



Classification of Hypertension Severity Levels Using the Random Forest Method

Sophy Awaliah¹, Anindita Septiarini², Novianti Puspitasari^{3*}

^{1,2,3} Program Studi Informatika, Universitas Mulawarman, Indonesia
Jl. Kuaro, Samarinda Ulu, Samarinda, Kalimantan Timur, Indonesia

E-Mail *: novipuspitasari@unmul.ac.id

ABSTRACT

The rising prevalence of hypertension in Indonesia specifically in Samarinda City, which recorded 215,206 cases in 2023 poses challenges for the field of Informatics Engineering regarding the efficiency and consistency of manual disease-level classification in high-patient-volume healthcare facilities. The limitations of conventional approaches create a need for a technology-based system that can classify hypertension severity quickly, accurately, and objectively. This study aims to develop a hypertension severity classification model using the Random Forest algorithm optimized via GridSearchCV and to identify dominant risk factors through Feature Importance analysis. The dataset comprises 5,373 adult patient medical records from the Remaja Community Health Center (Puskesmas) in Samarinda, covering January to December 2024; attributes include gender, age, smoking status, height, weight, systolic and diastolic blood pressure, and a BMI feature derived through feature engineering. Data imbalance was addressed using the SMOTE technique on the training data. Results indicate that the model achieved the highest accuracy of 99.75% with a 70:30 data split, while an 80:20 ratio yielded 99.62% accuracy; both figures exceed the minimum target of 90%. Evaluation using accuracy, precision, recall, and F1-score metrics demonstrates the model's superiority over manual classification. Feature Importance analysis identifies diastolic blood pressure (46.31%) and systolic blood pressure (45.04%) as the most dominant factors, followed by age, BMI, and other attributes. The resulting model can be implemented as a clinical decision support system to assist medical personnel in primary healthcare settings.

Keywords: hypertension, Random Forest, classification, feature importance, SMOTE

ABSTRAK

Peningkatan prevalensi hipertensi di Indonesia, khususnya di Kota Samarinda dengan 215.206 kasus pada tahun 2023, menimbulkan tantangan dalam bidang Teknik Informatika terkait efisiensi dan konsistensi klasifikasi tingkat penyakit secara manual pada fasilitas kesehatan dengan volume pasien tinggi. Keterbatasan pendekatan konvensional mendorong kebutuhan akan sistem berbasis teknologi yang mampu mengklasifikasikan tingkat hipertensi secara cepat, akurat, dan objektif. Penelitian ini bertujuan mengembangkan model klasifikasi tingkat hipertensi menggunakan algoritma *Random Forest* yang dioptimasi dengan *GridSearchCV*, serta mengidentifikasi faktor risiko dominan melalui analisis *Feature Importance*. Data yang digunakan terdiri dari 5.373 rekam medis pasien dewasa Puskesmas Remaja Samarinda periode Januari hingga Desember 2024, dengan atribut meliputi jenis kelamin, usia, status merokok, tinggi badan, berat badan, tekanan darah sistolik dan diastolik, serta fitur BMI yang ditambahkan melalui *feature engineering*. Ketidakseimbangan data ditangani menggunakan teknik SMOTE pada data latih. Hasil penelitian menunjukkan bahwa model dengan pembagian data 70:30 mencapai akurasi tertinggi sebesar 99,75%, sedangkan rasio 80:20 menghasilkan akurasi 99,62%, keduanya melampaui target minimal 90%. Evaluasi menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* membuktikan keunggulan model dibandingkan klasifikasi manual. Analisis *Feature Importance* menunjukkan tekanan darah diastolik (46,31%) dan sistolik (45,04%) sebagai faktor paling dominan, diikuti usia, BMI, dan atribut lainnya. Model yang dihasilkan dapat diimplementasikan sebagai sistem pendukung keputusan klinis untuk membantu tenaga medis di layanan kesehatan primer.

Kata Kunci: hipertensi, *Random Forest*, klasifikasi, *feature importance*, SMOTE



1. INTRODUCTION

Hypertension is a condition characterized by arterial blood pressure exceeding normal levels (120/80 mmHg); it is often asymptomatic yet carries the potential for serious complications such as heart disease, stroke, and kidney disorders or even death (Awalullaili et al., 2023; Bete et al., 2022; Qalsum & Abidin, 2023). The condition often goes undetected and is influenced by various factors and symptoms that contribute to high case rates and the risk of complications (Sapriani & Arisandi, 2024). In its chronic stage, hypertension thickens blood vessel walls, triggering ischemia and metabolic disturbances that can lead to neuronal damage and a decline in cognitive function (Deriana et al., 2025). Consequently, many individuals only become aware of the condition when symptoms have become severe and require medical attention at a healthcare facility (Nurzanah et al., 2022).

The World Health Organization (WHO) identifies hypertension as a leading cause of death globally, resulting in millions of cases of severe, life-threatening complications each year (Siregar et al., 2023). Clinically, hypertension is classified according to the 2025 American College of Cardiology/American Heart Association (AHA/ACC) guidelines, which categorize blood pressure into four main groups: normal (<120/80 mmHg), prehypertension (120–129/<80 mmHg), stage 1 hypertension (130–139/80–89 mmHg), and stage 2 hypertension ($\geq$ 140/90 mmHg) (Johnson et al., 2025). Prevalence increases with age and is influenced by risk factors such as obesity, physical inactivity, smoking, and unhealthy dietary habits (Novianti et al., 2024; Sanger & Lainsamputty, 2022).

The prevalence of hypertension in Indonesia continues to rise, including in East Kalimantan. According to 2023 data from the East Kalimantan Provincial Health Office, Samarinda City recorded the highest figure, with 215,206 cases about 44% of the province's total. This high figure is driven by a large population, an urban lifestyle, and better access to healthcare services, which facilitates the detection of more cases. This situation underscores the importance of optimal hypertension management at the primary care level.

Technological advancements in the healthcare sector have driven the adoption of digital innovations, such as machine learning, in medical services. This technology is widely used for applications such as cardiovascular risk analysis and heart disease detection (Dewi et al., 2022). Machine learning is a computational method that enables systems to learn from data without explicit programming, with performance improving as more data becomes available. The technology is rapidly evolving into a reliable instrument for various medical applications; it is effective for processing big data and conducting probability-based analysis to facilitate accurate decision-making (Nurkholifah et al., 2023; Santoso et al., 2021; Tarimana et al., 2024). Based on learning

methods, machine learning is categorized as supervised or unsupervised learning. Supervised learning models the relationship between known inputs and outputs, with classification grouping data into specific categories as one of its primary applications (Savitri et al., 2021).

Accurate early detection and classification of hypertension severity are crucial steps in determining appropriate medical interventions. However, manual classification has limitations in consistency and efficiency, particularly when handling large patient volumes (Sahila et al., 2024). Advances in machine learning offer a way to overcome these limitations, as the technology is highly effective and widely applied in data classification and clustering tasks. Previous studies have employed methods such as K-Nearest Neighbors, achieving 86.67% accuracy (Rahim, 2021) Support Vector Machines (SVM), achieving 97.15% accuracy (Putri, 2024). Despite these promising results, these methods still face challenges regarding high-dimensional data, the risk of overfitting, and a lack of feature importance analysis within the context of primary healthcare facilities in Indonesia.

One promising machine learning algorithm is Random Forest, an ensemble-based supervised learning method that combines multiple decision trees to generate more stable and accurate predictions (Prabowo & Kurniadi, 2023). In this process, each tree is built from a random subset of data and features, and classification results are determined through majority voting, thereby enhancing model accuracy (Novianti et al., 2024). This approach has also proven effective in handling complex datasets and mitigating the risk of overfitting, making Random Forest a reliable algorithm (Fadli & Saputra, 2023). Research by Aditya et al. (2024) demonstrates that Random Forest can achieve up to 100% accuracy in hypertension classification, outperforming KNN and SVM. Additionally, the algorithm features a "Feature Importance" analysis capability that helps objectively identify dominant risk factors.

Unlike most Random Forest-based hypertension classification studies that use public datasets such as those from Kaggle or the UCI Machine Learning Repository this study uses actual medical record data from the Remaja Samarinda Community Health Center (Puskesmas). Public datasets are generally not population-specific and fail to reflect the demographic characteristics or primary healthcare patterns of a particular region; consequently, models built from such data risk lacking relevance when applied to local populations. Data from the Remaja Samarinda Community Health Center directly reflects the characteristics of adult patients utilizing primary healthcare services in Samarinda—including age ranges, urban lifestyles, and visit patterns specific to the area. Thus, the resulting model is more representative and applicable for supporting clinical decision-making at the Remaja Samarinda Community Health Center and other primary healthcare facilities with similar population

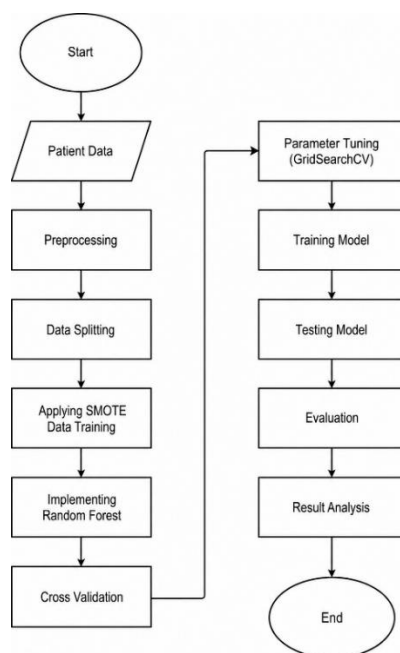
characteristics. Using local data also enhances the model's contextual validity in real-world healthcare settings, in contrast to models trained on public datasets, where data collection processes and attribute definitions are often not documented in detail.

This study aims to develop a classification model for hypertension severity categorized into four levels using the Random Forest algorithm optimized via GridSearchCV and to identify the most influential risk factors through Feature Importance analysis based on patient medical records from the Remaja Community Health Center (Puskesmas) in Samarinda. The study targets a minimum accuracy of 90% and is expected to support the development of clinical decision support systems in primary healthcare facilities.

2. METHOD

2.1 Research Workflow

Figure 1 visualizes the overall research workflow. The process begins with patient data input, followed by a preprocessing stage that includes data cleaning, feature engineering, and encoding. Next, the data is split into training and testing sets using two ratio scenarios. The training data is processed using SMOTE to address class imbalance, followed by implementing a Random Forest model within a unified pipeline. Cross-validation is conducted to evaluate model performance reliably, followed by parameter tuning using GridSearchCV to identify the optimal hyperparameter combination. The best-performing model is then used for final training and tested on the test set; performance is evaluated using various metrics, and the process concludes with an analysis of the results, including feature importance analysis.



Gambar 1. Research Workflow

2.2 Data Collection

The data used in this study were obtained from the medical records of patients at Puskesmas Remaja Samarinda for January to December 2024, in Microsoft Excel (.xlsx) format. After anonymization, we obtained 5,373 records of adult patients (aged 18 years or older) across all visit types. The dataset comprises seven input attributes: gender, age, smoking status, height, weight, systolic blood pressure, and diastolic blood pressure.

2.3 Data Preprocessing

Preprocessing is carried out in stages to ensure data quality before model training (Almufqi & Voutama, 2023). The first stage is data cleaning, which involves removing text units attached to numerical values (e.g., "cm" and "kg"), converting data types to the appropriate numerical format, and handling outliers using an Interquartile Range (IQR)-based capping method. The capping method was selected because it handles extreme values without requiring data deletion, thereby preserving the sample size. Values falling below the lower bound ($Q1 - 1.5 \times IQR$) are replaced with the lower bound value, while values exceeding the upper bound ($Q3 + 1.5 \times IQR$) are replaced with the upper bound value.

The second stage is feature engineering, which involves adding a new feature Body Mass Index (BMI) calculated by dividing weight (kg) by the square of height (m) (Hasibuan & A, 2021). Additionally, a categorical BMI status feature is created based on WHO standards: underweight (BMI < 18.5), normal (18.5–24.9), overweight (25–29.9), and obese (BMI > 30) (Wu et al., 2024). This feature addition aims to provide the model with extra information regarding the patient's weight status, a factor closely linked to hypertension risk. The third stage is encoding, which converts categorical variables into numerical representations using label encoding for the attributes of gender (0 = male, 1 = female) and smoking status (0 = non-smoker, 1 = smoker). The final stage involves constructing the target variable based on the combination of systolic and diastolic blood pressure values, in accordance with the 2025 AHA/ACC classification guidelines, as shown in Table 1.

Table 1. Hipertension Target Variable

Category	Sistolic	Diastolic	Label
Normal	< 120	< 80	0
Prehypertension	120 – 129	< 80	1
State 1 Hypertension	130 – 139	80 – 89	2
State 2 Hypertension	≥ 140	≥ 90	3

Systolic and diastolic attributes are included as inputs, even though they form the basis for the target labels in Table 1. This design choice reflects that, because the labels are derived deterministically from the 2025 AHA/ACC thresholds, the model essentially

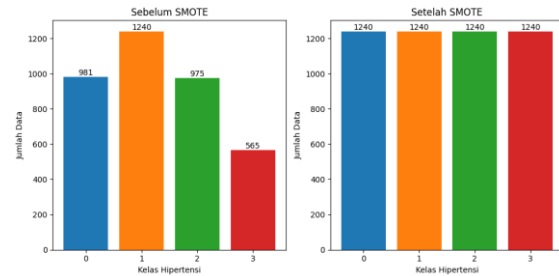
represents established clinical rules computationally rather than discovering new patterns. The objective is to quantitatively validate via feature importance analysis that these two attributes are indeed the primary determinants of hypertension classification, while also measuring the contribution of secondary attributes (age, BMI, sex, and smoking status); this serves as the foundation for an automation prototype capable of consistently replicating AHA/ACC rules across large datasets. Consequently, an accuracy approaching 100% reflects the model's consistency in learning these threshold-based rules, rather than an ability to uncover novel risk patterns.

2.4 Data Splitting

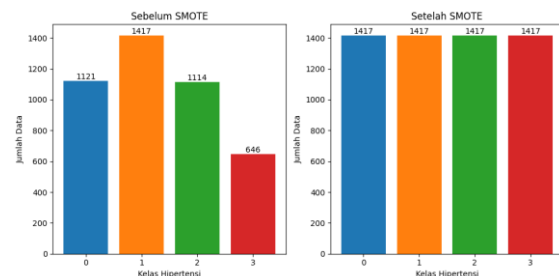
The dataset was split using two different train-test split scenarios to compare the effect of the data split ratio on model performance. The first scenario employed a 70:30 ratio, with 70% of the data (3,762 records) used for training and 30% (1,613 records) for testing. The second scenario used an 80:20 ratio, with 80% of the data (4,300 records) used for training and 20% (1,075 records) for testing. The split was performed randomly using a fixed random state to ensure reproducible results.

2.5 Handling Class Imbalance

Analysis of the class distribution reveals a significant imbalance, particularly regarding the Stage 2 Hypertension class, which contains fewer samples than the other classes. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was employed; this technique generates new synthetic samples for the minority class by interpolating between existing samples based on the K-Nearest Neighbor principle (Aprihartha et al., 2024; Nurussakinah et al., 2025). SMOTE was applied exclusively to the training data, not the test data; this prevents data leakage, which could otherwise lead to invalid model evaluations that are unrepresentative of real-world conditions. Specifically, the process sequence implemented in this study was as follows: (1) the dataset was first divided into training and test sets using a train-test split; (2) SMOTE was applied solely to the resulting training data; and (3) the test data remained in its original form without oversampling and was never involved in the synthetic sample generation process. Because SMOTE was not applied to the entire dataset before the split, no information leakage occurred from the test data into the model training process.



Gambar 2. Class Distribution Before and After SMOTE (70:30)



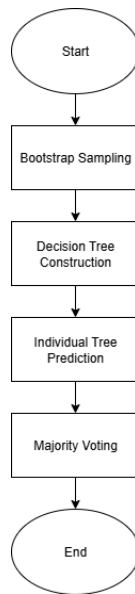
Gambar 3. Class Distribution Before and After SMOTE (80:20)

Figures 2 and 3 compare class distributions before and after applying SMOTE for each scenario. Before SMOTE, the class distribution for the 70:30 scenario consisted of 981 data points for Class 0 (Normal), 1,240 for Class 1 (Prehypertension), 975 for Class 2 (Stage 1 Hypertension), and only 565 for Class 3 (Stage 2 Hypertension). After applying SMOTE, all classes were balanced, with each containing 1,240 data points. A similar pattern was observed in the 80:20 scenario, where class distributions were successfully balanced to 1,417 data points per class after SMOTE.

2.6 Modeling and Hyperparameter Optimization

The Random Forest model was developed using the scikit-learn library in Python within the Google Colab environment. The entire workflow, from SMOTE application to model training, was encapsulated in a single, integrated pipeline, ensuring that oversampling occurred exclusively on the training data in each cross-validation fold. Hyperparameter optimization was performed using GridSearchCV with k-fold cross-validation (k=5), which involves partitioning the training data into five segments, with each segment serving as the validation set in turn. The optimized parameters included *n_estimators* (number of decision trees), *max_depth* (maximum tree depth), *min_samples_split* (minimum samples required to split a node), dan *min_samples_leaf* (minimum samples required at a leaf node). The GridSearchCV results yielded the optimal parameter combination, which was subsequently used to train the final model.

In general, Figure 4 visualizes the operational stages of the Random Forest algorithm in this modeling process.



Gambar 4. Stages of the Random Forest Algorithm

The algorithm operates through four main stages. First, bootstrap sampling generates multiple data subsets randomly selected with replacement from the training data, creating variation among trees and reducing overfitting. Second, each subset is used to construct a decision tree; during this process, feature selection at each node utilizes only a small, randomly chosen subset of features, specifically the square root of the total number of features ($\sqrt{m}$) to maintain diversity among the trees. Third, each constructed tree independently predicts the test data based on the patterns it has learned (*individual tree prediction*). Fourth, the predictions from all trees are aggregated via majority voting, where the class receiving the most votes becomes the model's final output, resulting in a classification that is more stable and accurate than a single tree's prediction.

2.7 Model Evaluation

We evaluated model performance using a confusion matrix and four key metrics. Accuracy measures the proportion of correct predictions relative to the total number of predictions. Precision measures the proportion of data predicted as a specific class that actually belongs to that class. Recall measures the proportion of data actually belonging to a specific class that the model successfully identified. The F1-score is the harmonic mean of precision and recall, providing a balanced view of both metrics. Additionally, we performed a feature importance analysis to measure each attribute's relative contribution to the model's decision-making process, calculated as the mean decrease in impurity across the trees in the ensemble.

3. DISCUSSION

3.1. Data Preprocessing Results

Following the data cleaning stage, all dataset columns were successfully converted to appropriate numeric data types. An outlier-handling process using the IQR-based capping method constrained extreme values in the systolic, diastolic, and BMI attributes without removing any data rows. Consequently, the dataset size remained at 5,373 records, while data quality improved and the distribution became more representative.

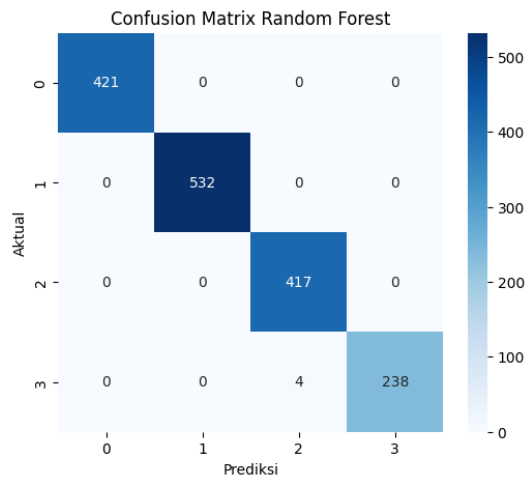
During feature engineering, BMI was successfully added as a new attribute, increasing the total number of input attributes from seven to eight (the original seven plus BMI and BMI status). The target variable was then constructed based on combinations of systolic and diastolic blood pressure values in accordance with 2025 AHA/ACC guidelines, resulting in the following class distribution: Normal (label 0): 1,404 records; Prehypertension (label 1): 1,773 records; Stage 1 Hypertension (label 2): 1,393 records; and Stage 2 Hypertension (label 3): 805 records. We then addressed this imbalanced distribution using SMOTE on the training data in each data-splitting scenario.

3.2. Hyperparameter Optimization Results

The GridSearchCV process yielded distinct optimal hyperparameter combinations for each data-splitting scenario. For the 70:30 scenario, the best parameters obtained were $n_estimators = 200$, $max_depth = 7$, $min_samples_split = 10$, and $min_samples_leaf = 5$. In the 80:20 scenario, the optimal parameters were similar, with only minor adjustments. Cross-validation produced near-perfect average scores in both scenarios. These exceptionally high cross-validation results are a logical consequence of the research design, in which the target variable is derived directly from the blood pressure values that also serve as input features. Consequently, the model essentially learns consistent clinical threshold rules rather than hidden, undefined patterns. This differs from conventional prediction scenarios, where labels are unknown beforehand, and the model must discover more complex patterns.

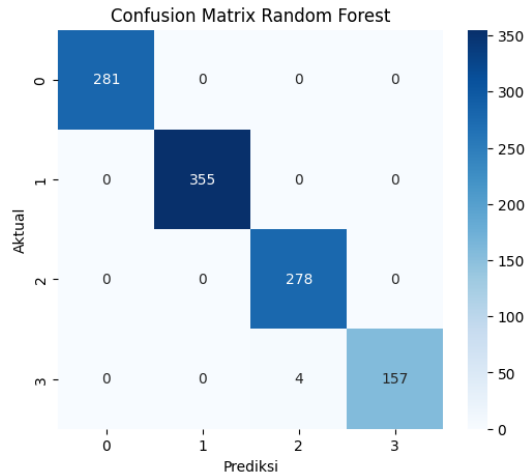
3.3. Model Test Results

Figures 4 and 5 present the model evaluation results on the test data using confusion matrices. A confusion matrix is a visual representation of a classification model's performance, showing the number of correct and incorrect predictions for each class. The rows of the matrix represent the actual classes (ground truth), while the columns represent the classes predicted by the model. Values along the main diagonal (from top-left to bottom-right) represent correct predictions, whereas values off the diagonal represent classification errors.



Gambar 5. Confusion Matrix Skenario 70:30

In the 70:30 scenario (Figure 4), the model successfully classified 421 Normal class data points (label 0), 532 Prehypertension data points (label 1), 417 Stage 1 Hypertension data points (label 2), and 238 Stage 2 Hypertension data points (label 3). Only four data points were misclassified—specifically, four Stage 2 Hypertension instances predicted as Stage 1 Hypertension. This error likely occurred in borderline cases where blood pressure values were very close to the threshold between the two classes. In this scenario, the total number of test data points was 1,612, resulting in an accuracy of 99.75%.



Gambar 6. Confusion Matrix Skenario 80:20

In the 80:20 scenario (Figure 5), the model successfully classified 281 Normal, 355 Prehypertension, 278 Stage 1 Hypertension, and 157 Stage 2 Hypertension data points. Similar to the previous scenario, four Stage 2 Hypertension cases were misclassified as Stage 1 Hypertension. The total number of test samples in this scenario was 1,075, resulting in 99.62% accuracy. The consistent error pattern across both scenarios indicates that borderline cases between Stage 1 and Stage 2 Hypertension pose a specific challenge that requires attention during clinical implementation.

Table 2. Comparison of Model Evaluation Results

Matrix	70:30	80:20
Accuracy	99,75%	99,62%
Precision	1,00	1,00
Recall	1,00	0,99
F1-score	1,00	0,99

3.4. Comparison with Previous Studies

Table 2 summarizes the comparison of evaluation results between the two scenarios. Both scenarios performed excellently, significantly exceeding the minimum 90% accuracy target established at the start of the study. The 70:30 scenario achieved an accuracy of 99.75%, with all macro-average metrics reaching a perfect score of 1.00. The 80:20 scenario achieved an accuracy of 99.62% with a macro-precision of 1.00; however, the macro-recall and F1-score were slightly lower at 0.99, reflecting minor misclassifications in the Stage 2 Hypertension class.

Although the accuracy difference between the two scenarios was only 0.13%, the 70:30 scenario is considered more optimal overall. This is because the larger proportion of test data (30%, or 1,612 samples) in the 70:30 scenario provides a more representative evaluation and better reflects the model's generalization capability regarding previously unseen data. The larger the test dataset used, the more reliable the estimate of the model's performance in real-world conditions becomes.

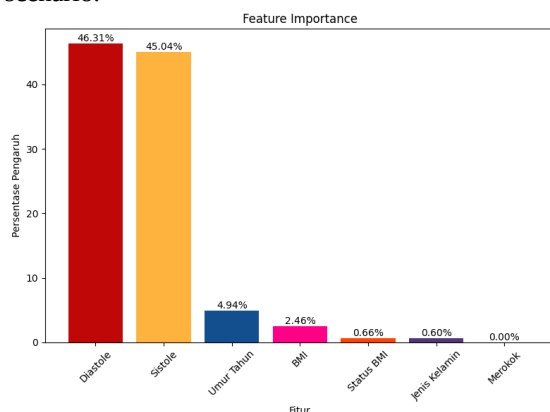
Tingginya akurasi pada kedua skenario dapat dipahami karena variabel target dalam penelitian ini dibentuk secara deterministik berdasarkan nilai tekanan darah sistolik dan diastolik sesuai panduan AHA/ACC. Hal ini berarti model pada dasarnya mempelajari aturan threshold yang sama dengan yang digunakan dalam pembentukan label, sehingga pola yang harus dipelajari bersifat eksplisit dan konsisten. Peningkatan performa model juga didukung oleh penerapan SMOTE yang memastikan model mendapatkan representasi yang seimbang dari seluruh kelas hipertensi selama pelatihan, serta *GridSearchCV* yang mengoptimalkan *Hyperparameter* secara sistematis.

Compared with previous studies, the results in this research demonstrate significant superiority. (Soraya et al., 2023) using KNN memperoleh akurasi 62,60%, Rahim (2021) with MKNN achieved 86,67%, and Purwono et al. (2022) achieved 85% using ANN while Awalullaili et al. (2023) achieved 89,22% using SVM. Several factors support the Random Forest model's superior performance in this study: ensemble capabilities that reduce variance and bias, systematic hyperparameter optimization using *GridSearchCV*, and effective handling of data imbalance via SMOTE.

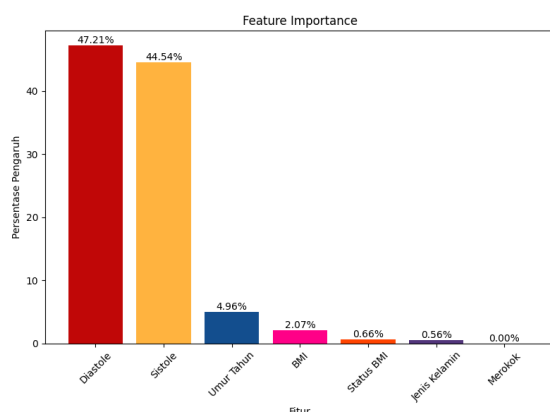
3.5. Feature Importance Analysis

One key advantage of the Random Forest algorithm is its ability to provide interpretable feature importance. Feature importance values are calculated

based on the mean decrease in Gini impurity across all decision trees in the ensemble. The higher its attribute's feature importance value, the greater its contribution to separating the classes and the more significant its role in the classification process. Figures 6 and 7 present the feature importance analysis results for each scenario.



Gambar 7. Scenario Feature Importance Graph 70:30



Gambar 8. Scenario Feature Importance Graph 80:20

Table 3. Comparison of Feature Importance

Attribute	70:30	80:20
Diastole	46,31%	47,21%
Systole	45,04%	44,54%
Age (Years)	4,94%	4,96%
BMI	2,46%	2,07%
BMI Status	0,66%	0,66%
Gender	0,60%	0,56%
Smoking Status	0,00%	0,00%

Based on Table 3 and Figures 6 and 7, the feature importance analysis results for both scenarios reveal a highly consistent pattern. Diastolic blood pressure is the most dominant factor, contributing 46.31% in the 70:30 scenario and 47.21% in the 80:20 scenario. Systolic blood pressure ranks second, contributing 45.04% and 44.54% in the respective scenarios. The combined contribution of these two blood pressure attributes exceeds 91%; this is expected, as the target variable in this study is defined directly from these

attributes in accordance with AHA/ACC guidelines. The consistency of results across scenarios also confirms that the developed model is stable and not dependent on a specific data-splitting ratio.

The age attribute ranks third, with contributions of 4.94% (70:30) and 4.96% (80:20). This aligns with established medical knowledge that hypertension risk increases progressively with age. After the age of 50, isolated systolic hypertension becomes the most common form due to arterial stiffening associated with aging. The significant contribution of age indicates that the model successfully captured this pattern from the medical record data.

The BMI attribute contributes 2.46% (70:30) and 2.07% (80:20). This contribution shows that excess weight helps distinguish hypertension classes, consistent with scientific evidence identifying obesity as an independent risk factor for hypertension. Weight gain increases blood volume and cardiac output, ultimately raising blood pressure. BMI status contributes 0.66% in both scenarios, while gender contributes 0.60% (70:30) and 0.56% (80:20). Despite their minor contribution, the model still identified both attributes as factors carrying relevant information for classification.

The smoking status attribute recorded a Feature Importance value of 0.00% in both scenarios. It is crucial to understand that this value does not imply smoking habits have no clinical impact on hypertension; medically, smoking is known to raise blood pressure through nicotine's effect of stimulating the sympathetic nervous system and causing vasoconstriction. The 0.00% value stems from an extreme imbalance in the distribution of smokers within the dataset: only 30 records of smokers out of a total of 5,373 (approximately 0.56%). Given this very small proportion, the model did not capture sufficient statistical signal to recognize the smoking attribute's contribution to the classification decision-making process. This represents a data limitation rather than a reflection of the clinical relevance of smoking habits as a hypertension risk factor.

This limited proportion of smoker data should be viewed from a health informatics perspective rather than merely as a statistical finding. The figure of 30 out of 5,373 records (0.56%) likely does not reflect the true prevalence of smoking in the service area of Puskesmas Remaja Samarinda; instead, it points to gaps in medical record documentation—such as the self-reported nature of smoking status (which is prone to under-reporting), the absence of a mandatory field for this attribute, and the lack of integrated, routine screening for behavioral risk factors. Consequently, the 0.00% Feature Importance value does not indicate that smoking lacks clinical relevance; rather, it indicates the need to improve data-recording systems in primary healthcare facilities so that this variable can be used optimally in future research.

4. CONCLUSION

This study developed a classification model for hypertension severity, categorized as Normal, Prehypertension, Stage 1 Hypertension, and Stage 2 Hypertension, using the Random Forest algorithm optimized via GridSearchCV on medical record data from the Remaja Community Health Center (Puskesmas) in Samarinda. The model using a 70:30 data split achieved the highest accuracy (99.75%), with macro-average precision, recall, and F1-score of 1.00 each. Meanwhile, the 80:20 scenario yielded an accuracy of 99.62%, with a macro-precision of 1.00 and macro-recall and F1-score of 0.99. Both scenarios significantly exceeded the minimum accuracy target of 90%, and the 70:30 scenario was selected as the optimal configuration because its generalization evaluation was more representative.

Feature Importance analysis revealed that diastolic (46.31%) and systolic (45.04%) blood pressure were the most dominant factors, contributing a combined total of over 91%, followed by age (4.94%) and BMI (2.46%). The consistency of Feature Importance results across scenarios confirms the model's reliability in identifying key hypertension factors. The negligible contribution of the smoking attribute (0.00%) was attributed to the imbalanced representation of smokers in the dataset, rather than a lack of clinical relevance for the factor itself.

The developed model could be implemented as a machine learning-based clinical decision support system, helping medical personnel rapidly and objectively classify hypertension severity in primary healthcare facilities and improving service efficiency and the quality of patient care. For future research, it is recommended to compare the model with other algorithms such as XGBoost, LightGBM, or Support Vector Machines to identify the optimal method using larger, more diverse datasets from multiple healthcare institutions, thereby enhancing generalizability. Additionally, incorporating relevant clinical variables such as family history, physical activity, salt intake patterns, and duration of hypertension would enrich the information the model learns. As a further methodological validation, future studies should also conduct experimental scenarios that exclude systolic and diastolic blood pressure attributes, predicting hypertension levels based solely on secondary attributes (age, gender, BMI, and smoking status). Comparing accuracy between scenarios with and without blood pressure attributes would provide independent quantitative evidence of the contribution of these secondary attributes and strengthen the model's overall validity; however, because this scenario was not performed in the current study, it remains an explicit recommendation for future development.

REFERENCES

- Aditya, M. F. R., Azizah, N. L., & Indahyanti, U. (2024). Prediksi Penyakit Hipertensi Menggunakan Metode Decison Tree dan *Random Forest*. *Jurnal Ilmiah Komputasi*, 23(1).
<https://doi.org/10.32409/jikstik.23.1.3503>
- Almufqi, F. M., & Voutama, A. (2023). PERBANDINGAN METODE DATA MINING UNTUK MEMPREDIKSI PRESTASI AKADEMIK SISWA. *Jurnal Teknika*, 15(1), 61–66.
<https://doi.org/10.30736/jt.v15i1.929>
- Aprihartha, M. A., Putrawan, Z., Zulhan, D., & Nurfaizal, F. A. (2024). Algoritma Synthetic Minority Oversampling Technique dan C5.0 dalam Mengatasi Ketidakseimbangan Data pada Klasifikasi Kelulusan Siswa. *UPGRADE : Jurnal Pendidikan Teknologi Informasi*, 2(1), 1–10.
<https://doi.org/10.30812/upgrade.v2i1.4148>
- Awalullaili, F. O., Ispriyanti, D., & Widiharini, T. (2023). Klasifikasi Penyakit Hipertensi Menggunakan Metode SVM Grid Search Dan SVM Genetic Algorithm (GA). *Jurnal Gaussian*, 11(4), 488–498.
<https://doi.org/10.14710/j.gauss.11.4.488-498>
- Bete, D., Kurniyanti, M. A., & Mayasari, S. I. (2022). Terapi Tertawa Terhadap Tekanan Darah Pada Lansia Dengan Hipertensi. *Jurnal Keperawatan*.
<http://journal.stikeskendal.ac.id/index.php/Kepawatan>
- Deriana, I. G., Gaghauna, E. E. M., & Santoso, B. R. (2025). Hubungan Klasifikasi Hipertensi Terhadap Penurunan Fungsi Kognitif Pasien Hipertensi di Puskesmas Pekauman Kota Banjarmasin. *Journal of Health (JoH)*, 12(1), 010–018.
<https://doi.org/10.30590/joh.v12n1.704>
- Dewi, P., Purwono, P., & Kurniawan, S. D. (2022). Pemanfaatan Teknologi *Machine learning* pada Klasifikasi Jenis Hipertensi Berdasarkan Fitur Pribadi. *Smart Comp: Jurnalnya Orang Pintar Komputer*.
- Fadli, M., & Saputra, R. A. (2023). Klasifikasi Dan Evaluasi Performa Model *Random Forest* Untuk Prediksi Stroke. *JT : Jurnal Teknik*, 12.
<http://jurnal.umat.ac.id/index.php/jt/index>
- Hasibuan, M. U. Z., & A, Palmizal. (2021). Sosialisasi Penerapan Indeks Massa Tubuh (IMT) di Suta Club. *Jurnal Cerdas Sifa Pendidikan*, 10, 19–24. <https://online-journal.unja.ac.id/csp>
- Johnson, H. M., Shimbo, D., Abdalla, M., Altieri, M. M., Bress, A. P., Carter, J., Commodore-Mensah, Y., Egan, B., Melnyk, B. M., Ogunniyi, M. O., Schott, S. L., Watson, K. E., Williamson, J. D., Jones, D. W., Ferdinand, K. C., Taler, S. J., Bansal, N., Bello, N. A., Cohen,

- J. B., ... Whelton, P. K. (2025). 2025 AHA/ACC/AANP/AAPA/ABC/ACCP/ACPM/AGS/AMA/ASPC/NMA/PCNA/SGIM Guideline for the Prevention, Detection, Evaluation and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*, 152(11), e114–e218.
<https://doi.org/10.1161/CIR.00000000000001356>
- Novianti, N., Putri Agustini Alkadri, S., & Fakhruzi, I. (2024). Klasifikasi Penyakit Hipertensi Menggunakan Metode *Random Forest*. *Progresif: Jurnal Ilmiah Komputer*.
- Nurkholifah, M., Jasmarizal, Umar, Y., & Rahmaddeni. (2023). Analisa Performa Algoritma *Machine learning* Dalam Prediksi Penyakit Liver. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, 4(1), 164–172.
<https://doi.org/10.35870/jimik.v4i1.149>
- Nurussakinah, N., Faisal, M., & Santoso, I. B. (2025). Algoritma *Random Forest* dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes. In *Jurnal Informatika Sunan Kalijaga* (Vol. 10, Number 2). MEI.
<https://doi.org/10.14421/jiska.2025.10.2.221-234>
- Nurzanah, S. C., Alam, S., & Hermanto, T. I. (2022). Analisis Association Rule Untuk Identifikasi Pola Gejala Penyakit Hipertensi Menggunakan Algoritma Apriori (Studi Kasus: Klinik Rafina Medical Center). *Jurnal Informatika Dan Komputer) Akreditasi KEMENRISTEKDIKTI*, 5(2). <https://doi.org/10.33387/jiko>
- Prabowo, A. S., & Kurniadi, F. I. (2023). Analisis Perbandingan Kinerja Algoritma Klasifikasi dalam Mendeteksi Penyakit Jantung. *Jurnal Sistem Komputer Dan Kecerdasan Buatan*.
<https://doi.org/https://doi.org/10.47970/siskom-kb.v7i1.468>
- Purwono, P., Dewi, P., Wibisono, S. K., Putra Dewa, B., Informatika, P., Harapan Bangsa, U., Keperawatan, P., & Bangsa, H. (2022). Model Prediksi Otomatis Jenis Penyakit Hipertensi dengan Pemanfaatan Algoritma *Machine learning* Artificial Neural Network. *INSECT (Informatics and Security): Jurnal Teknik Informatika*, 7(2), p.
- Putri, H. Z. E. (2024). Implementasi Metode *Support Vector Machine* Pada Klasifikasi Diagnosis Penyakit Hipertensi [Universitas Islam Negeri Maulana Malik Ibrahim]. <http://etheses.uin-malang.ac.id/id/eprint/64256>
- Qalsum, U., & Abidin, W. (2023). Klasifikasi Penyakit Hipertensi Menggunakan Metode K-Means Clustering. In *Jurnal Matematika dan Statistika serta Aplikasinya* (Vol. 11, Number 2).
- Rahim, H. (2021). *Klasifikasi Tingkat Risiko Hipertensi Menggunakan Metode Modified K-Nearest Neighbor* [Universitas Brawijaya].
<https://repository.ub.ac.id/id/eprint/184436>
- Sahila, R., Widiharhi, T., & Utami, I. T. (2024). Analisis Klasifikasi Menggunakan Regresi Logistik Biner Dan Algoritma Naive Bayes Classifier Pada Penyakit Hipertensi. *Jurnal Gaussian*, 13(2), 319–327.
<https://doi.org/10.14710/j.gauss.13.2.319-327>
- Sanger, A. Y., & Lainsamputty, F. (2022). Stres Dan Komponen Tekanan Darah Pada Penderita Hipertensi. *Nutrix Journal*.
<https://doi.org/https://doi.org/10.37771/nj.Vol6.Iss1.604>
- Santoso, P., Abijono, H., & Anggreini, N. L. (2021). Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data. *Jurnal Teknologi Terapan*, 4(2).
<https://doi.org/https://doi.org/10.33379/gtech.v4i2.635>
- Sapriani, A., & Arisandi, Y. (2024). Faktor-Faktor Resiko Yang Mempengaruhi Terhadap Penderita Hipertensi. *Jurnal Ilmiah Multi Science Kesehatan*, 16(1), 218–226.
<https://jurnal.stikes-aisyiyah-palembang.ac.id/index.php/Kep/article/view/>
- Savitri, N. L. P. C., Rahman, R. A., Venyutzky, R., & Rakhmawati, N. A. (2021). Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning. *Jurnal Teknik Informatika Dan Sistem Informasi*, 7(1).
<https://doi.org/10.28932/jutisi.v7i1.3216>
- Siregar, A. P., Purba, D. P., Pasaribu, J. P., & Bakara, K. R. (2023). Implementasi Algoritma *Random Forest* Dalam Klasifikasi Diagnosis Penyakit Stroke. *Jurnal Penelitian Rumpun Ilmu Teknik*, 2(4), 155–164.
<https://doi.org/10.55606/juprit.v2i4.3039>
- Soraya, R., Pratama, P., Hayati, M. N., & Goejantoro, R. (2023). Klasifikasi Status Hipertensi Pasien UPTD Puskesmas Sempaja, Kota Samarinda Menggunakan Metode K-Nearest Neighbor. *Jurnal EKSPONENSIAL*, 14(2).
<http://jurnal.fmipa.unmul.ac.id/index.php/expo-nensial>
- Tarimana, A. A., Fajar, M. R. S., Saktiawan, M. A., & Saputra, R. A. (2024). Prediksi Penyakit Hipertensi Menggunakan *Machine learning* Dengan Algoritma Regresi Logistik. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Number 6).
- Wu, Y., Li, D., & Vermund, S. H. (2024). Advantages and Limitations of the Body Mass Index (BMI) to Assess Adult Obesity. In *International Journal of Environmental Research and Public Health* (Vol. 21, Number 6). Multidisciplinary Digital Publishing Institute (MDPI).
<https://doi.org/10.3390/ijerph21060757>

HALAMAN INI SENGAJA DIKOSONGKAN