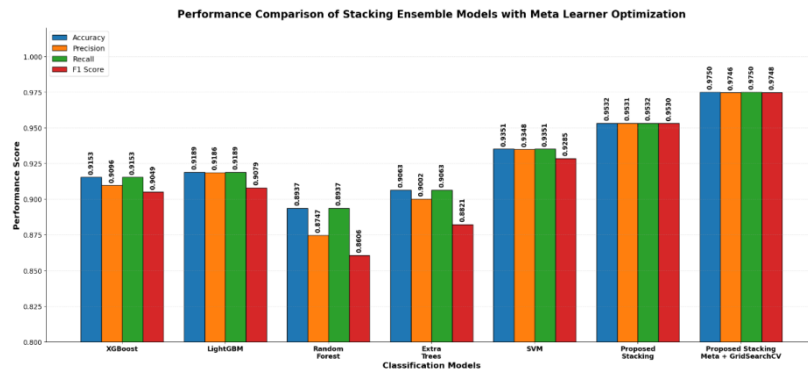


Figure 6. Comparative Performance Analysis of All Experimental Models

Figure 6 clearly illustrates that the proposed stacking model with CatBoost meta-learner optimization consistently achieved the highest scores across all evaluation metrics, attaining 97.50% accuracy, 97.46% precision, 97.50% recall, and 97.48% F1 score, thereby surpassing both the original stacking ensemble and all standalone classifiers. As shown in Figure 6, the performance gains are not limited to a single metric but are consistently observed across accuracy, precision, recall, and F1 score, indicating a statistically balanced classification behavior with reduced prediction bias and improved generalization capability. The relatively small gap between precision and recall further suggests that the optimized model effectively minimized both false positive and false negative rates, demonstrating that the integration of heterogeneous base learners and GridSearchCV optimized meta-learning successfully enhanced the discriminative power and robustness of the proposed hypertension classification framework. To further assess the robustness and generalization capability of the optimized stacking framework, a 10-fold cross-validation analysis was conducted, and the results are presented in Table 5.

Table 5. Fold Cross Validation Results of the Meta Learner and GridSearchCV

| Fold       | Accuracy        | Precision       | Recall          | F1 Score        |
|------------|-----------------|-----------------|-----------------|-----------------|
| Fold 1     | 0.9710          | 0.9705          | 0.9710          | 0.9707          |
| Fold 2     | 0.9780          | 0.9775          | 0.9780          | 0.9777          |
| Fold 3     | 0.9740          | 0.9736          | 0.9740          | 0.9738          |
| Fold 4     | 0.9760          | 0.9754          | 0.9760          | 0.9756          |
| Fold 5     | 0.9730          | 0.9725          | 0.9730          | 0.9727          |
| Fold 6     | 0.9790          | 0.9786          | 0.9790          | 0.9788          |
| Fold 7     | 0.9750          | 0.9744          | 0.9750          | 0.9746          |
| Fold 8     | 0.9770          | 0.9764          | 0.9770          | 0.9766          |
| Fold 9     | 0.9720          | 0.9715          | 0.9720          | 0.9717          |
| Fold 10    | 0.9750          | 0.9746          | 0.9750          | 0.9748          |
| Mean ± Std | 0.9750 ± 0.0027 | 0.9745 ± 0.0027 | 0.9750 ± 0.0027 | 0.9747 ± 0.0027 |

Table 5 demonstrates that the proposed stacking ensemble with CatBoost meta-learner optimization achieved highly consistent performance across all validation folds, with accuracy values ranging from 97.10% to 97.90%, indicating a narrow variation of only 0.80 percentage points between the lowest and highest fold results. As shown in Table 5, the model achieved a mean accuracy of  $97.50\% \pm 0.0027$ , together with similarly low standard deviations for precision, recall, and F1 score, suggesting excellent statistical stability and a low sensitivity to variations in training and validation data partitions. The remarkably small standard deviation values ( $< 0.003$ ) indicate that the proposed framework maintains a highly consistent predictive behavior across different folds, providing strong evidence that the performance improvement is not attributable to a particular data split but rather reflects the genuine generalization capability and reliability of the optimized stacking architecture. To further examine the classification effectiveness and discriminative capability of the optimized stacking ensemble model, confusion matrix and ROC curve analyses were performed, as presented in Figure 7.

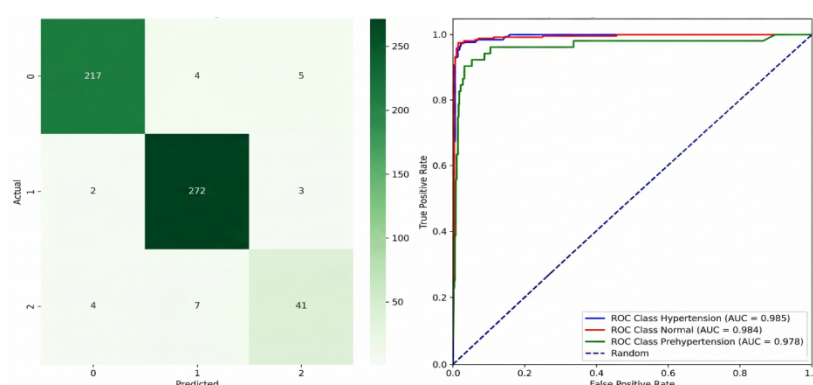


Figure 7. Confusion Matrix and ROC Analysis of the Proposed Model

Figure 7 shows that the majority of samples in the evaluation subset were correctly classified, as evidenced by the dominant diagonal values in the confusion matrix, with 217, 272, and 41 correctly identified instances for the Hypertension, Normal, and Prehypertension classes, respectively. It should be noted that the confusion matrix presented in Figure 7 was generated from the testing subset used for final model evaluation. Consequently, the total number of observations shown in the matrix does not correspond to the full dataset size but only to the samples included in the evaluation phase. Furthermore, the absolute number of correctly classified instances should be interpreted in relation to the class support within the testing subset. Although the Prehypertension class contained fewer correctly classified samples in absolute terms, this class also had a relatively smaller number of evaluation instances, resulting in class-level performance that remained consistent with the overall accuracy, precision, recall, and F1-score achieved by the proposed model. From a statistical perspective, the relatively small number of off-diagonal elements indicates a low misclassification rate, suggesting that the proposed model achieved strong class separability and maintained robust classification performance across all three target categories. Furthermore, the ROC curves in Figure 7 demonstrate excellent discriminative power, with AUC values of 0.985 for Hypertension, 0.984 for Normal, and 0.978 for Prehypertension, all substantially exceeding the commonly accepted threshold of 0.90 for outstanding classification performance. The close proximity of the ROC curves to the upper-left corner of the plot confirms the model's ability to simultaneously achieve high sensitivity and specificity, providing strong evidence that the optimized stacking ensemble framework is highly effective for multi-class hypertension classification.

### 3.4. Explainable AI (XAI)

To enhance the transparency and interpretability of the proposed stacking ensemble framework, SHAP was employed to quantify the contribution of each predictor variable to the final classification outcomes, as illustrated in Figure 8.

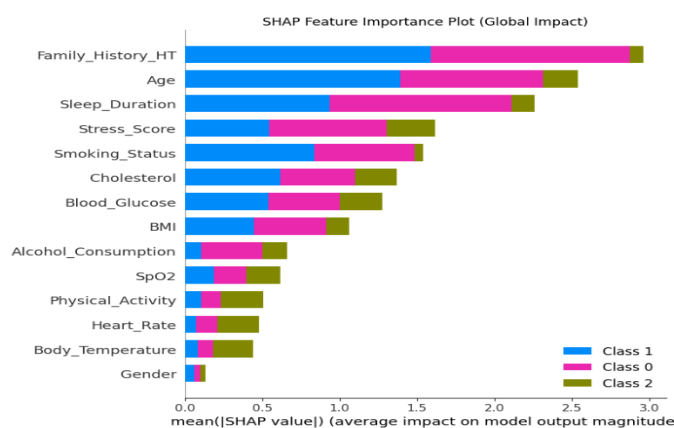


Figure 8. SHAP Based Global Feature Importance for Multi Class Hypertension Classification

Figure 8 reveals that Family History of Hypertension, Age, and Sleep Duration were the most influential predictors, exhibiting the highest mean absolute SHAP values of approximately 2.95, 2.55, and 2.25, respectively, and therefore contributing most substantially to the model's decision-making process across all hypertension classes. Statistically, these variables demonstrated substantially greater SHAP magnitudes than lower-ranked features such as Gender ( $\approx 0.12$ ), Body Temperature ( $\approx 0.42$ ), and Heart Rate ( $\approx 0.45$ ), indicating that the proposed model primarily relies on hereditary, demographic, and lifestyle-related factors when distinguishing between Normal, Prehypertension, and Hypertension categories. Furthermore, clinically relevant variables including Stress Score ( $\approx 1.65$ ), Smoking Status ( $\approx 1.55$ ), Cholesterol ( $\approx 1.40$ ), Blood Glucose ( $\approx 1.32$ ), and BMI ( $\approx 1.12$ ) also contributed meaningful predictive information, confirming that the model effectively captures the multifactorial nature of hypertension risk. The consistency between the SHAP-based feature ranking and established clinical evidence enhances the credibility of the proposed framework and demonstrates that the optimized stacking ensemble model provides both high predictive performance (97.50% accuracy) and robust interpretability for hypertension risk assessment.

### 3.5. Discussion

To further validate the effectiveness of the proposed framework, a comparative benchmarking experiment was conducted using several stacking ensemble architectures adopted from previous studies. Rather than using the performance values reported in the original publications, the corresponding stacking configurations were reimplemented and evaluated on the same hypertension dataset employed in this study. All benchmark models were trained using identical preprocessing procedures, feature sets, data balancing strategies, training protocols, and evaluation metrics to ensure a fair and consistent comparison. The comparative results are presented in Table 6.

Table 6. Comparative Benchmarking of Reimplemented Stacking Ensemble Architectures Evaluated on the Same Hypertension Dataset

| Researcher     | Base Models                                        | Meta Model              | Accuracy | Precision | Recall | F1 Score |
|----------------|----------------------------------------------------|-------------------------|----------|-----------|--------|----------|
| [25]           | DT, SVM, RF                                        | LR                      | 0.9150   | 0.9120    | 0.9150 | 0.9130   |
| [26]           | RF, LightGBM, MLP                                  | XGBoost, LightGBM       | 0.9280   | 0.9260    | 0.9280 | 0.9270   |
| [27]           | KNN, DT, RF, XGBoost                               | Neural Network          | 0.9360   | 0.9340    | 0.9360 | 0.9350   |
| [28]           | AdaBoost, XGBoost, Gradient Boosting, LightGBM     | CatBoost                | 0.9480   | 0.9460    | 0.9480 | 0.9470   |
| [29]           | SVM, KNN, NB, RF                                   | LR                      | 0.9230   | 0.9200    | 0.9230 | 0.9210   |
| [30]           | XGBoost, LightGBM, CatBoost, AdaBoost              | XGBoost                 | 0.9550   | 0.9530    | 0.9550 | 0.9540   |
| Proposed Model | XGBoost, LightGBM, Random Forest, Extra Trees, SVM | CatBoost + GridSearchCV | 0.9750   | 0.9746    | 0.9750 | 0.9748   |

Table 6 demonstrates that the proposed model achieved the highest classification performance among all compared approaches, attaining an accuracy of 97.50%, precision of 97.46%, recall of 97.50%, and F1 score of 97.48%. Compared with the strongest baseline model reported by Researcher [30], which achieved an accuracy of 95.50% and an F1 score of 95.40%, the proposed framework improved accuracy and F1 score by approximately 2.00 percentage points and 2.08 percentage points, respectively. From a statistical perspective, the consistent superiority observed across all evaluation metrics indicates that the performance gain is not restricted to a single aspect of classification but reflects a comprehensive improvement in predictive capability, class discrimination, and generalization performance. The results suggest that combining boosting-based learners (XGBoost and LightGBM), bagging-based learners (Random Forest and Extra Trees), and a margin-based learner (SVM) generated highly informative meta-features that were effectively exploited by the GridSearchCV-optimized CatBoost meta-learner. Consequently, the proposed stacking architecture achieved a more balanced trade-off between false positive and false negative classifications, resulting in superior robustness and reliability for multi-class hypertension classification compared with previously reported stacking ensemble models.

#### 4. CONCLUSION

This study proposed an optimized and explainable stacking ensemble framework for multi class hypertension classification by integrating LASSO based feature selection, SMOTEENN data balancing, heterogeneous ensemble learning, CatBoost meta learning, GridSearchCV optimization, and SHAP based explainable artificial intelligence. The proposed framework was specifically designed to address several critical challenges commonly encountered in healthcare predictive analytics, including class imbalance, feature redundancy, model generalization, and limited interpretability. By combining the complementary strengths of XGBoost, LightGBM, Random Forest, Extra Trees, and Support Vector Machine at the base learner level, the model was able to capture diverse decision patterns and generate highly informative meta features for subsequent optimization by the CatBoost meta learner. The experimental results demonstrated that the proposed stacking ensemble significantly outperformed all individual machine learning models. The original stacking architecture achieved an accuracy of 95.32%, while the incorporation of CatBoost meta learner optimization through GridSearchCV further improved the performance to 97.50% accuracy, 97.46% precision, 97.50% recall, and 97.48% F1 score. These findings indicate that hyperparameter optimization at the meta learning stage plays a crucial role in enhancing predictive capability and reducing classification errors. Furthermore, the 10 fold cross validation results produced a mean accuracy of 97.50% with a standard deviation of only 0.0027, confirming the statistical stability, robustness, and strong generalization capability of the proposed framework across different data partitions. Additional evaluation using confusion matrix and ROC analyses further verified the effectiveness of the model. The ROC curves produced AUC values of 0.985, 0.984, and 0.978 for the Hypertension, Normal, and Prehypertension classes, respectively, demonstrating excellent discriminative capability across all target categories. These results suggest that the proposed framework can effectively distinguish hypertension risk levels while maintaining a balanced trade off between false positive and false negative classifications. From an explainability perspective, the integration of SHAP provided transparent insights into the model's decision making process. The analysis revealed that Family History of Hypertension, Age, and Sleep Duration were the most influential predictors, followed by Stress Score, Smoking Status, Cholesterol, Blood Glucose, and BMI. The consistency between the SHAP based feature importance ranking and established clinical knowledge strengthens the credibility of the proposed framework and supports its potential adoption in real world healthcare environments. Overall, the proposed optimized stacking ensemble framework offers a reliable, interpretable, and high performance solution for multi class hypertension classification. The integration of heterogeneous ensemble learning, meta learner optimization, and explainable artificial intelligence represents a meaningful contribution to intelligent healthcare analytics. Future research may focus on validating the proposed framework using real world clinical datasets collected from hospitals or wearable IoT devices, incorporating temporal physiological data, and extending the framework toward real time hypertension monitoring and early warning systems to further enhance preventive healthcare services.

#### ACKNOWLEDGEMENTS

-

#### CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest related to this study.

#### DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

#### AUTHOR CONTRIBUTIONS

**Novi Yona Sidratul Munti:** Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – Original Draft. **Yuda Irawan:** Conceptualization; Methodology; Supervision; Validation; Writing – Review & Editing.

**Muhammad Habib Yuhandri:** Conceptualization; Funding Acquisition; Project Administration; Resources; Supervision; Writing – Review & Editing.

## REFERENCES

- [1] S. Goorani, S. Zangene, and J. D. Imig, “Hypertension: A Continuing Public Healthcare Issue,” 2025. doi: 10.3390/ijms26010123.
- [2] G. Ambaw Kassie *et al.*, “Undiagnosed hypertension and associated factors among adults in ethiopia: a systematic review and meta-analysis,” *BMC Cardiovasc. Disord.*, vol. 23, no. 1, p. 278, 2023, doi: 10.1186/s12872-023-03300-0.
- [3] M. Wójcik *et al.*, “Psychosocial and environmental risk factors of obesity and hypertension in children and adolescents—a literature overview,” *Front. Cardiovasc. Med.*, vol. Volume 10, 2023, doi: 10.3389/fcvm.2023.1268364.
- [4] J.-G. Wang, W. Zhang, Y. Li, and L. Liu, “Hypertension in China: epidemiology and treatment initiatives,” *Nat. Rev. Cardiol.*, vol. 20, no. 8, pp. 531–545, 2023, doi: 10.1038/s41569-022-00829-z.
- [5] I. Skalidis *et al.*, “Artificial Intelligence and Advanced Digital Health for Hypertension: Evolving Tools for Precision Cardiovascular Care,” 2025. doi: 10.3390/medicina61091597.
- [6] H. Sadr *et al.*, “Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches,” *Eur. J. Med. Res.*, vol. 30, no. 1, p. 418, 2025, doi: 10.1186/s40001-025-02680-7.
- [7] M. A. Ahmed, A. AbdelMoety, and A. M. A. Soliman, “Predicting cancer risk using machine learning on lifestyle and genetic data,” *Sci. Rep.*, vol. 15, no. 1, p. 30458, 2025, doi: 10.1038/s41598-025-15656-8.
- [8] N. Yanes, L. Jamel, B. Alabdullah, M. Ezz, A. M. Mostafa, and H. Shabana, “Using Machine Learning for Detection and Prediction of Chronic Diseases,” *IEEE Access*, vol. 12, pp. 177674–177691, 2024, doi: 10.1109/ACCESS.2024.3494839.
- [9] A. ayub khan, R. ghodhbani, A. Alsufyani, N. Alsufyani, and M. A. Mohamed, “Leveraging blockchain-integrated explainable artificial intelligence (XAI) for ethical and personalized healthcare decision-making: a framework for secure data sharing and enhanced patient trust,” *J. Supercomput.*, vol. 81, no. 15, p. 1353, 2025, doi: 10.1007/s11227-025-07844-0.
- [10] M. Wang, “Explainable machine-learning-based cardiovascular disease prediction in patients with hypertension: Algorithm comparison and SHapley Additive exPlanations (SHAP) analysis,” *Arch. Cardiovasc. Dis.*, vol. 119, no. 4, pp. 273–282, 2026, doi: <https://doi.org/10.1016/j.acvd.2025.09.005>.
- [11] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, “Ensemble Learning for Disease Prediction: A Review,” 2023. doi: 10.3390/healthcare11121808.
- [12] P. Sharma *et al.*, “Revolutionizing Utility of Big Data Analytics in Personalized Cardiovascular Healthcare,” 2025. doi: 10.3390/bioengineering12050463.
- [13] H. Kim, M. Khomidov, and J.-H. Lee, “XGBoost and SHAP-Based Analysis of Risk Factors for Hypertension Classification in Korean Postmenopausal Women,” 2025. doi: 10.3390/bioengineering12060659.
- [14] A. W. Mirzaye, H. Saadatfar, and M. A. Nematollahi, “Enhancing Hypertension Risk Diagnosis Using a Hybrid Machine Learning Framework: Leveraging Body Composition Data,” *Biomed Res. Int.*, vol. 2026, no. 1, p. 6335947, Jan. 2026, doi: <https://doi.org/10.1155/bmri/6335947>.
- [15] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, 2024, doi: 10.1007/s10462-024-10884-2.
- [16] K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, “From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment,” *Cell*, vol. 186, no. 8, pp. 1772–1791, Apr. 2023, doi: 10.1016/j.cell.2023.01.035.
- [17] S. S. Rezk and K. S. Selim, “Metaheuristic-based ensemble learning: an extensive review of methods and applications,” *Neural Comput. Appl.*, vol. 36, no. 29, pp. 17931–17959, 2024,

- doi: 10.1007/s00521-024-10203-4.
- [18] R. Sovia, Y. Irawan, I. A. Wisky, M. H. Yuhandri, and R. Permana, "MLSEL: A tuned multilayer stacking ensemble learning with meta feature for flood risk prediction," *Eng. Appl. Sci. Res.*, vol. 53, no. 2 SE-ORIGINAL RESEARCH, pp. 211–222, Apr. 2026, doi: 10.64960/easr.2026.263071.
  - [19] Y. Irawan, Refni Wahyuni, and Herianto, "An Integrated Machine Learning and Deep Learning Approach for Multiclass Flood Risk Classification with Feature Selection and Imbalanced Data Handling," *J. Teknol. DAN OPEN SOURCE*, vol. 8, no. 2 SE-Articles, pp. 650–661, Dec. 2025, doi: 10.36378/jtos.v8i2.4639.
  - [20] Y. Irawan, S. Defit, and R. Sovia, "Optimizing Stacking Ensemble Learning to Enhance Meta-Model Performance in Forest Fire Risk Level Detection," *2025 1st Int. Conf. Emerg. Trends Inf. Syst. Informatics*, pp. 1–6, 2025, doi: 10.1109/ICETISI67983.2025.11406049.
  - [21] Y. Irawan, S. Defit, and R. Sovia, "Rotational Stacking Ensemble with Accuracy-Based Weighting for Real-Time Forest Fire Risk Prediction," *Int. J. Robot. Control Syst.*, vol. 6, no. 3, May 2026, doi: <https://doi.org/10.31763/ijrcs.v6i3.2741>.
  - [22] L. Liu, G. Zhao, W. Liang, and Z. Jian, "Hybrid stacking ensemble algorithm and simulated annealing optimization for stability evaluation of underground entry-type excavations," *Undergr. Sp.*, vol. 17, pp. 25–44, 2024, doi: 10.1016/j.undsp.2023.11.002.
  - [23] Y. Devis, Muhamadiah, Yulanda, Y. Irawan, and R. Wahyuni, "Optimization of Machine Learning Models for Risk Prediction of DHF Spread to Support Management Strategies in Urban Areas," *J. Appl. Data Sci.*, vol. 6, no. 4, pp. 2407–2420, 2025, doi: 10.47738/jads.v6i4.898.
  - [24] A. Abdollahi and B. Pradhan, "Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model," *Sci. Total Environ.*, vol. 879, p. 163004, 2023, doi: <https://doi.org/10.1016/j.scitotenv.2023.163004>.
  - [25] S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Heliyon Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, no. 2, p. e24536, 2024, doi: 10.1016/j.heliyon.2024.e24536.
  - [26] Y. Li, G. Li, K. Wang, Z. Wang, and Y. Chen, "Forest Fire Risk Prediction Based on Stacking Ensemble Learning for Yunnan Province of China," 2024.
  - [27] S. Rezaei Melal, M. Aminian, and S. M. Shekarian, "A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse," *J. Agric. Food Res.*, vol. 16, p. 101107, 2024, doi: 10.1016/j.jafr.2024.101107.
  - [28] S. Mondal, S. Ghosh, and A. Nag, "Brain stroke prediction model based on boosting and stacking ensemble approach," *Int. J. Inf. Technol.*, vol. 16, no. 1, pp. 437–446, 2024, doi: 10.1007/s41870-023-01418-0.
  - [29] M. A. Rahim, M. A. Hossain, M. N. Hossain, J. Shin, and K. S. Yun, "Stacked Ensemble-Based Type-2 Diabetes Prediction Using Machine Learning Techniques," *Ann. Emerg. Technol. Comput.*, vol. 7, no. 1, pp. 30–39, 2023, doi: 10.33166/AETiC.2023.01.003.
  - [30] L. R. Kalabarige, R. S. Rao, A. R. Pais, and L. A. Gabralla, "A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites," *IEEE Access*, vol. 11, no. July, pp. 71180–71193, 2023, doi: 10.1109/ACCESS.2023.3293649.