



Analisis Komparatif Algoritma Pembelajaran Mesin untuk Prediksi Risiko Hipertensi: Gradient Boosting, Random Forest, XGBoost, Regresi Logistik, dan Support Vector Machine

Comparative Analysis of Machine Learning Algorithms for Hypertension Risk Prediction: Gradient Boosting, Random Forest, XGBoost, Logistic Regression, and Support Vector Machine

Theo Vhaldino¹, Arpa Pauziah², Khoirin³, Rudiansyah⁴, Debbi Susita Yuzzaki⁵, Ahmad Syarif⁶

^{1,2,6}Teknologi Informasi, Fakultas Kesehatan dan Teknologi, Universitas Aisyiyah Palembang

³Ilmu Keperawatan, Fakultas Kesehatan dan Teknologi, Universitas Aisyiyah Palembang

^{4,5}Manajemen Informasi Kesehatan, Fakultas Kesehatan dan Teknologi, Universitas Aisyiyah Palembang

*Korespondensi Email: theovhaldino@unisa-palembang.ac.id

Abstract

Hypertension is a major contributor to cardiovascular morbidity, while community screening may be constrained by irregular health checks and limited access to early risk assessment. This study aims to compare the performance of Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and Support Vector Machine for hypertension risk prediction and to analyze the contribution of dominant predictors to model performance. The study uses the Hypertension Risk Prediction Dataset obtained from the Kaggle platform, comprising 1,985 patient records with ten demographic, clinical, and lifestyle predictors. Numerical features were standardized, categorical features were transformed using one-hot encoding, and missing medication values were imputed using the mode. Evaluation was conducted using an 80:20 stratified train-test split and 5-fold stratified cross-validation. The best-performing model was optimized using RandomizedSearchCV and subsequently analyzed through feature importance and ablation analysis. Tuned Gradient Boosting achieved a cross-validation F1-score of 0.9994 and ROC-AUC of 1.0000. On 397 test records, the model achieved an accuracy of 0.9975, precision of 1.0000, recall of 0.9951, F1-score of 0.9976, and ROC-AUC of 1.0000. Feature importance identified blood pressure history as the dominant predictor, while its removal reduced the F1-score to 0.7681. The contribution of this study lies in comparing five machine learning algorithms and analyzing the dependence of model performance on dominant predictors through ablation analysis. The model can support early hypertension screening, but external clinical validation and calibration are required before clinical or diagnostic use.

Keywords: Hypertension; Machine Learning; Gradient Boosting; XGBoost; Risk Prediction

Abstrak

Hipertensi merupakan salah satu kontributor utama morbiditas kardiovaskular, sementara skrining komunitas masih dapat terkendala oleh pemeriksaan kesehatan yang tidak rutin dan keterbatasan akses terhadap penilaian risiko dini. Penelitian ini bertujuan membandingkan kinerja Logistic Regression, Random Forest, Gradient Boosting, XGBoost, dan Support Vector Machine untuk prediksi risiko hipertensi serta menganalisis kontribusi prediktor dominan terhadap kinerja model. Penelitian menggunakan Hypertension Risk Prediction Dataset yang diperoleh dari platform Kaggle, terdiri atas 1.985 data pasien dengan sepuluh prediktor demografis, klinis, dan gaya hidup. Fitur numerik distandarisasi, fitur kategorikal ditransformasikan menggunakan one-hot encoding, dan nilai medication yang hilang diimputasi menggunakan modus. Evaluasi dilakukan menggunakan pembagian data latih-uji stratified 80:20 dan stratified cross-validation 5-fold. Model terbaik dioptimalkan menggunakan RandomizedSearchCV, kemudian dianalisis melalui feature importance dan ablation. Tuned Gradient Boosting memperoleh F1-score cross-validation sebesar 0,9994 dan ROC-AUC sebesar 1,0000. Pada 397 data uji, model memperoleh akurasi 0,9975, precision 1,0000, recall 0,9951, F1-score 0,9976, dan ROC-AUC 1,0000. Feature importance menunjukkan blood pressure history sebagai prediktor dominan, sedangkan penghapusannya menurunkan F1-score menjadi 0,7681. Kontribusi penelitian terletak pada perbandingan lima algoritma dan analisis ketergantungan kinerja terhadap prediktor dominan melalui ablation. Model dapat mendukung skrining awal hipertensi, tetapi memerlukan validasi dan kalibrasi klinis eksternal sebelum penggunaan klinis atau diagnostik.

Kata kunci : Pembelajaran Mesin; Gradient Boosting; XGBoost; Prediksi Risiko.

1. Introduction

Hypertension is a major noncommunicable disease because its clinical consequences accumulate gradually while the condition itself is frequently unnoticed by patients. Sustained elevation of blood pressure increases the probability of stroke, coronary heart disease, heart failure, kidney disease, and premature cardiovascular mortality. The World Health Organization describes hypertension as a global public health priority and emphasizes that many affected individuals remain undiagnosed or inadequately controlled [1]. Large population analyses also show that hypertension prevalence and treatment coverage remain uneven across countries, with persistent gaps in low- and middle-income settings [2]. This burden makes early identification of high-risk individuals a public health necessity rather than merely a clinical routine.

Early identification of high-risk individuals is especially relevant in Indonesia. The national Basic Health Research report documented a high prevalence of hypertension among adults, suggesting that more than one third of the adult population may be affected [3]. This epidemiological condition is clinically important because global evidence consistently links elevated blood pressure with cardiovascular events and preventable mortality [4], [5]. However, community-level screening still faces practical constraints, including irregular health checks, limited awareness of asymptomatic risk, uneven access to trained personnel, and the absence of simple digital instruments that can support early referral. The Indonesian screening problem is therefore not only epidemiological; it is also an information-systems problem.

An information-systems problem requires an analytical approach that can transform routinely collected patient attributes into actionable risk information. Conventional blood pressure measurement remains indispensable for diagnosis, but a predictive screening layer can help identify individuals who should be prioritized for measurement, counseling, or medical consultation. In this role, the model does not replace clinical judgment; instead, it functions as a triage-oriented decision support component that summarizes risk based on accessible patient variables. Such a component is valuable only if it is accurate, transparent enough to be interpreted, and simple

enough to be deployed in community-facing applications.

A model that is accurate, interpretable, and deployable is difficult to obtain with a single algorithmic assumption. Linear models are useful baselines because they offer a simple decision surface, yet they may fail to capture interactions among age, body mass index, lifestyle behavior, family history, medication status, and prior blood pressure condition. Kernel-based and ensemble models can capture more complex decision boundaries, but their performance depends on data preprocessing, feature representation, and validation design. For that reason, a comparative experimental design is more defensible than selecting one algorithm a priori.

A comparative experimental design is also consistent with recent developments in medical machine learning. Reviews of machine-learning-based diagnosis have shown that supervised learning models are increasingly being applied to support disease risk prediction and clinical decision-making[6]. In the context of hypertension, previous studies have investigated various machine learning approaches using different predictor compositions, population-based datasets, feature selection techniques, and risk prediction systems. These studies have demonstrated the potential of machine learning for hypertension risk prediction while also reporting varying levels of predictive performance across different algorithms and validation settings. Sheikh Mohammed Shariful Islam et al.[7] used six machine learning algorithms to predict hypertension in population data from Bangladesh, Nepal, and India. XGBoost, GBM, LR, and LDA achieved an accuracy of 90%, a recall of 100%, and an F1-score of 95% however, this study still relied on conventional machine learning and did not explore the use of Transformers for tabular data classification. Huanhuan Zhao et al.[8] compared RF, CatBoost, MLP, and LR for predicting the risk of hypertension. RF achieved an AUC of 92%, an accuracy of 82%, a sensitivity of 83%, and a specificity of 81%; however, its prediction results still showed relatively low accuracy and specificity values. Yan Peng et al.[9] used hybrid feature selection and XGBoost to predict hypertension and achieved an accuracy of 77.2% and an AUC of 84.6%; however, these accuracy and AUC results are still relatively low. Jinsong Du et al.[10] compared 10 machine learning algorithms for predicting the risk of hypertension and identified LightGBM as the best-

performing model; however, the study's results did not provide specific predictive performance metrics, such as accuracy, F1-score, or AUC. Seung Ha Hwang et al.[11] used an ensemble of Adaptive Boosting and Logistic Regression to predict hypertension, achieving a balanced accuracy of 81.2%, a sensitivity of 80.6%, a specificity of 81.8%, and an AUC of 90.1%; however, the balanced accuracy and sensitivity values remain relatively low. Collectively, these studies demonstrate that predictive performance varies according to algorithm, predictor composition, population characteristics, and validation setting. However, the existing studies do not directly address how the contribution of clinically proximal predictors, particularly previous blood pressure history, affects the observed performance of different machine learning algorithms when evaluated within a common experimental framework.

This limitation creates an important research gap in hypertension risk prediction. A high-performing model may derive its predictive capability primarily from variables that are closely related to the target outcome rather than from a broad combination of demographic and lifestyle characteristics. Previous studies have generally focused on comparing algorithms or identifying predictive factors, but do not necessarily combine algorithmic comparison with a controlled ablation analysis to quantify the contribution of such clinically proximal predictors. Consequently, it remains unclear whether the predictive performance of a hypertension screening model is maintained when previous blood pressure history is excluded and prediction relies more heavily on lifestyle and demographic information. Addressing this gap is important for interpreting model performance and determining the appropriate scope of a screening-oriented application.

The relationship between predictor composition and performance is therefore central to hypertension modeling. Some variables, such as age, smoking status, body mass index, sleep duration, salt intake, and stress score, represent demographic, behavioral, or lifestyle characteristics that may support prevention and risk stratification. Other variables, particularly previous blood pressure history, are clinically proximal to the target outcome and may substantially increase predictive performance. When such a proximal feature is included, high model performance may reflect useful

discriminative information, but it may also obscure the extent to which lifestyle and demographic variables independently contribute to prediction. This issue motivates the use of ablation analysis in the present study.

In this research, ablation analysis is used to quantify the change in model performance after selected predictors are removed. By comparing the full feature set with reduced feature scenarios, the study examines whether the observed predictive performance is broadly supported by demographic and lifestyle characteristics or is predominantly driven by previous blood pressure history. This analysis provides an additional basis for interpreting model performance and defining the appropriate use of the resulting screening model.

To address this gap, this study conducts a comparative analysis of Gradient Boosting, Random Forest, XGBoost, Logistic Regression, and Support Vector Machine for hypertension risk prediction. The study evaluates model performance using stratified cross-validation and independent test data, tunes the strongest candidate, examines feature importance, performs controlled ablation analysis, and translates the selected pipeline into a web-based screening prototype. The contribution of this study is not limited to comparing the predictive performance of multiple algorithms, but also includes quantifying the contribution of dominant predictors and examining how model performance changes when clinically proximal information is excluded. This approach provides a more transparent basis for interpreting high predictive performance and for defining the use of machine learning as an early hypertension screening support tool.

2. Methodology

Research Design

The study used a quantitative experimental design. The full program was implemented in Python using pandas and scikit-learn for preprocessing, modeling, and evaluation [12], and XGBoost for scalable gradient-boosted tree modeling [13]. The experimental process consisted of data acquisition, data preprocessing, comparative model training, cross-validation, hyperparameter tuning, test evaluation, feature importance analysis, model serialization, and implementation in a Streamlit web interface. The research flow is shown in Figure 1.



Figure 1. Research flow of the hypertension risk prediction study

The design follows the logic of applied prediction-model research. First, the dataset was inspected to identify feature types, missing values, and target class distribution. Second, a preprocessing pipeline was constructed so that every model was trained on the same representation of numerical and categorical variables. Third, five algorithms with different learning assumptions were evaluated under an identical train-test split and cross-validation scheme. Fourth, the best-performing baseline model was tuned to obtain a stronger final classifier. Fifth, the final model was examined using feature importance and ablation analysis to determine whether performance depended on a small set of dominant predictors. This sequence is important because a prediction model should be assessed not only by its score, but also by the credibility of the modeling process that produced the score.

The research objective was not to establish a clinical diagnostic rule, but to evaluate whether machine learning can provide a reliable early screening signal from simple health attributes. Therefore, the methodological emphasis was placed on reproducible preprocessing, fair algorithm comparison, and careful interpretation of very high performance values. This orientation is aligned with reporting guidance for prediction models, which emphasizes clear descriptions of data, predictors, outcomes, modeling strategy, and intended use [14], [15].

Dataset

The dataset used in this study was the Hypertension Risk Prediction Dataset, obtained from the Kaggle platform and publicly available through the Kaggle dataset repository. The dataset contains 1,985 patient records with ten predictors and one binary target variable, Has_Hypertension. The predictors consist of five numerical variables and five categorical variables. The numerical variables are Age, Salt_Intake, Stress_Score, Sleep_Duration, and BMI, while the categorical variables are BP_History, Medication,

Family_History, Exercise_Level, and Smoking_Status. The target variable was encoded as 1 for hypertension risk and 0 for no risk.

Table 1. Dataset summary

Aspect	Description
Number of records	1,985 patient records
Number of predictors	10 predictors and one binary target variable
Target distribution	Yes: 1,032 records (51.99%); No: 953 records (48.01%)
Train-test split	1,588 training records and 397 testing records using stratification
Missing values	Medication contained 799 missing values before preprocessing
Imputation rule	Missing Medication values were replaced with the mode, Beta Blocker

The target distribution is relatively balanced, with a difference of 79 records between the two classes. This balance reduces the risk that the model simply follows the majority class. At the same time, the inclusion of BP_History makes the dataset clinically strong but also requires cautious interpretation because previous blood pressure condition is a proximal predictor of the target outcome.

The numerical variables show a wide enough range to represent heterogeneous patient profiles. Age ranges from 18 to 84 years, BMI ranges from 11.9 to 41.9, salt intake ranges from 2.5 to 16.4 grams per day, stress score ranges from 0 to 10, and sleep duration ranges from 1.5 to 11.4 hours per day. These ranges make standardization necessary for algorithms that are sensitive to scale, especially SVM and Logistic Regression. They also make the dataset suitable for tree-based models, which can partition nonlinear thresholds without assuming a linear relationship between predictors and the target.

Table 2. Predictor variables used in the model

Feature	Type	Description
Age	Numeric	Patient age in years
Salt_Intake	Numeric	Estimated daily salt consumption in grams
Stress_Score	Numeric	Stress score on a 0-10 scale
Sleep_Duration	Numeric	Average sleep duration in hours per day
BMI	Numeric	Body mass index
BP_History	Categorical	Previous blood pressure category
Medication	Categorical	Medication status or type

Family_History	Categorical	Family history of hypertension
Exercise_Level	Categorical	Physical activity level
Smoking_Status	Categorical	Current smoking status

The categorical predictors represent clinically interpretable groups. BP_History has three values: Normal, Prehypertension, and Hypertension. Family_History has two values: Yes and No. Exercise_Level has three values: Low, Moderate, and High. Smoking_Status has two values in the dataset: Smoker and Non-Smoker. Medication contains four observed medication categories plus missing records. These attributes were retained because they represent variables that can realistically be collected through a health questionnaire or screening form.

The study used a quantitative experimental design. The full program was implemented in Python using pandas and scikit-learn for preprocessing

Data Preprocessing

Data preprocessing was performed before model training to ensure that all algorithms received a consistent input representation. Missing values were identified only in the Medication field, with 799 missing records. Because Medication is a categorical variable, missing values were handled using mode imputation, in which each missing value was replaced by the most frequently observed medication category in the training data. Mode imputation was selected because it is appropriate for nominal categorical variables, preserves the categorical structure of the feature, and does not introduce artificial numerical relationships between medication categories. The mode was calculated only from the training data to prevent information from the validation or test data from influencing the imputation process. Numerical variables were standardized using StandardScaler, while categorical variables were converted into dummy variables using OneHotEncoder with the first category dropped to avoid redundant columns. Unknown categories were ignored during transformation, which makes the saved pipeline more robust when receiving new user input through the web interface.

The preprocessing process was embedded in a ColumnTransformer so that transformation rules were learned only from the training data during validation. This detail is important because applying scaling or encoding rules to the complete

dataset before splitting can leak information from the test subset into model training. The implemented pipeline prevents this leakage by fitting the imputer, scaler, and encoder within each training fold during cross-validation and then applying the learned transformations to the corresponding validation fold. The same principle was used for the final train-test evaluation.

The data were split into training and testing subsets using an 80:20 stratified split with a random state of 42. The stratification step preserved the target class proportion in both subsets. Model selection was conducted only on the training subset using 5-fold stratified cross-validation, and the final selected model was evaluated on the independent test subset.

After preprocessing, the five numerical variables remained as standardized continuous predictors, whereas the categorical variables were transformed into binary indicator columns. The resulting feature matrix was therefore appropriate for all selected algorithms. This unified representation ensured that differences in model performance were caused primarily by the learning algorithms and not by inconsistent preprocessing choices.

Classification Algorithms

Five classification algorithms were compared. Logistic Regression was included as a linear baseline because it is widely used in medical risk modeling and provides an interpretable comparison point for nonlinear algorithms [16]. Random Forest was selected because it combines multiple decision trees through bagging and is known for stable performance on heterogeneous tabular data [17]. Gradient Boosting was selected because it iteratively improves weak learners by minimizing residual errors, often producing strong results for structured classification problems [18]. XGBoost was included as an efficient implementation of gradient-boosted trees [13]. Support Vector Machine was included because it can construct separating hyperplanes in high-dimensional transformed feature spaces [19].

The inclusion of these algorithms was deliberate. Logistic Regression represents a classical statistical learning baseline and allows the study to determine whether nonlinear modeling is necessary. SVM represents margin-based learning and is useful for evaluating whether separation in a transformed feature space can improve classification. Random Forest evaluates a

bagging-based ensemble strategy, whereas Gradient Boosting and XGBoost evaluate sequential boosting strategies. Together, these models represent linear, margin-based, bagging, and boosting families, which makes the comparison broader than a simple contest among similar tree models.

After the initial comparison, Gradient Boosting was tuned using RandomizedSearchCV with 40 candidate parameter settings and F1-score as the optimization metric. F1-score was prioritized because screening models must balance precision and recall; missing a high-risk case can be more consequential than recommending additional follow-up for a low-risk user.

Table 3. Hyperparameters selected for the tuned Gradient Boosting model

Parameter	Selected value
n_estimators	236
learning_rate	0.1802
max_depth	2
min_samples_s	10
split	
min_samples_l	13
leaf	
subsample	0.9209

The hyperparameter search space covered estimator count, learning rate, tree depth, minimum split size, minimum leaf size, and subsampling rate. These parameters control the bias-variance behavior of Gradient Boosting. A deeper tree can capture more complex interactions but may overfit; a lower learning rate can improve generalization but requires more estimators; and subsampling introduces stochasticity that can reduce variance. The selected configuration uses shallow trees with a moderate learning rate and subsampling, indicating that the model learned compact interactions rather than unrestricted tree structures.

Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy measures the overall proportion of correct classifications, whereas precision measures the proportion of predicted risk cases that are truly risk cases. Recall measures the ability to capture actual risk cases, and F1-score combines precision and recall into a single harmonic mean. ROC-AUC summarizes the discrimination ability of the model over all possible decision thresholds and is widely used in binary classification evaluation [20].

For hypertension screening, recall and F1-score are particularly important. A false negative means that a patient with hypertension risk is incorrectly classified as no risk, which may delay further measurement or consultation. A false positive is also undesirable because it can create unnecessary concern, but it is generally less harmful in a screening context than missing a high-risk individual. For this reason, F1-score was used as the tuning metric because it balances precision and recall, while ROC-AUC was used to evaluate threshold-independent discrimination.

The confusion matrix was also inspected because aggregate metrics alone can hide clinically relevant error patterns. Specifically, the analysis considered true negatives, false positives, false negatives, and true positives. This inspection allows the study to explain not only how high the model score is, but what type of classification error remains in the final model.

Ablation Study Design

An ablation study was added to test the dependence of the best model on specific predictor groups. The full model used all ten predictors. Four reduced scenarios were then evaluated using the same tuned Gradient Boosting configuration: without BP_History, lifestyle-demographic variables only, without the top three predictors, and without behavioral predictors. The purpose was to determine whether the near-perfect result of the full model was driven by one clinically proximal variable or by a more distributed risk pattern.

The ablation scenarios were evaluated using the same 80:20 stratified split, the same 5-fold cross-validation approach, and the same test metrics as the full model. This controlled design makes the comparison interpretable because the only intentional difference between scenarios is feature availability. In publication terms, the ablation study strengthens internal validity by showing how the conclusion changes when dominant predictors are removed.

Model Serialization and Web Prototype

After the best model was selected, the complete preprocessing-classification pipeline was serialized into a pickle file. This design choice ensures that the deployed model receives raw user input in the same schema as the training data and applies identical preprocessing before prediction.

The web prototype was built using Streamlit and contains input controls for age, BMI, salt intake, stress score, sleep duration, blood pressure history, medication, family history, exercise level, and smoking status.

The application returns the predicted class, risk probability, and screening-oriented recommendation. It also contains a medical disclaimer indicating that the output is for early screening only and is not a substitute for professional diagnosis. This distinction is essential because prediction models for health applications must be integrated into a clear clinical pathway rather than presented as standalone diagnostic authorities.

3. Results and Discussion

Results

Comparative Model Performance

Table 4 presents the cross-validation results of the five algorithms and the tuned Gradient Boosting model. Before tuning, Gradient Boosting already produced the strongest baseline result with an F1-score of 0.9856 and ROC-AUC of 0.9990. XGBoost followed closely with an F1-score of 0.9811 and ROC-AUC of 0.9985. Random Forest also performed strongly, while SVM and Logistic Regression showed lower F1-score values. This pattern indicates that tree-based ensemble models are more suitable for capturing the nonlinear interactions in the hypertension risk dataset.

The ranking in Table 4 shows a clear separation between boosted tree models and the non-boosting baselines. Logistic Regression produced the lowest F1-score, which indicates that a single linear decision boundary was not sufficient to represent the interaction among the available predictors. SVM improved upon Logistic Regression but still remained below the ensemble methods. Random Forest performed strongly, confirming the value of tree aggregation, but its F1-score remained below Gradient Boosting and XGBoost. These results suggest that sequential error correction, which is the core principle of boosting, is particularly suitable for this dataset.

Table 4. Cross-validation performance comparison

Model	F1-score	Std	ROC-AUC	Std
Gradient Boosting (Tuned)	0.9994	0.0000	1.0000	0.0000
Gradient Boosting	0.9856	0.0100	0.9990	0.0012
XGBoost	0.9811	0.0024	0.9985	0.0004
Random Forest	0.9482	0.0070	0.9894	0.0027
SVM	0.9060	0.0096	0.9696	0.0040
Logistic Regression	0.8825	0.0133	0.9528	0.0107

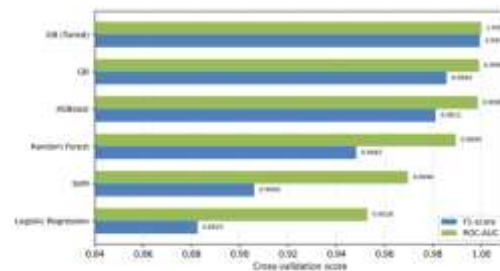


Figure 2. Cross-validation F1-score and ROC-AUC comparison across algorithms

The improvement obtained by tuning Gradient Boosting is substantial. The tuned model increased the cross-validation F1-score from 0.9856 to 0.9994, suggesting that the selected combination of tree depth, learning rate, estimator count, and subsampling helped reduce residual classification errors. The very small standard deviation also indicates stable cross-validation performance across folds. However, this near-perfect score should not be interpreted as proof of clinical readiness by itself; it instead indicates that the available dataset contains highly discriminative patterns.

The small standard deviation should also be interpreted carefully. In one sense, it is positive because it shows that the model did not depend on one favorable fold. In another sense, it raises a methodological question: if all folds are almost perfectly separable, then at least one predictor or predictor combination may be very close to the target concept. This observation motivates the later ablation analysis. Without ablation, the study would risk presenting a high score without explaining why the score is high.

Best Model Evaluation on Test Data

The tuned Gradient Boosting model was retrained on the full training subset and then evaluated on 397 independent test records. The model correctly classified 396 records and misclassified only one risk case as no risk. It produced no false-positive prediction on the test subset. The resulting confusion matrix is shown in Figure 3, and the test metrics are summarized in Table 5.

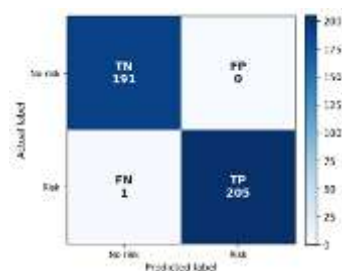


Figure 3. Confusion matrix of the tuned Gradient Boosting model on test data

Table 5. Test performance of the tuned Gradient Boosting model

Metric	Value
Accuracy	0.9975
Precision	1.0000
Recall	0.9951
F1-score	0.9976
ROC-AUC	1.0000

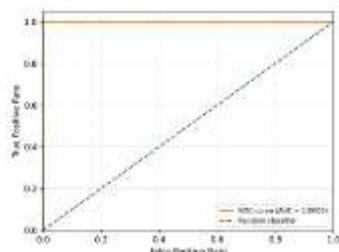


Figure 4. ROC curve of the tuned Gradient Boosting model

From a screening perspective, the most important observation is the high recall value of 0.9951. This means that almost all actual risk cases in the test subset were captured. The precision value of 1.0000 indicates that every positive prediction in this test subset corresponded to an actual risk case. The ROC-AUC of 1.0000 reflects perfect class separation under the tested threshold range. These results are stronger than many previous machine learning studies on hypertension risk prediction [7], [8], [9], [10], [11], but the difference should be interpreted in relation to dataset structure, input variables, and validation design.

The confusion matrix provides a more concrete interpretation of these values. The model correctly identified 191 no-risk records and 205 risk records. It generated zero false positives and one false negative. In public health screening, the single false negative is the most clinically relevant error because it represents a risk case that would not be flagged by the system. However, the number is very small relative to the 206 actual risk cases in the test subset, which explains the high recall

value. The absence of false positives explains the precision value of 1.0000.

A critical explanation is that BP_History is part of the predictor set. This feature provides information very close to the target condition, so the model can distinguish risk classes more easily than if it relied only on lifestyle factors such as salt intake, sleep duration, smoking, BMI, or stress. For this reason, the model is appropriate as an experimental screening support prototype, but it should be externally validated on local primary-care or community data before being claimed as a deployable clinical decision-support model.

The test result is therefore best understood as evidence of internal predictive validity under the available dataset, not as evidence of external clinical generalization. For SINTA-indexed publication, this distinction matters because reviewers may question overly perfect results in health prediction studies. The appropriate scientific response is not to weaken the result, but to interpret it transparently, test its dependence on dominant predictors, and clearly state the requirement for external validation.

Dominant Predictive Factors

The feature importance analysis of the tuned Gradient Boosting model showed that blood pressure history was the strongest discriminator, followed by stress score, age, family history, smoking status, BMI, sleep duration, and salt intake. Because categorical features were one-hot encoded with one category dropped, the importance values for BP_History_Normal and BP_History_Prehypertension should be understood as discriminative contributions relative to the omitted BP_History category, not as direct causal effects.

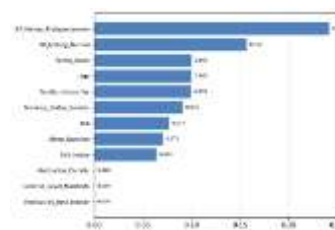


Figure 5. Dominant features in the tuned Gradient Boosting model

Table 6. Top feature importance values

Feature	Importance	Interpretation
---------	------------	----------------

BP_History_Prehypertension	0.2415	Strong discriminator of previous blood pressure status
BP_History_Normal	0.1566	Strong discriminator relative to the omitted category
Stress_Score	0.1001	Psychosocial stress contributes to risk separation
Age	0.0999	Age captures demographic risk gradient
Family_History_Years	0.0994	Family history contributes to inherited or shared risk
Smoking_Status_Smoker	0.0909	Smoking status supports lifestyle-related discrimination
BMI	0.0707	Body composition contributes to hypertension risk
Sleep_Duration		Sleep duration adds behavioral information
Salt_Intake	0.0638	Salt intake contributes to lifestyle risk profile

The resulting factor pattern is clinically plausible. Previous blood pressure condition, age, family history, smoking, BMI, salt intake, stress, and sleep behavior are all variables commonly associated with cardiometabolic risk. In the application workflow, these variables are collected through a web form and passed to the saved preprocessing-model pipeline. The output consists of a class prediction, risk probability, and basic prevention-oriented recommendations. The system is therefore best framed as an early screening assistant rather than an autonomous diagnostic tool.

The first two importance values are attached to one-hot encoded BP_History categories. This means that the model used previous blood pressure category as the primary partitioning signal. The next group of predictors, consisting of stress score, age, family history, smoking status, BMI, sleep duration, and salt intake, represents a mixture of demographic and lifestyle risk. Their presence in the top-ranked variables supports the practical value of collecting behavioral data, even though behavioral variables alone may not fully reproduce the performance of the full model.

From a health promotion perspective, the feature pattern has a useful implication. Variables such as salt intake, smoking status, sleep duration, stress score, and BMI are connected to modifiable risk behaviors. Even if they are not the strongest discriminators individually, they can support targeted education after a user receives a screening result. Thus, the model output should not only say whether a patient is at risk, but should also guide the user toward measurable prevention actions.

Ablation Study Results

The ablation study was conducted to examine how model performance changed when important predictor groups were removed. This analysis is essential because the full model achieved near-perfect performance. If the full model remained strong after removing BP_History, the result would suggest that lifestyle and demographic attributes alone are sufficient. Conversely, if the performance declined sharply, the result would show that previous blood pressure history is the main explanatory driver.

Table 7. Ablation study of the tuned Gradient Boosting model

Scenario	Removed features	F1-score	ROC-AUC
Full feature set	None	0.9976	1.0000
Without BP_History	BP_History	0.7681	0.8702
Lifestyle-demographic only	BP_History and Medication	0.7640	0.8652
Without top three predictors	BP_History, Stress_Score, Age	0.7343	0.8003
Without behavioral predictors	Salt_Intake, Stress_Score, Sleep_Duration, BMI, Exercise_Level, Smoking_Status	0.7866	0.8714

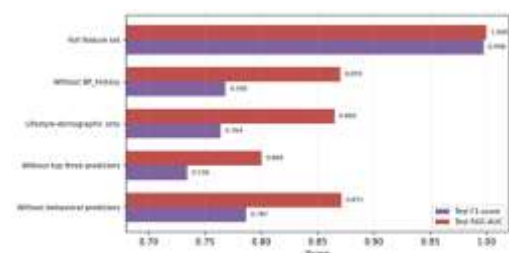


Figure 6. Performance changes under ablation scenarios

The ablation results show that BP_History is the dominant source of predictive separation. Removing only BP_History reduced the test F1-score from 0.9976 to 0.7681 and ROC-AUC from 1.0000 to 0.8702. The lifestyle-demographic scenario produced a similar result, with F1-score of 0.7640 and ROC-AUC of 0.8652. This means that lifestyle and demographic variables still contain meaningful information, but they do not reproduce the near-perfect performance of the full model.

The strongest decline appeared when BP_History, Stress_Score, and Age were removed

together. In this scenario, the test F1-score decreased to 0.7343 and ROC-AUC to 0.8003. This confirms that the model depends on a hierarchy of predictors: previous blood pressure category is the strongest signal, while stress and age provide additional separation. Interestingly, the model without behavioral predictors still achieved F1-score of 0.7866 because it retained Age, BP_History, Medication, and Family_History. This further confirms that clinical and family-history variables carry substantial predictive weight.

The ablation study changes the interpretation of the entire experiment. The full model is still the best model for a screening prototype when BP_History is available. However, if the system is intended for users who do not know their previous blood pressure status, a lifestyle-demographic version should be developed and validated separately. In other words, the system can have two future modes: a clinically informed mode with BP_History and a community self-screening mode without BP_History.

Discussion

Comparison with Related Studies and Methodological Implications

The results of this study are consistent with previous evidence that ensemble-based learning can perform strongly in hypertension risk prediction. Islam et al. [7] showed that population-level risk factors can be modeled using machine learning across South Asian countries. Zhao et al. [8] emphasized the feasibility of using easy-to-collect variables, which is aligned with the current study's reliance on simple demographic, clinical, and lifestyle inputs. Peng et al. [9] and Du et al. [10] further demonstrated that XGBoost and SHAP-based visualization can support hypertension risk modeling, while Hwang et al. [11] highlighted the importance of validating prediction models with large health screening cohorts.

Table 8. Position of the current study among related hypertension prediction studies

Study	Main approach	Relation to current study
Islam et al. [7]	Machine learning on population-level data	Supports demographic and behavioral predictors
Zhao et al. [8]	Risk prediction using easy-to-collect factors	Confirms practical feature design

Peng et al. [9]	Feature selection and XGBoost	Shows boosted trees suit tabular risk data
Du et al. [10]	Visualization model with SHAP	Supports interpretable risk outputs
Hwang et al. [11]	Nationwide cohort model development	Highlights external validation needs

The methodological implication is that algorithm ranking cannot be separated from feature structure. In this dataset, the strong performance of tuned Gradient Boosting is likely driven by the interaction between tree-based learning and highly informative clinical predictors. If BP_History were removed, the model would probably rely more heavily on lifestyle predictors and the performance gap among algorithms might become smaller. For future publication-quality validation, an ablation experiment should therefore be conducted by comparing two scenarios: models with BP_History and models without BP_History. Such an experiment would clarify how much of the predictive power comes from previous clinical status versus modifiable lifestyle factors.

Compared with previous studies, the distinctive element of the present article is not merely the high score of the full model, but the combination of algorithm comparison, deployment pipeline, feature importance, and ablation analysis. This combination provides a more complete basis for decision making. If a health facility has previous blood pressure category data, the tuned Gradient Boosting pipeline is a strong candidate for screening support. If the target users are community members who only know lifestyle information, the ablation result warns that a different validation standard is needed.

Another implication concerns deployment ethics. A web application can improve accessibility, but screening output must be communicated carefully to avoid self-diagnosis. The safest implementation is to present probability as a triage signal that encourages blood pressure measurement and medical consultation. In addition, model monitoring should be planned if the system is used repeatedly, because changes in user population, medication patterns, or measurement behavior can reduce model calibration over time.

Implementation as a Web-Based Screening Prototype

The selected model was serialized as a pipeline using pickle so that preprocessing and prediction steps remain consistent during deployment. The Streamlit interface collects the same ten input features used during training and returns the predicted class and probability of hypertension risk. This implementation is important because a high-performing model has limited practical value if it cannot be used by nontechnical health workers or community users. A simple web interface also makes the model easier to demonstrate, revise, and validate in future field studies.

Several safeguards are necessary before operational use. First, the prediction result should be presented as screening support, not as medical diagnosis. Second, users should be advised to verify risk status with blood pressure measurement and professional consultation. Third, the model should be recalibrated and validated with Indonesian primary-care data, especially because public datasets may not represent local clinical, behavioral, and demographic distributions. Finally, future work should evaluate calibration,

from 0.9976 to 0.7681. This finding confirms that the full model is most appropriate when previous blood pressure category is available, while a lifestyle-demographic version should be developed separately for users who do not know their blood pressure history. The study therefore supports machine learning as an early hypertension screening assistant, but not as a diagnostic substitute. External validation, calibration testing, prospective evaluation, and testing with local Indonesian community health data are required before broader clinical implementation.

References

- [1] W. H. Organization, *Global report on hypertension: The race against a silent killer*. World Health Organization, 2023.
- [2] B. Zhou *et al.*, “Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants,” *Lancet*, vol. 398, no. 10304, pp. 957–980, 2021.
- [3] T. Riskesdas, “Laporan nasional RISKESDAS 2018,” *Jakarta Lemb. Penerbit Badan Penelit. dan Pengemb. Kesehat.*, 2019.
- [4] K. T. Mills *et al.*, “Global disparities of hypertension prevalence and control: a decision thresholds, fairness across demographic groups, and prospective clinical usefulness.
- [5] G. A. Mensah *et al.*, “Global burden of cardiovascular diseases and risks, 1990–2022,” *J. Am. Coll. Cardiol.*, vol. 82, no. 25, pp. 2350–2473, 2023.
- [6] M. M. Ahsan, S. A. Luna, and Z. Siddique, “Machine-learning-based disease diagnosis: A comprehensive review,” in *Healthcare*, 2022, vol. 10, no. 3, p. 541.
- [7] 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, p. 839379, 2022.
- [8] H. Zhao *et al.*, “Predicting the risk of hypertension based on several easy-to-collect risk factors: a machine learning method,” *Front. public Heal.*, vol. 9, p. 429, 2021.
- [9] Y. Peng, J. Xu, L. Ma, and J. Wang, “Prediction of hypertension risks with feature selection and XGBoost,” *J. Mech. Med. Biol.*, vol. 21, no. 05, p. 2140028, 2021.
- [10] J. Du *et al.*, “Developing a hypertension visualization risk prediction system utilizing machine learning and health check-up data,” *Sci. Rep.*, vol. 13, no. 1, p.

4. Conclusion

This study compared five machine learning algorithms for hypertension risk prediction using 1,985 records and ten demographic, clinical, and lifestyle predictors. Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and Support Vector Machine were evaluated through a consistent preprocessing pipeline, 5-fold stratified cross-validation, and independent test evaluation. The tuned Gradient Boosting model achieved the best performance, with a cross-validation F1-score of 0.9994 and ROC-AUC of 1.0000. On the test subset, the model obtained 0.9975 accuracy, 1.0000 precision, 0.9951 recall, 0.9976 F1-score, and 1.0000 ROC-AUC.

The feature importance and ablation analyses provide the main scientific interpretation of the result. Blood pressure history was the dominant predictor; removing it reduced the test F1-score

systematic analysis of population-based studies from 90 countries,” *Circulation*, vol. 134, no. 6, pp. 441–450, 2016.

- 18953, 2023.
- [11] S. H. Hwang *et al.*, “Machine learning–based prediction for incident hypertension based on regular health checkup data: derivation and validation in 2 independent nationwide cohorts in South Korea and Japan,” *J. Med. Internet Res.*, vol. 26, p. e52794, 2024.
- [12] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [13] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [14] G. S. Collins *et al.*, “TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods,” *bmj*, vol. 385, 2024.
- [15] R. F. Wolff *et al.*, “PROBAST: a tool to assess the risk of bias and applicability of prediction model studies,” *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019.
- [16] J. C. Stoltzfus, “Logistic regression: a brief primer,” *Acad. Emerg. Med.*, vol. 18, no. 10, pp. 1099–1104, 2011.
- [17] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [18] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Ann. Stat.*, pp. 1189–1232, 2001.
- [19] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [20] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.