



Cross-Dataset Evaluation of Boosting Models for Hypertension Prediction

Bety Wulan Sari^{1*}, Dewi Ayu Murtiningsih², Donni Prabowo³, Yoga Pristyanto⁴, Ika Nur Fajri⁵, Ike Verawati⁶
^{1,2,3,4,5}Department of Information System, Faculty of Computer Science, Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia

⁶Department of Informatics, Faculty of Computer Science, Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia

*bety@amikom.ac.id

Abstract

Hypertension remains a significant global risk factor for cardiovascular disease and related mortality, necessitating reliable early risk prediction models. Although boosting algorithms have demonstrated strong performance in structured medical data, limited studies have examined their consistency across heterogeneous datasets. This study aims to evaluate the cross-dataset performance and stability of three boosting models, such as XGBoost, LightGBM, and CatBoost, for hypertension prediction under multiple train–test split ratios. Two independent structured datasets were analyzed using 60:40, 70:30, 80:20, and 90:10 splits. To identify the optimal hyperparameters, grid search was performed using repeated stratified 5-fold cross-validation with three repetitions. Model effectiveness was measured using the evaluation metrics of accuracy, precision, recall, F1-score, and AUC. Results show that Dataset 1 gained consistently high predictive performance (accuracy > 0.98; AUC ≈ 1.00), indicating strong and well-separated predictive signals, whereas Dataset 2 demonstrated substantially lower discriminative ability (accuracy ≈ 0.71–0.72; AUC ≈ 0.50), suggesting limited predictive structure. Across both datasets, CatBoost consistently obtained the highest accuracy, particularly at the 90:10 split ratio. These findings demonstrate that dataset characteristics critically determine model effectiveness and that among the evaluated boosting algorithms, CatBoost delivered the strongest overall predictive performance.

Keywords: AUC; boosting algorithms; cross-dataset evaluation; hypertension prediction; hyperparameter tuning

How to Cite: B. W. Sari, D. A. Murtiningsih, D. Prabowo, Y. Pristyanto, I. N. Fajri, and I. Verawati, “Cross-Dataset Evaluation of Boosting Models for Hypertension Prediction”, *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 888 - 897, Jul. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i4.7620>

Received: March 3, 2026

Accepted: June 21, 2026

Available Online: July 23, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

The global burden of cardiovascular disease is strongly influenced by hypertension, which continues to be a major public health concern [1]. Early identification of individuals at risk is essential to support preventive strategies and informed clinical decision-making. In this context, the use of modifiable, non-invasive risk factors, such as body mass index (BMI), smoking status, alcohol intake, and physical activity level, offers practical advantages, as these variables can be easily obtained without laboratory testing. This makes them particularly suitable for cost-effective, broad-scale screening, especially where medical resources are limited. Machine learning has been extensively employed to enhance hypertension risk prediction models [2]. Specifically, gradient boosting methods such as XGBoost, LightGBM, and CatBoost have shown promising results on structured tabular medical datasets owing to their ability to capture complex

nonlinear patterns [3-5]. Prior studies have reported promising results using these approaches, as well as other machine learning models, depending on the dataset characteristics and feature selection [6-10]. These findings also highlight the potential of non-invasive variables to support early hypertension screening in real-world settings. However, existing studies present several important limitations. Most research relies on a single dataset, which restricts the evaluation of model robustness and generalizability. In addition, some studies lack rigorous validation strategies, such as cross-validation or external dataset testing, which may lead to biased or overestimated performance [6-10]. Other limitations include dataset-specific constraints, such as restricted populations, missing key variables, and imbalanced data distributions, all of which can significantly influence model outcomes [8], [9], [11].

Importantly, model performance has been shown to vary depending on dataset characteristics. For example, Kurniawan et al. [11] reported that Logistic Regression outperformed more complex models in an Indonesian dataset, indicating that predictive performance is highly data-dependent. These findings suggest that strong performance on a single dataset does not necessarily indicate robustness to unseen datasets across different populations or data distributions.

Despite these advances, systematic cross-dataset evaluation remains limited, particularly for boosting-based models in hypertension prediction. It remains unclear which algorithm provides the most stable and consistent performance across datasets with different characteristics. This issue is especially relevant for population-specific applications, as models developed using global or simulated datasets may not fully capture local demographic and lifestyle patterns, including those of the Indonesian population.

To overcome this research gap, this study undertakes a systematic cross-dataset evaluation of boosting-based models for hypertension risk classification. Specifically, this study compares the performance of XGBoost, LightGBM, and CatBoost across two independent datasets with different characteristics. The analysis focuses on model performance, stability across multiple train-test split ratios, and generalizability between datasets. By emphasizing cross-dataset validation and the use of non-invasive risk factors, the objective of this study is to enhance the understanding of how machine learning models can be effectively applied to hypertension prediction, including their relevance to the Indonesian context.

The remainder of this paper is organized as follows. Section 2 describes the methodology and experimental design. Section 3 presents the results and discussion, including cross-dataset performance analysis. Finally, Section 4 concludes the study and outlines future research directions.

2. Methods

This paper uses a cross-dataset experimental framework to examine the generalization ability of boosting-based machine learning models for predicting hypertension. Data collection, EDA, data preprocessing, addressing data imbalance, feature selection, model training, hyperparameter tuning, evaluation, and model interpretability are all included in the overall procedure. Figure 1 depicts the methodological framework.

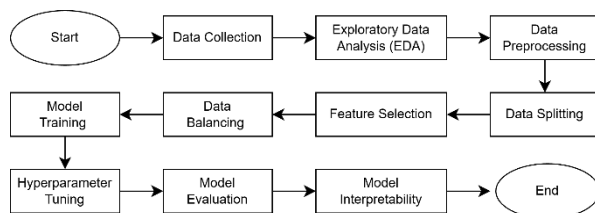


Figure 1. Flowchart of the research experiment

Two separate datasets are used for model evaluation to assess the model's generalizability across different data distributions and feature sets.

2.1 Data Collection

Clinical and demographic information on hypertension risk is included in the structured CSV file used for the analysis. The data-loading process is used to access clinical and demographic hypertension risk data in an analytical setting. To preserve structural visibility and prevent shortened data in exploratory analysis, every column was made visible during the initial examination. Let the dataset be defined as Equation 1.

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \quad (1)$$

$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$ denotes the feature vector of the i -th individual and $y_i \in 0,1$ denotes the hypertension label. Additionally, feature names were translated from Spanish to English using a predefined mapping dictionary to standardize terminology and improve interpretability.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was used to explore data distribution and statistical properties prior to modeling. Class distribution analysis was performed using frequency counts, as shown in Equation 2.

$$P(y = c) = \frac{n_c}{n} \quad (2)$$

n_c denotes the number of samples in class c .

A correlation matrix was computed using Pearson's correlation coefficient, as defined in Equation 3.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3)$$

Among the traits, this analysis finds multicollinearity and potential redundancy. To examine how the characteristics varied across the various classes, summary statistics, including means, standard deviations, and percentage distributions, were generated.

2.3 Data Preprocessing

Data preprocessing prepares the dataset for machine learning modelling. This stage included column renaming, removal of non-informative features, and verification of data types [12].

Feature scaling was considered to normalize numerical ranges where appropriate. Standardization transformed each feature, as expressed in Equation 4.

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

μ represents the feature mean and σ the standard deviation. This transformation ensures that features contribute proportionally during model optimization.

The target variable was converted into a binary outcome, as defined in Equation 5.

$$y = \begin{cases} 1, & \text{hypertensive} \\ 0, & \text{non-hypertension} \end{cases} \quad (5)$$

2.4 Data Splitting

The dataset was split into training and test sets to ensure unbiased model evaluation. The feature matrix X and target vector y were separated, as defined in Equation 6.

$$X = [x_{ij}]_{n \times p}, y = [y_i]_{n \times 1} \quad (6)$$

Stratified sampling was applied to preserve class proportions [13], as described in equation 7.

$$P_{train}(y) \approx P_{test}(y) \quad (7)$$

To evaluate robustness and sensitivity to training size, multiple split ratios were explored: 60:40, 70:30, 80:20, and 90:10. This approach ensures that model performance is not dependent on a single partition and provides a more reliable assessment of generalization across different data allocations.

2.5 Feature Selection

To improve model generalization and reduce dimensionality, three feature selection strategy was implemented using three independent methods.

Boruta selects relevant features by comparing their importance against randomized shadow features [14]. Feature importance from Random Forest was calculated as defined in Equation 8.

$$Imp_j = \frac{1}{T} \sum_{t=1}^T I_{j,t} \quad (8)$$

T denotes the number of trees and $I_{j,t}$ the importance of feature j in tree t .

Recursive Feature Elimination (RFE) iteratively removes the least important features based on model importance ranking until a predefined number of features is reached [15], [16]. SelectKBest employed the ANOVA F-test to rank features [17], as formulated in Equation 9.

$$F = \frac{\text{Between-class variance}}{\text{Within-class variance}} \quad (9)$$

A voting mechanism was applied, retaining features selected by two or more methods (as defined in Equation 10) for model development.

$$F_{final} = f_j: f_j \in \text{at least 2 methods} \quad (10)$$

This ensemble selection strategy enhances feature stability and robustness.

2.6 Data Balancing

Given class imbalance, synthetic oversampling techniques were applied to the training data. SMOTE generated new minority samples using linear interpolation [18], as expressed in Equation 11.

$$\mathbf{x}_{new} = \mathbf{x}_i + \lambda(\mathbf{x}_{zi} - \mathbf{x}_i), \lambda \sim U(0,1) \quad (11)$$

Oversampling was restricted to the training set to prevent information leakage.

2.7 Model Training

Three classification models were evaluated: XGBoost, LightGBM, CatBoost, the additive framework of gradient boosting is defined in Equation 12.

$$\hat{y}_i = \sum_{t=1}^T f_t(\mathbf{x}_i) \quad (12)$$

The objective function minimized during training is defined in equation 13. This objective consists of the empirical loss component and a regularization term that controls model complexity to prevent overfitting.

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \Omega(f) \quad (13)$$

The loss function l corresponds to the logistic loss, which is formulated in Equation 14.

$$l(y, \hat{y}) = -[y \log(\hat{p}) + (1 - y) \log(1 - \hat{p})] \quad (14)$$

This study employs a comparative cross-dataset evaluation approach, in which each dataset is used independently for model training and testing. The performance results obtained from each dataset are then compared to assess model consistency and generalizability across datasets with different characteristics.

For each dataset, the data are split into training and testing sets using multiple ratios (60:40, 70:30, 80:20, and 90:10) to evaluate model stability under different data distributions. Each model is trained and tested within the same dataset to ensure a fair comparison.

Three boosting-based algorithms, namely XGBoost, LightGBM, and CatBoost, are implemented and evaluated on both datasets using the same experimental settings. Model performance is assessed using standard evaluation metrics, including accuracy, precision, recall, F1-score, and AUC.

2.8 Hyperparameter Tuning

Hyperparameter optimization was conducted using GridSearchCV combined with Repeated Stratified K-Fold cross-validation [19], [20].

The dataset is partitioned into k folds for cross-validation, as described in Equation 15.

$$\mathcal{D} = \bigcup_{i=1}^k \mathcal{D}_i \quad (15)$$

The optimal set of parameters was identified through systematic search and evaluation, as defined in Equation 17.

$$\theta^* = \arg \max_{\theta \in \Theta} \text{ROC-AUC}(\theta) \quad (16)$$

Repeated cross-validation reduces variance in performance estimation and improves robustness.

2.9 Model Evaluation

The optimized models were evaluated on the independent test set. The primary assessment metrics include accuracy, precision, recall, and the F1-score [21].

Receiver Operating Characteristic (ROC) curve analysis was used to evaluate the model's discrimination ability across varying classification thresholds [22], [23] as expressed in Equation 17.

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN} \quad (17)$$

The AUC quantifies overall classification performance [24].

2.10 Model Interpretability

To enhance transparency, SHAP (SHapley Additive exPlanations) was applied to the best-performing model [25], [26]. SHAP summary plots (beeswarm plots) were generated to visualize the global impact of features across all samples, illustrating both the magnitude and direction of their contributions. In addition, SHAP feature importance bar plots were used to rank features by average absolute contribution, providing insight into the most influential predictors of hypertension risk.

3. Results and Discussions

3.1 Dataset Characteristics and Class Distribution

This section presents the statistical characteristics of both datasets after preprocessing. Two datasets were used to evaluate the generalizability of the proposed models.

Dataset 1 was derived from the National Health and Nutrition Survey (Encuesta Nacional de Salud y Nutrición – ENSANUT) conducted by the Instituto Nacional de Salud Pública, Mexico. The dataset was accessed through the Kaggle repository titled “Hipertensión Arterial México” (<https://www.kaggle.com/datasets/frederickfelix/hipertension-arterial-mxico>), which aggregates three ENSANUT-based datasets. It represents real-world survey data and primarily contains anthropometric, clinical, and biochemical indicators associated with hypertension risk.

Dataset 2 was obtained from the Kaggle repository titled “Hypertension Risk Prediction Dataset” (<https://www.kaggle.com/datasets/ankushpanday1/hypertension-risk-prediction-dataset>). This dataset includes 174,982 records with 24 variables and is synthetically generated to simulate hypertension risk based on demographic, clinical, and lifestyle attributes from multiple countries, including Indonesia. Although the dataset is multinational, the inclusion of Indonesian samples enables preliminary analysis of model applicability in the Indonesian context. However, due to its simulated nature, the results should be interpreted with caution for population-specific conclusions.

Despite originating from different sources, both datasets were used to evaluate model robustness across varying feature structures and sample sizes. Table 1 summarizes the dataset statistics, including dataset size, feature count, and class distribution.

Table 1. Datasets Statistic

Metric	Dataset 1	Dataset 2
Total Samples	4363	174982
Number of Features	36	23
Class 0	1547	49201
Class 1	2816	125781

3.2 Exploratory Data Analysis (EDA) Results

Exploratory analysis confirmed the absence of duplicate records and structural inconsistencies. Correlation analysis revealed moderate associations among lipid-related variables and blood pressure indicators, consistent with medical knowledge. Class distribution analysis showed a noticeable imbalance in both datasets, with hypertensive cases dominating. This justified the need for oversampling strategies. Figure 2 illustrates the class distribution before balancing for both datasets.

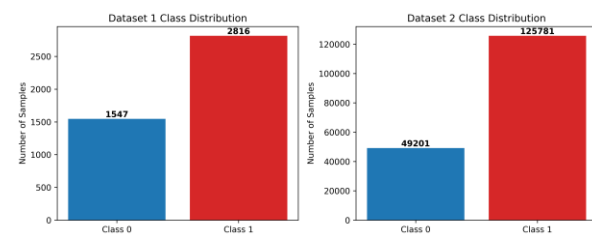


Figure 2. Class Distribution

The descriptive statistics indicate that Dataset 1 exhibits greater variability due to its larger sample size, whereas Dataset 2 demonstrates more compact statistical dispersion. These structural differences are expected to influence model generalization performance.

Figure 3 reports the distribution of the Age feature in the Hypertension Dataset and the Mexico Hypertension Dataset. The Hypertension Dataset exhibits a relatively uniform distribution, indicating that age groups are evenly represented. In contrast, the Mexico dataset follows a bell-shaped distribution, with a higher concentration of individuals in the middle-age range (around 45–60 years).

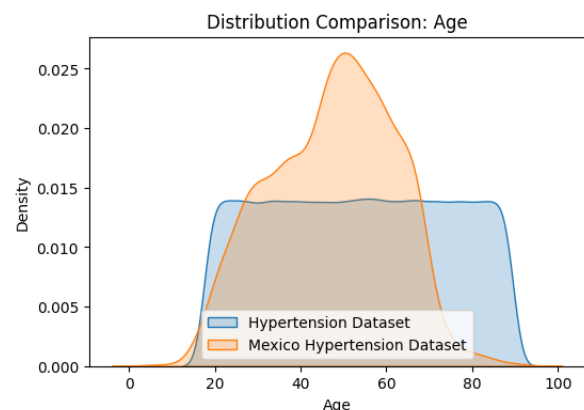


Figure 3. Distribution Comparison: Age

This difference indicates a distribution shift between the datasets, which may affect model generalization, as models trained on uniformly distributed data may not

perform optimally on more naturally distributed populations.

3.3 Data Preprocessing Results

To refine the feature sets for predictive modeling, non-informative and administrative variables were removed. In Dataset 1, the FOLIO_I feature was dropped because it served primarily as an internal identifier and lacked clinical relevance. Similarly, in Dataset 2, the Country feature was excluded to ensure the model's predictions were based solely on biological and behavioral risk factors, rather than on geographic labels that could introduce regional bias.

Following these structural adjustments, the data underwent numerical transformation to prepare for the feature selection and training stages. A StandardScaler was applied to all numeric features (type int64 and float64) to ensure they followed a standard normal distribution.

3.4 Data Splitting

To assess model robustness, both datasets were evaluated across several stratified train-test split ratios, including 60:40, 70:30, 80:20, and 90:10. This approach was intended to determine whether model performance and ranking were sensitive to variations in the training data proportion.

Across all configurations, performance trends remained consistent. Although larger training proportions slightly improved accuracy and F1-score, the relative ranking of the boosting models did not change. This confirms that the comparative results are not dependent on a single random partition.

In all experiments, stratified sampling was employed to maintain the original class distribution in both the training and test sets, thereby ensuring a representative and unbiased evaluation.

3.5 Feature Selection Results

The hybrid feature selection strategy combining Boruta, RFE, and SelectKBest resulted in a refined subset of predictive features. Each method selected a partially overlapping set of variables, indicating both shared and method-specific relevance patterns.

The voting mechanism retained features selected by at least two methods, ensuring robustness and reducing bias from any single selection approach. Table 2 compares selected features across the three methods.

Based on the comparisons in Table 2, a collective voting mechanism was used to filter the most significant variables. Table 3 shows the final voting-based feature subset used for model training.

The selected features include clinically relevant predictors such as age, lipid profile indicators, blood pressure measurements, and lifestyle-related variables. Their consistent selection across methods reinforces their predictive importance for hypertension risk.

Table 2. Selected Features

Dataset 1	
Boruta Selected Features	Age, LDL, Triglyceride Level, Body Weight, Height, Waist Circumference, Body Weight (2nd), Calf Circumference, Waist Circumference (2nd), Blood Pressure, BMI, Total Physical Activity
RFE Selected Features	Age, Total Cholesterol Level, Triglyceride Level, Body Weight, Height, Waist Circumference, Waist Circumference (2nd), Blood Pressure, BMI, Total Physical Activity
SelectKBest Selected Features	Age, Triglyceride Level, C-Reactive Protein, Body Weight, Height, Waist Circumference, Body Weight (2nd), Calf Circumference, Waist Circumference (2nd), BMI
Dataset 2	
Boruta Selected Features	Age, BMI, Cholesterol, Systolic_BP, Diastolic_BP, Smoking_Status, Alcohol_Intake, Physical_Activity_Level, Stress_Level, Salt_Intake, Sleep_Duration, Heart_Rate, LDL, HDL, Triglycerides, Glucose, Education_Level, Employment_Status
RFE Selected Features	Age, Cholesterol, Systolic_BP, Smoking_Status, Physical_Activity_Level, Stress_Level, Sleep_Duration, LDL, Education_Level, Employment_Status
SelectKBest Selected Features	BMI, Cholesterol, Systolic_BP, Diastolic_BP, Smoking_Status, Family_History, Sleep_Duration, HDL, Glucose, Employment_Status

Table 3. Final Voting-based Feature

Dataset	Final Selected Features	Total Features
Dataset 1	Age, Triglyceride Level, Body Weight, Height, Waist Circumference, Body Weight (Second Measurement), Calf Circumference, Waist Circumference (Second Measurement), Blood Pressure, BMI, Total Physical Activity	11
Dataset 2	Age, BMI, Cholesterol, Systolic_BP, Diastolic_BP, Smoking_Status, Physical_Activity_Level, Stress_Level, Sleep_Duration, LDL, HDL, Glucose, Education_Level, Employment_Status	14

3.6 Data Balancing Techniques

To address the inherent class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was limited to the training data to avoid information leakage. This process generated synthetic observations for the hypertensive class, successfully equalizing class distributions within the training set without altering or biasing the test set's original distribution. Table 4 summarizes the class proportions before and after this oversampling process.

Table 4. Class Distribution Before and After SMOTE

Dataset	Class	Before SMOTE (Imbalanced)	After SMOTE (Balanced)
Dataset 1	Class 0	928	1689
	Class 1	1689	1689
Dataset 2	Class 0	34441	88046
	Class 1	88046	88046

Performance comparisons indicated that models trained on the balanced data exhibited significant improvements in Recall and F1-score compared to models trained on the raw, imbalanced data. By providing the algorithms with an equal number of

samples for both classes, SMOTE allowed the models to better learn the specific patterns of the minority class. This resulted in a more robust decision boundary, which is critical for medical classification where the cost of missing a hypertensive case is high.

3.7 Model Training Results

Three boosting models were evaluated: XGBoost, LightGBM, and CatBoost. Table 5 depicts the performance comparison of the boosting models across multiple split ratios for Dataset 1. Overall, all three models demonstrated exceptionally high performance across all evaluation metrics. Accuracy consistently exceeded 0.98 for all split configurations, while AUC values approached 1.0, indicating near-perfect discriminative capability. The results remained stable across varying train-test proportions (60:40 to 90:10), suggesting that the models are robust to different data partition strategies. Among the evaluated methods, CatBoost ranked first in terms of accuracy and F1-score at the 90:10 split ratio, although the performance differences among the boosting algorithms were marginal.

Table 5. Performance Comparison of Boosting Models Across Multiple Split Ratios (Dataset 1)

Model	Accuracy	Precision	Recall	F1-Score	AUC
Split Ratio : 60:40					
XGBoost	0.9868	0.9946	0.9849	0.9997	0.9997
LGBM	0.9885	0.9982	0.9840	0.9911	0.9974
CatBoost	0.9845	0.9955	0.9805	0.9879	0.9980
Split Ratio : 70:30					
XGBoost	0.9878	0.9964	0.9846	0.9905	0.9990
LGBM	0.9893	0.9976	0.9858	0.9917	0.9986
CatBoost	0.9924	0.9976	0.9905	0.9941	0.9987
Split Ratio : 80:20					
XGBoost	0.9885	0.9885	0.9885	0.9885	0.9885
LGBM	0.9964	0.9964	0.9964	0.9964	0.9964
CatBoost	0.9858	0.9858	0.9858	0.9858	0.9858
Split Ratio : 90:10					
XGBoost	0.9931	1.0000	0.9894	0.9947	0.9996
LGBM	0.9931	1.0000	0.9894	0.9947	0.9996
CatBoost	0.9954	1.0000	0.9929	0.9964	0.9995

Table 6. Performance Comparison of Boosting Models Across Multiple Split Ratios (Dataset 2)

Model	Accuracy	Precision	Recall	F1-Score	AUC
Split Ratio : 60:40					
XGBoost	0.6995	0.7184	0.9570	0.8207	0.5037
LGBM	0.7186	0.7187	0.9997	0.8362	0.4973
CatBoost	0.7170	0.7187	0.9960	0.8349	0.4958
Split Ratio : 70:30					
XGBoost	0.7070	0.7201	0.9689	0.8262	0.5120
LGBM	0.7186	0.7187	0.9997	0.8362	0.5091
CatBoost	0.7170	0.7187	0.9960	0.8349	0.5093
Split Ratio : 80:20					
XGBoost	0.7085	0.7191	0.9755	0.8279	0.5045
LGBM	0.7188	0.7188	0.9998	0.8363	0.5000
CatBoost	0.7172	0.7185	0.9972	0.8352	0.5065
Split Ratio : 90:10					
XGBoost	0.7058	0.7175	0.9741	0.8263	0.5016
LGBM	0.7188	0.7188	1.0000	0.8363	0.5014
CatBoost	0.7178	0.7192	0.9963	0.8354	0.4890

In contrast, Table 6 shows substantially lower performance on Dataset 2, with accuracy around 0.70–0.72 and AUC values close to 0.50, indicating limited

discriminative ability. Although LightGBM and CatBoost slightly outperform XGBoost, the overall performance pattern remains consistent across split ratios.

Notably, all models exhibit very high recall (≈ 0.98 –1.00) but near-random AUC, indicating that the model tended to overpredict the positive class. This indicates weak feature-label relationships or potential class imbalance, which limits the models' ability to effectively distinguish between classes. This further highlights the strong dependency of model performance on dataset characteristics.

3.8 Hyperparameter Tuning Results

Grid search combined with repeated stratified 5-fold cross-validation (three repetitions) was used to optimize model performance and stability. This approach ensures that the selected hyperparameters are robust to sampling variability while maintaining consistent class distributions across folds. The predefined hyperparameter search space for each model is described as follows.

For XGBoost and LightGBM, the hyperparameter tuning explored `n_estimators` (200, 400), `max_depth` (4–8 for XGBoost and -1 to 8 for LightGBM), and `learning_rate` (0.05–0.1), along with `subsample` and `colsample_bytree` values ranging from 0.8 to 1.0. For CatBoost, the tuning considered iterations (200, 400), `depth` (6–10), and `learning_rate` (0.05–0.1).

The predefined hyperparameter search space for each boosting model is summarized in Table 7.

Table 7. Comparison of Optimal Hyperparameters Across Datasets

Model	Configuration Details
XGBoost	Dataset 1
	Best ROC-AUC: 0.9977
	Best Parameters: <code>n_estimators</code> = 200; <code>max_depth</code> = 6; <code>learning_rate</code> = 0.05; <code>subsample</code> = 1.0; <code>colsample_bytree</code> = 1.0
	Dataset 2
LightGBM	Best ROC-AUC: 0.8145
	Best Parameters: <code>n_estimators</code> = 200; <code>max_depth</code> = 8; <code>learning_rate</code> = 0.1; <code>subsample</code> = 0.8; <code>colsample_bytree</code> = 1.0
	Dataset 1
	Best ROC-AUC: 0.9974
CatBoost	Best Parameters: <code>n_estimators</code> = 200; <code>max_depth</code> = -1; <code>learning_rate</code> = 0.05; <code>subsample</code> = 0.8; <code>colsample_bytree</code> = 0.8
	Dataset 2
	Best ROC-AUC: 0.8102
	Best Parameters: <code>n_estimators</code> = 200; <code>max_depth</code> = 6; <code>learning_rate</code> = 0.1; <code>subsample</code> = 0.8; <code>colsample_bytree</code> = 0.8
CatBoost	Dataset 1
	Best ROC-AUC: 0.9977
	Best Parameters: <code>iterations</code> = 200; <code>depth</code> = 10; <code>learning_rate</code> = 0.1
	Dataset 2
CatBoost	Best ROC-AUC: 0.8100
	Best Parameters: <code>iterations</code> = 200; <code>depth</code> = 8; <code>learning_rate</code> = 0.1

Table 7 shows that the optimal hyperparameter configurations differ between Dataset 1 and Dataset 2,

indicating that boosting performance is strongly dataset-dependent. For XGBoost, Dataset 1 favored a lower learning rate and moderate depth, whereas Dataset 2 required deeper trees and a higher learning rate, indicating greater complexity needs. Similarly, LightGBM selected unrestricted depth for Dataset 1 but constrained depth for Dataset 2, suggesting different regularization requirements. For CatBoost, both datasets selected the same number of iterations, but Dataset 1 preferred deeper trees than Dataset 2, consistent with its stronger predictive performance.

Overall, these differences confirm that optimal hyperparameters do not generalize uniformly across datasets and must be tuned according to dataset characteristics. The performance of the tuned models for Dataset 1 and Dataset 2 is presented in Tables 8 and 9, respectively.

Table 8. After Hyperparameter Tuning for Dataset 1

Model	Accuracy	Precision	Recall	F1-Score	AUC
Split Ratio : 60:40					
XGBoost	0.9868	0.9946	0.9849	0.9997	0.9997
LGBM	0.9885	0.9982	0.9840	0.9911	0.9974
CatBoost	0.9845	0.9955	0.9805	0.9879	0.9980
Split Ratio : 70:30					
XGBoost	0.9878	0.9964	0.9846	0.9905	0.9990
LGBM	0.9893	0.9976	0.9858	0.9917	0.9986
CatBoost	0.9924	0.9976	0.9905	0.9941	0.9987
Split Ratio : 80:20					
XGBoost	0.9885	0.9964	0.9858	0.9911	0.9996
LGBM	0.9897	0.9964	0.9876	0.9920	0.9995
CatBoost	0.9908	0.9982	0.9876	0.9929	0.9997
Split Ratio : 90:10					
XGBoost	0.9931	1.000	0.9894	0.9947	0.9996
LGBM	0.9931	1.000	0.9894	0.9947	0.9996
CatBoost	0.9954	1.000	0.9929	0.9964	0.9995

Table 8 shows that hyperparameter tuning preserved the strong performance of Dataset 1, with accuracy consistently above 0.98 and AUC values close to 1.0 across all split ratios, indicating high stability. The best performance was obtained at the 90:10 split, where CatBoost achieved the highest accuracy, while XGBoost and LightGBM performed comparably.

Table 9. After Hyperparameter Tuning for Dataset 2

Model	Accuracy	Precision	Recall	F1-Score	AUC
Split Ratio : 60:40					
XGBoost	0.7111	0.7190	0.9819	0.8301	0.5002
LGBM	0.7118	0.7185	0.9848	0.8308	0.4989
CatBoost	0.7166	0.7189	0.9949	0.8346	0.4891
Split Ratio : 70:30					
XGBoost	0.7133	0.7190	0.9868	0.8319	0.5098
LGBM	0.7177	0.7187	0.9978	0.8356	0.5016
CatBoost	0.7177	0.7190	0.9968	0.8354	0.5080
Split Ratio : 80:20					
XGBoost	0.7139	0.7187	0.9890	0.8325	0.5031
LGBM	0.7148	0.7191	0.9900	0.8331	0.5073
CatBoost	0.7176	0.7190	0.9967	0.8354	0.5036
Split Ratio : 90:10					
XGBoost	0.7114	0.7192	0.9819	0.8303	0.4918
LGBM	0.7176	0.7185	0.9981	0.8355	0.4956
CatBoost	0.7190	0.7190	0.9997	0.8365	0.4929

Table 9 indicates that Dataset 2 maintained moderate accuracy (≈ 0.71 – 0.72) and AUC values near 0.50

across splits, with only marginal improvement after tuning. Although CatBoost and LightGBM slightly outperformed XGBoost, overall performance appears constrained by dataset characteristics rather than hyperparameter configuration.

Grid search with repeated stratified 5-fold cross-validation (3 repetitions) identified dataset-specific optimal hyperparameters, confirming that boosting configurations do not generalize uniformly across datasets. For Dataset 1, all tuned models achieved consistently high performance (accuracy > 0.98 ; AUC ≈ 1.0) across all split ratios, demonstrating strong discriminative ability and stability. In contrast, Dataset 2 showed limited improvement after tuning (accuracy ≈ 0.71 – 0.72 ; AUC ≈ 0.50), indicating that its predictive performance is primarily constrained by inherent dataset characteristics rather than hyperparameter configuration. Notably, for both datasets, the highest accuracy was observed at the 90:10 split ratio, where CatBoost emerged as the best-performing model among the evaluated boosting algorithms. This indicates that improving model configuration alone is insufficient when the underlying data lacks strong predictive signals. Table 10 provides a cross-dataset performance comparison of all models based on their best results after hyperparameter tuning.

Table 10. Cross-Dataset Performance Comparison (Best Results After Tuning)

Model	Dataset 1 (Best)	Dataset 2 (Best)
XGBoost	Acc: 0.9931,	Acc: 0.7139,
	AUC: 0.9996	AUC: 0.5031
LightGBM	Acc: 0.9931,	Acc: 0.7177,
	AUC: 0.9996	AUC: 0.5073
CatBoost	Acc: 0.9954,	Acc: 0.7190,
	AUC: 0.9995	AUC: 0.5080

To evaluate model robustness, a cross-dataset comparison was conducted using the best-performing hyperparameter settings. As shown in Table 10, all models achieved excellent performance on Dataset 1 (accuracy > 0.99 ; AUC ≈ 1.0). In contrast, performance dropped substantially on Dataset 2 (accuracy ≈ 0.71 – 0.72 ; AUC ≈ 0.5), indicating limited discriminative ability. This gap suggests that model performance is strongly influenced by dataset characteristics. Dataset 1 likely contains clearer feature-label relationships, whereas Dataset 2 may introduce noise or weaker patterns due to its simulated nature. Overall, these findings demonstrate that performance measured on a single dataset can overestimate model effectiveness, underscoring the importance of cross-dataset evaluation for evaluating generalizability.

3.9 Model Evaluation

Based on the results of hyperparameter tuning and split-ratio analysis, CatBoost with the 90:10 train–test configuration was selected as the final model for both datasets. The evaluation was conducted on the independent test sets to assess generalization performance.

Figure 4 presents the confusion matrix of CatBoost at the 90:10 split ratio for Dataset 1, while Figure 5 shows the corresponding confusion matrix for Dataset 2. The confusion matrices illustrate the classification outcomes, including true positives, true negatives, false positives, and false negatives, for the independent test sets.

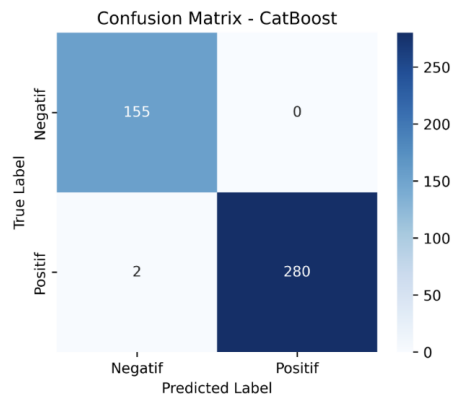


Figure 4. Confusion Matrix of Dataset 1

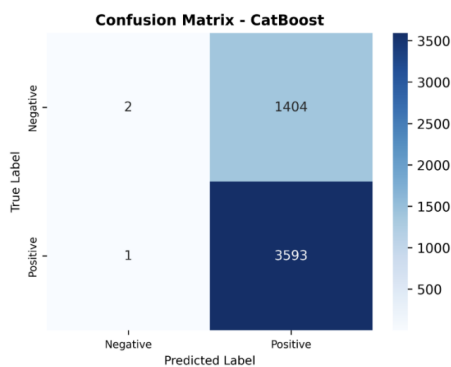


Figure 5. Confusion Matrix of Dataset 2

Figure 6 displays the ROC curve for Dataset 1, and Figure 7 presents the ROC curve for Dataset 2. The ROC analysis highlights the strong discriminative capability achieved in Dataset 1 and the comparatively limited separability observed in Dataset 2.

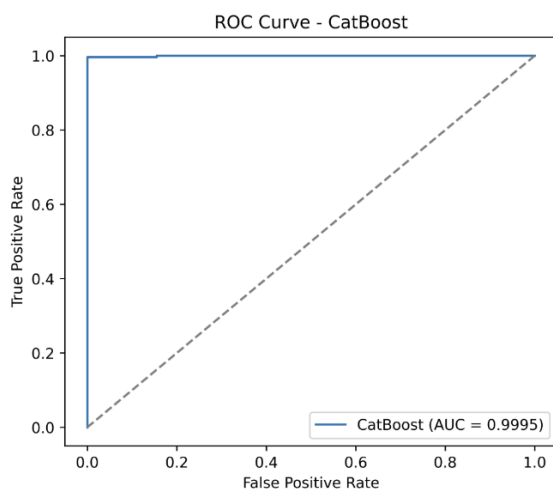


Figure 6. ROC Curve of Dataset 1

The ROC analysis confirms the strong separability achieved in Dataset 1, while demonstrating the comparatively limited discrimination capability observed in Dataset 2.

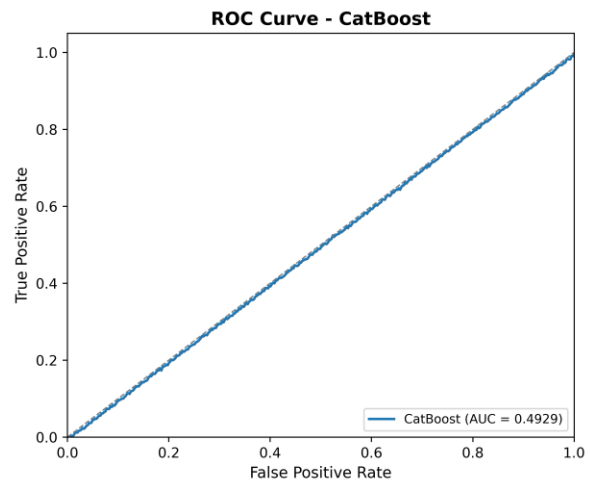


Figure 7. ROC Curve of Dataset 2

3.10 Model Interpretability

Model transparency and clinical interpretability were enhanced using SHAP (Shapley Additive exPlanations) for the best-performing model, CatBoost. SHAP summary (beeswarm) plots were generated to visualize the overall impact of each feature across all samples. These plots illustrate not only the magnitude of a feature's influence but also its direction, whether it increases or decreases hypertension risk. Figure 8 presents the SHAP summary plot, while Figure 9 shows the SHAP feature importance ranking.

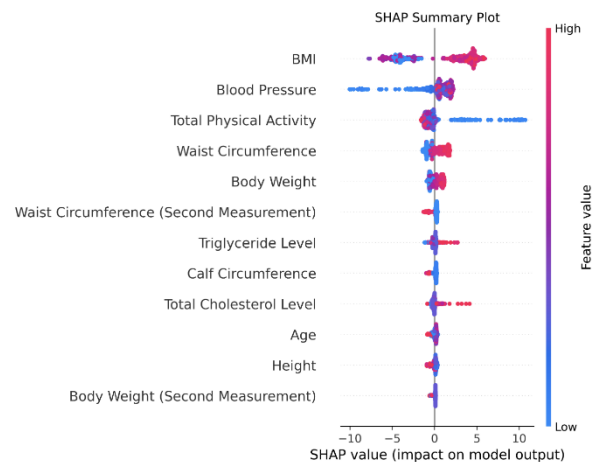


Figure 8. SHAP of Dataset 1

Figure 8 shows that BMI and Blood Pressure are dominant predictors, with wide SHAP distributions indicating strong and clearly separable feature effects. In contrast, Figure 9 identifies Stress Level and Smoking Status as important features, but with narrower and more centered SHAP ranges, reflecting weaker contributions.

The contrast between the two SHAP plots shows that Dataset 1 contains stronger, more clearly separable

predictive signals, whereas Dataset 2 exhibits more subtle, overlapping feature effects. This explains the substantial performance gap between the two datasets and supports the conclusion that dataset characteristics strongly influence boosting model effectiveness.

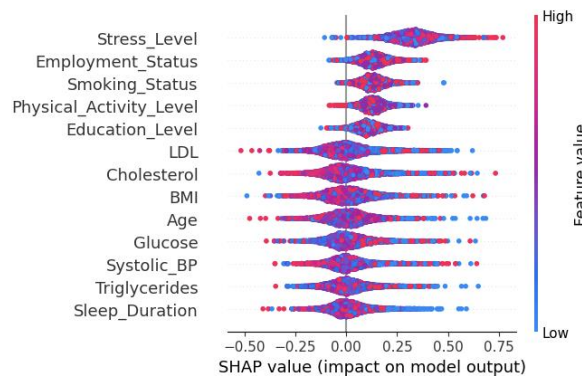


Figure 9. SHAP of Dataset 2

3.11 Comparative Analysis

To position the proposed model within the existing body of hypertension prediction research, Table 11 presents a comparative analysis of reported best-performing models from prior studies alongside the results of this study. The comparison focuses on classification accuracy to provide a clear benchmark of predictive performance.

Table 11. Comparative Analysis

Author	Model	Accuracy
Montagna et al. (2023) [9]	Random Forest	0.681
Islam et al. (2022) [10]	XGBoost, Gradient Boosting Machine	0.900
Kurniawan et al. (2023) [11]	Logistic Regression	0.898
Sakka et al. (2023) [27]	SMOTE + K-Nearest Neighbor	0.839
Fang et al. (2023) [5]	KNN + LightGBM	0.860
Islam et al. (2023) [28]	XGBoost	0.888
Tanaka et al. (2025) [4]	XGBoost	0.827
This Study	CatBoost	0.995

Table 11 shows that the CatBoost model attained the highest accuracy of 0.995, outperforming previous studies, which reported results ranging from 0.681 to 0.900. While prior research suggests that ensemble and boosting methods generally deliver superior performance compared to traditional models, none achieved performance at the level observed in this study.

However, the findings of Kurniawan et al. (2023), where Logistic Regression outperformed other models [11], differ from this study, where boosting algorithms achieved higher performance. This discrepancy may be due to differences in dataset characteristics and feature relationships, indicating that model performance is highly data-dependent.

The findings demonstrate that the proposed model achieves strong predictive performance while underscoring the need to consider dataset

characteristics when evaluating performance outcomes. In addition, the findings demonstrate the potential of using modifiable, non-invasive risk factors for predicting hypertension. Since these variables are easy to obtain and do not require clinical procedures, the proposed model is well-suited for early screening and preventive healthcare applications. This is particularly relevant for large-scale public health settings, where accessibility and cost-effectiveness are critical considerations.

3.12 Limitations and Future Work

This study has several limitations. First, the datasets were obtained from publicly available repositories. Dataset 1 is derived from the National Health and Nutrition Survey (ENSANUT) conducted by the Instituto Nacional de Salud Pública and accessed through the Kaggle repository, while Dataset 2 is a simulated dataset designed for machine learning experimentation. Although using two datasets enables cross-dataset evaluation, the simulated nature of Dataset 2 may not fully reflect real-world clinical variability. Second, the study focuses only on boosting-based algorithms (XGBoost, LightGBM, and CatBoost). Third, the use of a multinational dataset. Since the dataset is multinational, the results should be interpreted with caution for population-specific conclusions. However, the presence of Indonesian samples provides initial insight into the model's applicability in the Indonesian context. Future studies could further validate the proposed approach by incorporating a broader range of real-world clinical datasets and exploring other machine learning or deep learning approaches to further assess model generalizability and predictive performance.

4. Conclusions

This study systematically evaluated the performance and generalizability of XGBoost, LightGBM, and CatBoost for hypertension prediction across two independent datasets using multiple train-test split ratios and repeated stratified cross-validation. The results show that model effectiveness is highly dependent on dataset characteristics. For Dataset 1, all tuned models achieved consistently high performance (accuracy > 0.98; AUC $\approx$ 1.0), with CatBoost providing the best overall results, particularly at the 90:10 split ratio. In contrast, Dataset 2 demonstrated substantially lower and stable performance (accuracy $\approx$ 0.71–0.72; AUC $\approx$ 0.50) even after hyperparameter optimization, indicating limited feature separability. SHAP analysis further confirmed stronger, more distinct predictive signals in Dataset 1, whereas Dataset 2 exhibited more overlapping feature effects, which explains the performance gap.

These findings confirm that boosting configurations do not generalize uniformly across heterogeneous datasets and that predictive performance is primarily constrained by inherent data structure rather than model complexity alone. The main contribution of this study

is its cross-dataset evaluation framework, which provides empirical evidence on the robustness and transferability of boosting models in hypertension prediction. Practically, the results emphasize the need for dataset-specific validation prior to clinical deployment. Future research should focus on feature enrichment and broader external validation to improve model generalizability across diverse populations.

Acknowledgements

The authors would like to acknowledge the valuable funding provided by the Department of Research and Community Service, Universitas Amikom Yogyakarta.

References

- [1] S. Goorani, S. Zangene, and J. D. Imig, "Hypertension: A Continuing Public Healthcare Issue," *Int. J. Mol. Sci.*, vol. 26, no. 1, p. 123, Dec. 2024, doi: [10.3390/ijms26010123](https://doi.org/10.3390/ijms26010123).
- [2] H. Kohjitani, H. Koshimizu, K. Nakamura, and Y. Okuno, "Recent developments in machine learning modeling methods for hypertension treatment," *Hypertension Research*, vol. 47, no. 3, pp. 700–707, Mar. 2024, doi: [10.1038/s41440-023-01547-w](https://doi.org/10.1038/s41440-023-01547-w).
- [3] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments," *Energies (Basel)*, vol. 14, no. 16, p. 5196, Aug. 2021, doi: [10.3390/en14165196](https://doi.org/10.3390/en14165196).
- [4] M. Tanaka et al., "Machine learning-based analyses of contributing factors for the development of hypertension: a comparative study," *Clin. Exp. Hypertens.*, vol. 47, no. 1, 2025, doi: [10.1080/10641963.2025.2449613](https://doi.org/10.1080/10641963.2025.2449613).
- [5] M. Fang et al., "A hybrid machine learning approach for hypertension risk prediction," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14487–14497, Jul. 2023, doi: [10.1007/s00521-021-06060-0](https://doi.org/10.1007/s00521-021-06060-0).
- [6] V. E. Gutiérrez-Carmona, M. A. Luna-Montes, A. Guzmán-Pando, J. Camarillo-Cisneros, and N. G. Sámano-Lira, "Hypertension Risk Prediction in Mexican Population Using Machine Learning Models," 2026, pp. 79–88. doi: [10.1007/978-3-032-13729-6_9](https://doi.org/10.1007/978-3-032-13729-6_9).
- [7] E. Septian et al., "Prediction of Personalised Hypertension Using Machine Learning in Indonesian Population," *J. Med. Syst.*, vol. 49, no. 1, Dec. 2025, doi: [10.1007/s10916-025-02253-5](https://doi.org/10.1007/s10916-025-02253-5).
- [8] Y. Wang et al., "Prediction Model of Intradialytic Hypertension in Hemodialysis Patients Based on Machine Learning," *J. Med. Syst.*, vol. 49, no. 1, Dec. 2025, doi: [10.1007/s10916-025-02237-5](https://doi.org/10.1007/s10916-025-02237-5).
- [9] S. Montagna et al., "Machine Learning in Hypertension Detection: A Study on World Hypertension Day Data," *J. Med. Syst.*, vol. 47, no. 1, Dec. 2023, doi: [10.1007/s10916-022-01900-5](https://doi.org/10.1007/s10916-022-01900-5).
- [10] S. M. S. Islam et al., "Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries," *Front. Cardiovasc. Med.*, vol. 9, Mar. 2022, doi: [10.3389/fcvm.2022.839379](https://doi.org/10.3389/fcvm.2022.839379).
- [11] R. Kurniawan et al., "Hypertension prediction using machine learning algorithm among Indonesian adults," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 776–784, Jun. 2023, doi: [10.11591/ijai.v12.i2.pp776-784](https://doi.org/10.11591/ijai.v12.i2.pp776-784).
- [12] R. Oulhiq, K. Benjelloun, Y. Kali, and M. Saad, "A data mining based approach for process identification using historical data," *International Journal of Modelling and Simulation*, vol. 42, no. 2, pp. 335–349, Mar. 2022, doi: [10.1080/02286203.2021.1905375](https://doi.org/10.1080/02286203.2021.1905375).
- [13] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset," *Electronics (Basel)*, vol. 12, no. 21, p. 4423, Oct. 2023, doi: [10.3390/electronics12214423](https://doi.org/10.3390/electronics12214423).
- [14] T. Yousefi and Ö. Varliklar, "New hybrid multi-objective feature selection: Boruta-XGBoost," *PeerJ Comput. Sci.*, vol. 12, p. e3463, Jan. 2026, doi: [10.7717/peerj-cs.3463](https://doi.org/10.7717/peerj-cs.3463).
- [15] Y. Pristyanto, A. F. Nugraha, B. Wulan Sari, M. Sulistiyono, and A. Sunyoto, "Enhancement of Machine Learning Models Using Intersection Filtering Model Based on Recursive Feature Elimination on Specified Android Malware Classification," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 3, pp. 836–852, Apr. 2025, doi: [10.22266/ijies2025.0430.56](https://doi.org/10.22266/ijies2025.0430.56).
- [16] B. Wulan Sari and D. Prabowo, "Optimization of Thyroid Disease Prediction Using Feature Selection with Dynamic Threshold," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 8, pp. 168–183, Sep. 2025, doi: [10.22266/ijies2025.0930.12](https://doi.org/10.22266/ijies2025.0930.12).
- [17] S. Abdumalikov, J. Kim, and Y. Yoon, "Performance Analysis and Improvement of Machine Learning with Various Feature Selection Methods for EEG-Based Emotion Classification," *Applied Sciences*, vol. 14, no. 22, p. 10511, Nov. 2024, doi: [10.3390/app142210511](https://doi.org/10.3390/app142210511).
- [18] A. A. Engda, A. O. Salau, and O. Ajala, "Classical Machine Learning Approaches for Early Hypertension Risk Prediction: A Systematic Review," Oct. 01, 2025, *John Wiley and Sons Inc.* doi: [10.1002/aii2.70005](https://doi.org/10.1002/aii2.70005).
- [19] N. L. Fitriyani et al., "A Novel Approach Utilizing Bagging, Histogram Gradient Boosting, and Advanced Feature Selection for Predicting the Onset of Cardiovascular Diseases," *Mathematics*, vol. 13, no. 13, p. 2194, Jul. 2025, doi: [10.3390/math13132194](https://doi.org/10.3390/math13132194).
- [20] K. Thomas, A. Khandakar, P. Chelvanathan, B. Aissa, and M. I. Hossain, "Enhanced prediction and optimization of thin metal film optical properties using optimized ensemble learning models," *Sci. Rep.*, vol. 15, no. 1, p. 43523, Dec. 2025, doi: [10.1038/s41598-025-27524-6](https://doi.org/10.1038/s41598-025-27524-6).
- [21] D. Prabowo, Bety Wulan Sari, Yoga Pristyanto, and Afrig Aminuddin, "Enhanced Predictive Modeling for Non-Invasive Liver Disease Diagnosis," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 4, pp. 745–753, Aug. 2025, doi: [10.29207/resti.v9i4.6449](https://doi.org/10.29207/resti.v9i4.6449).
- [22] M.-L. Leitner and M. Arendasy, "Comparing machine learning and artificial neural network models in psychological research: a ROC-based analysis," *Front. Psychol.*, vol. 17, Feb. 2026, doi: [10.3389/fpsyg.2026.1746479](https://doi.org/10.3389/fpsyg.2026.1746479).
- [23] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification," *PLoS One*, vol. 19, no. 12, p. e0316019, Dec. 2024, doi: [10.1371/journal.pone.0316019](https://doi.org/10.1371/journal.pone.0316019).
- [24] H. Han, "The Utility of Receiver Operating Characteristic Curve in Educational Assessment: Performance Prediction," *Mathematics*, vol. 10, no. 9, p. 1493, Apr. 2022, doi: [10.3390/math10091493](https://doi.org/10.3390/math10091493).
- [25] K. Merabet et al., "Predicting water quality variables using gradient boosting machine: global versus local explainability using SHapley Additive Explanations (SHAP)," *Earth Sci. Inform.*, vol. 18, no. 3, p. 298, Sep. 2025, doi: [10.1007/s12145-025-01796-y](https://doi.org/10.1007/s12145-025-01796-y).
- [26] B. Oral, H. A. Maddah, and R. Yildirim, "Predictive Modeling and SHAP (SHapley Additive Explanations) Analysis for Enhancing Natural Dye-Sensitized Solar Cell Performance," *Solar RRL*, vol. 8, no. 22, Nov. 2024, doi: [10.1002/solr.202400432](https://doi.org/10.1002/solr.202400432).
- [27] Y. Sakka, D. Qarashai, and A. Altarawneh, "Predicting Hypertension using Machine Learning: A Case Study at Petra University," 2023, doi: [10.14569/IJACSA.2023.0140368](https://doi.org/10.14569/IJACSA.2023.0140368).
- [28] M. M. Islam et al., "Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia," *PLoS One*, vol. 18, no. 8 August, Aug. 2023, doi: [10.1371/journal.pone.0289613](https://doi.org/10.1371/journal.pone.0289613).