Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 5, No. 2, May 2026, pp. 344-352
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI: https://doi.org/10.69916/jkbti.v5i2.493

PREDICTION OF HYPERTENSION COMORBIDITIES USING THE RANDOM FOREST ALGORITHM (CASE STUDY AT PUSKESMAS KASIHAN 2)

R. Muhammad Luqman Harjito[*1], Ahmad Subhan Yazid[2], Dita Danianti[3], Dhina Puspasari Wijaya[4]

[1,2,3,4]Informatics, Faculty of Science, Engineering, and Technology, Alma Ata University, Yogyakarta, Indonesia
Email: [1]223200256@almaata.ac.id, [2]subhan@almaata.ac.id, [3]dita@almaata.ac.id, [4]dhina.puspa@almaata.ac.id

(Received: May 15, 2026; Revised: May 21, 2026; Published: May 25, 2026)

Abstract

Hypertension is a chronic non-communicable disease often accompanied by comorbidities, which can increase complication risks and reduce patients' quality of life. Currently, the identification of comorbidities in hypertension patients is commonly performed manually, making the process time-consuming and highly dependent on the attentiveness of healthcare personnel. This study aims to develop a predictive model for hypertension comorbidities using a machine learning-based Random Forest algorithm, designed as an early screening tool for the general population. The research method follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which includes six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Clinical data were collected from medical records, focusing on predicting eight primary comorbidities using a multi-label classification approach. Data preprocessing consisted of data cleaning, transformation, train-test splitting, and handling class imbalance. The Random Forest model was trained and evaluated using subset accuracy, hamming loss, micro precision, micro recall and micro F1-score metrics. The results demonstrate that the Random Forest algorithm successfully predicts hypertension comorbidities with a subset accuracy of 0.8684, hamming loss of 0.0227, micro precision of 0.9032, micro recall of 0.9749 and micro F1-score of 0.9377. The model was successfully deployed into a Streamlit-based web application, enabling healthcare professionals to obtain direct prediction results. This system is expected to assist in the early screening and monitoring of hypertension patients.

Keywords: hypertension; comorbidity; random forest; multi-label classification; CRISP-DM.

1. INTRODUCTION

Hypertension, or high blood pressure, is a significant global health issue. It is often referred to as a "silent killer" because it frequently presents no clear symptoms but can lead to severe complications such as stroke, heart disease, and kidney failure [1], [2]. The high prevalence of hypertension requires attention not only to the detection of the disease itself but also to the comorbidities that frequently accompany it. The presence of comorbidities significantly increases the burden of care and demands comprehensive monitoring.

At primary healthcare facilities, the identification of comorbidity risks is still largely conducted manually through diagnostic examinations and patient interviews. This process highly depends on the experience and accuracy of medical personnel, which risks delaying detection and timely intervention. Therefore, an effective early screening tool is needed that can predict the risk of comorbidities in the general population, allowing high-risk patients to be identified and managed more effectively. According to a global report released by the World Health Organization (WHO) in 2023, the prevalence of hypertension has reached a significant level globally, with an estimated 1.3 billion adults aged 30–79 worldwide living with hypertension [3].

At the provincial level, hypertension remains a significant health issue in the Special Region of Yogyakarta (DIY). The 2024 Government Performance Report (LKIP) of the DIY Health Office noted that the prevalence rate of non-communicable diseases (NCDs) in DIY is above the national average, with the number of hypertension patients in DIY recorded at 143,382 people in 2023 and 131,221 people in 2024 [4]. When examined by regency/city, the DIY Health Profile published in 2024 shows that the estimated number of hypertension patients aged more than 15 years in DIY reached 191,573 people. Bantul Regency had the highest number with 49,306 people, followed by Sleman Regency with 47,084 people, Kulon Progo Regency with 33,985 people, Yogyakarta City with 32,972 people, and Gunungkidul Regency with 28,226 people [5]. The high prevalence of hypertension warrants significant attention, as it is closely associated with various comorbidities and cardiometabolic complications (e.g., coronary heart disease and stroke) that can exacerbate patient health outcomes[6]. At the Puskesmas (primary healthcare center) level, hypertensive patients frequently present with comorbidities, with diabetes mellitus being one of the most commonly encountered. This is evidenced by studies based on Puskesmas medical records, which report diabetes as the most frequent comorbidity among hypertensive patients [7]. Regarding comorbidities in hypertensive patients, the

interviewee stated that the most frequently encountered conditions include high cholesterol (ICD-10 E78), heart disease (I11), type 2 diabetes mellitus (E11), and renal failure (N19). In addition to the interview findings, the determination of comorbidity labels was also established during the Data Understanding stage, based on the distribution of diagnoses within the medical record data. Given that comorbidities can increase the healthcare burden and necessitate more comprehensive monitoring, data-driven approaches such as machine learning can be leveraged to develop comorbidity prediction systems. This enables high-risk patients to be identified earlier and receive more targeted interventions. [8]. Previous studies have demonstrated the efficacy of machine learning algorithms, particularly Random Forest, in predicting hypertension with high accuracy [9].

Previous studies have demonstrated the effectiveness of machine learning algorithms in predicting hypertension. One study, titled "Classification of Hypertension Using the Random Forest Method," developed a prediction model using the Random Forest algorithm based on the PPG-BP Database, which contains 219 data points. The dataset was randomly split into 80% for training and 20% for testing. The evaluation yielded high performance, achieving 98% accuracy on the training data and 95% on the testing data. This indicates that the Random Forest model generalizes well to unseen data [10]. Another study, titled "Implementation of Support Vector Machine (SVM) and Random Forest Algorithms for Hypertension Classification Based on Health Data," utilized medical check-up records from the Anggadita Public Health Center. The researchers initially collected 2,500 patient records, which were reduced to 2,462 after preprocessing and removing duplicates. The dataset was split into 80% training (1,969 records) and 20% testing (493 records). Evaluation using a confusion matrix showed that Random Forest outperformed SVM with an accuracy of 96% compared to SVM's 93%. Consequently, Random Forest is recommended for classifying normal and hypertensive patients using healthcare service data [11] Furthermore, a study titled "Comparison of Random Forest and Naïve Bayes Algorithm Accuracy in Predicting Hypertension Risk" utilized a health dataset covering over 1,000 patients. During the modeling phase, the dataset was split into 80% training data and 20% testing data. The evaluation results showed that Random Forest achieved the best performance with 91% accuracy, outperforming Naïve Bayes which achieved 85% accuracy. Thus, Random Forest was concluded to be superior and more consistent for predicting hypertension risk [9].

However, these studies primarily focus on predicting the occurrence or classification of hypertension status and have not specifically addressed the simultaneous prediction of primary comorbidities. This research bridges the gap by developing a multi-label predictive model using the Random Forest algorithm. The objective of this study is to build a machine learning model capable of predicting the probability of multiple comorbidities in hypertension patients within the general population.

2. RESEARCH METHODS

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. The methodology consists of the following phases:

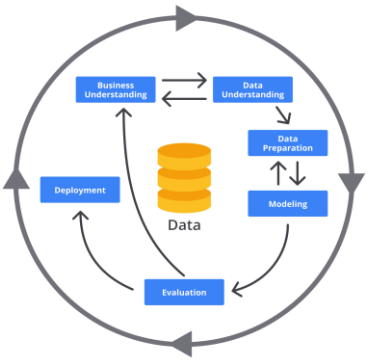

Figure 1. CRISP-DM Process Flow

The Cross-Industry Standard Process for Data Mining (CRISP-DM) serves as a standardized process model and framework for the systematic and structured execution of data mining projects. CRISP-DM encompasses six fundamental stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [12].

2.1. Business Understanding

The initial stage of system development begins with understanding the requirements (business understanding), which involves identifying the problem: the absence of a prediction model for hypertension comorbidities. At this stage, the researcher analyzes the main objective of the system, which is to serve as an initial screening tool for comorbidity risks in hypertensive patients based on the data available in medical records.

2.2. Data Understanding

This stage focuses on understanding the characteristics of patient medical record data used to develop the comorbidity prediction model. At this stage, the researcher identifies the data sources, the types and number of available attributes, and performs an initial descriptive analysis to observe the distribution of the data. The researcher also assesses data quality by checking for the presence of missing values. The outcomes of this stage serve as the foundation for determining the data cleaning and selection procedures, which are subsequently executed more technically during the Data Preparation stage. This research uses the patient dataset taken from Puskesmas Kasihan 2 (Primary Healthcare Center). The dataset consists of 597 records, 6 features, and 8 labels. The details of the features and labels are presented in Table 1.

Table 1. Features and Labels for the Comorbidity Prediction Model

Features	Age, JK, TDS, TDD, IMT, Heart Rate
Labels	E11, E78, I11, N19, J00, M17, M79.1, M54.5

2.3. Data Preparation

The data underwent cleaning, where incomplete attributes (e.g., Fasting Blood Sugar with >50% missing values) were dropped, and minor missing values in BMI and Heart Rate were imputed using the median. The data was then split into an 80% training set and a 20% testing set. To address class imbalance across different comorbidity labels, the SMOTE technique was applied.

2.4. Modeling

The modeling stage is carried out by applying the Random Forest algorithm to predict the presence or absence of comorbidities in hypertensive patients. During this stage, the model parameters were configured, the model was trained using the training dataset, and parameter tuning was performed if necessary. The resulting model is subsequently tested using the testing dataset and evaluated based on performance metrics such as subset accuracy and Hamming loss.

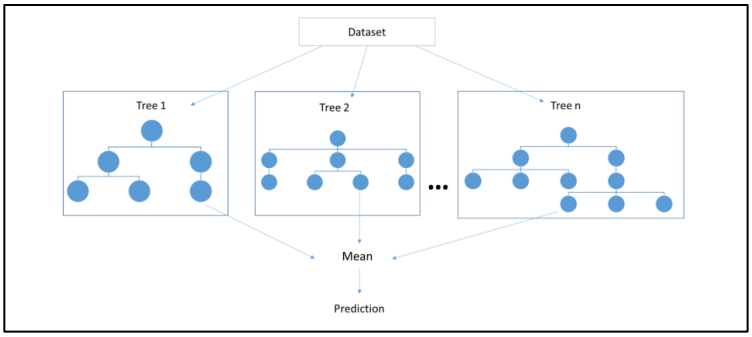

Figure 2. Random Forest Diagram

Random Forest is an ensemble method based on multiple decision trees. It is a machine learning technique widely utilized for classifying large datasets [13]. The objective of this method is to construct multiple decision trees and aggregate their predictions to map each observation into a predefined class [14].

2.5. Evaluation

The evaluation results are utilized to assess the feasibility of the model as a product developed within this R&D study. Specifically, the model's performance is analyzed using two primary evaluation metrics: subset accuracy and Hamming loss. These metrics are applied to determine whether the model is adequately robust for use as a predictive screening tool for hypertension comorbidities across the general population, or if it requires further refinement. A model deemed to exhibit good performance is then advanced to the next stage.

$$subsetacc(h) = \frac{1}{p}\sum_{i=1}^{p}[h(x_i) = Y_i] \qquad (1)$$

The formula above defines subset accuracy, which measures the proportion of instances where the predicted set of labels perfectly matches the true set of labels. The calculation evaluates the entire dataset by checking each patient's record individually. A prediction is deemed successful only if all predicted comorbidities for a specific patient are exactly identical to their actual diagnosed comorbidities; any partial mismatch results in a score of zero for that instance. In the context of multi-label classification, subset accuracy provides a strict evaluation metric, as it requires

the model to correctly identify the exact combination of a patient's comorbidities without any false positives or false negatives.

$$hamloss(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}|h(x_i) \triangle Y_i| \qquad (2)$$

The following equation represents Hamming loss, which evaluates the fraction of individual labels that are incorrectly predicted. Unlike subset accuracy, this metric accommodates partial matches. The formula calculates the error rate by examining the total number of possible comorbidity categories and identifying the mismatches situations where a comorbidity was predicted by the model but is actually absent, or present in the patient but not predicted. By averaging these individual errors across all labels and all evaluated patient records, Hamming loss provides a comprehensive measure of the model's overall misclassification rate, where a value closer to zero indicates superior predictive performance.

2.6. Deployment

At this stage, the model that successfully passed the evaluation process is deployed into a practical web application utilizing the Streamlit framework. Streamlit allows users to transform data scripts into interactive web applications, making information delivery more efficient and improving accessibility[15]. The system was designed with a user-friendly interface to facilitate seamless interaction. The platform allows healthcare professionals to input the clinical data of hypertensive patients and instantly obtain real-time comorbidity prediction results, making it an accessible and effective tool for early screening.

3. RESULTS AND DISCUSSION

3.1. Business Understanding

This study aims to address the limitations of manual identification of hypertension comorbidities at Puskesmas Kasihan 2 by developing a machine learning predictive model. Utilizing the Random Forest algorithm (multi-label classification) on medical record data, the model was evaluated using subset accuracy and hamming loss metrics. The developed model was subsequently integrated into a Streamlit -based web application prototype to serve as an early screening tool, enabling faster, more consistent early detection of comorbidities.

3.2. Data Understanding

The data for this study were collected from the medical records of Puskesmas Kasihan 2 between January and September 2025. The dataset combines two primary sources: demographic data exported to a Microsoft Excel file (comprising 22 patient history attributes) and clinical diagnostic data extracted directly from the health center's information system. Out of the 22 patient history attributes originally exported, only specific variables were selected for further processing, as detailed in Table 2.

Table 2. Attributes used in Excel files

No.	Attribute Name	Example Attribute Input	Description
1	ICDX	E11, I10 / I10, E11, E78	Used
2	Diagnosis	Non-insulin-dependent diabetes mellitus, Essential (primary) hypertension / Essential (primary) hypertension, Non-insulin-dependent diabetes mellitus, Pure hypercholesterolaemia	Used
3	Age	61 / 48	Used
4	Gender	F / M	Used

Table 2 illustrates the attributes utilized in the Excel file for the dataset. It details the specific variables, provides examples of the data entries for each attribute, and confirms whether these attributes are actively used ("Used") in the data processing or research.

3.3. Data Preparation

In the Data Preparation stage, the data analyzed in the previous phase is prepared to ensure its suitability for the modeling process. This stage encompasses data selection and integration, comorbidity label generation, data cleaning, data transformation, analyzing attribute relationships, splitting the dataset into training and testing sets, and handling imbalanced data.

a. Data Selection and Integration

The research data originated from two sources: patient demographic data (an Excel file) and clinical data from the health center's information system. Data collection began by identifying patients in the first source with a recorded diagnosis of hypertension (I10) and accompanying comorbidities. The identified patient data was then matched with the second source to obtain the corresponding clinical information. This matching process utilized patient identification and examination details from the original data (e.g., patient name, examination date, and age) to ensure that both sources represented the exact same patients. The matched results were subsequently merged into a single integrated dataset. After filtering hypertension (I10) cases, a modeling dataset consisting of 597 records was obtained.

b. Comorbidity Label Generation (Multi-label)

The target labels in this study were derived from the accompanying diagnoses found in the medical records of hypertensive patients. Since a single patient may have more than one comorbidity, the targets were formulated using a multi-label approach. The eight labels used are E11 (Type 2 diabetes mellitus), E78 (High cholesterol), I11 (Heart disease), N19 (Kidney failure), J00 (Acute nasopharyngitis), M17 (Gonarthrosis), M79.1 (Myalgia), and M54.5 (Low back pain). Each label is represented in a binary format, where 1 indicates that the patient has the comorbidity and 0 indicates otherwise.

c. Cleaning Data

After data selection and label generation, a data cleaning phase was conducted to enhance the overall quality of the dataset. The data cleaning process involved several steps, such as the elimination of irrelevant attributes, checking for duplicates, and the treatment of missing values.

d. Data Transformation

After the data was cleaned, it was transformed into a format suitable for modeling by encoding categorical attributes. For example, gender was numerically encoded, with 1 for male and 0 for female.

e. Attribute Relationships

Linear relationships among features in the dataset were analyzed to observe attribute correlations before the train-test split process. This step is intended to pinpoint attribute pairs with high correlations and to understand the underlying patterns among numerical features. For easier interpretation, the analysis results are visualized as a correlation matrix in the form of a heatmap, making the correlation values between pairs of attributes distinctly visible, as illustrated in Figure 3.

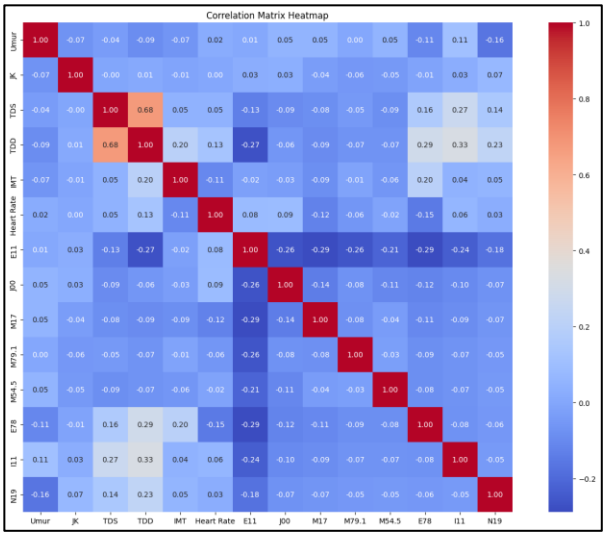

Figure 3. Correlation Matrix Heatmap

Based on the heatmap, the most notable linear relationship is the positive correlation between Systolic Blood Pressure (TDS) and Diastolic Blood Pressure (TDD) at 0.68. The majority of the other clinical and demographic attributes exhibit very weak correlations near zero. This indicates that most features are independent, which helps minimize the risk of multicollinearity in the predictive model.

f. Handling Imbalanced Data

This research used the SMOTE method because it effectively handles class imbalance and can help improve model performance [16]. To better understand the impact of SMOTE on the dataset, the number of samples in each

label was examined before and after its application. The following figures illustrate the distribution of labels and show how the class balance changed after the resampling process.

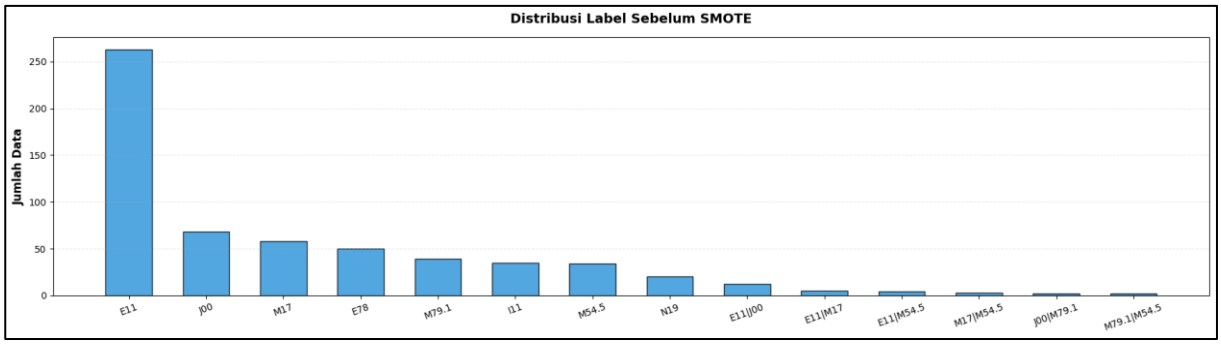

Figure 4. Distribution Data Label (Before SMOTE)

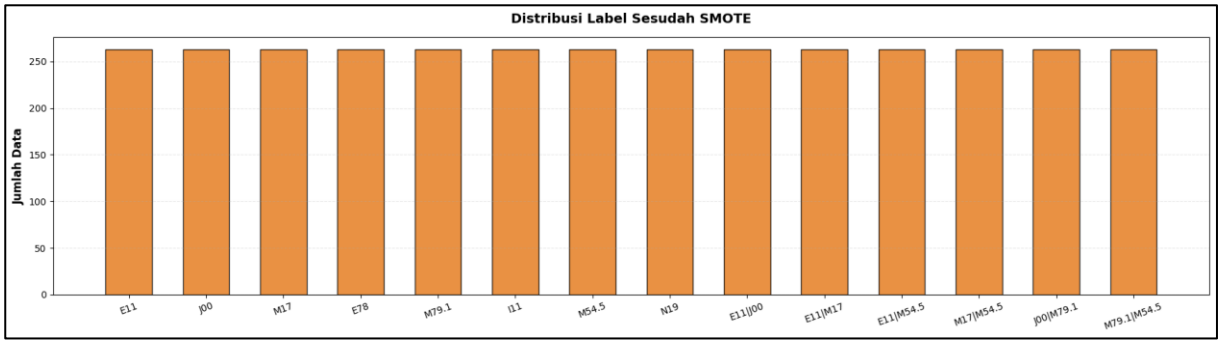

Figure 5. Distribution Data Label (After SMOTE)

Figures 4 and 5 show the label distribution before and after applying SMOTE. Before SMOTE, the dataset consisted of 597 patient records and several labels had noticeably fewer samples than others, resulting in an imbalanced distribution. After SMOTE was applied, the dataset size increased to 3,682 records because SMOTE was applied to generate synthetic samples for minority labels. the distribution of samples across labels became more balanced. This helped reduce class imbalance and provided a more balanced dataset for the model training process.

g. Train and Test Data Split

The finalized dataset was subsequently divided into training and testing sets to build and evaluate the predictive model. The data split was conducted using an 80:20 ratio, allocating 80% for the training data and 20% for the testing data. This partition resulted in 2945 records for the training set and 737 records for the testing set. The training data was utilized to train the Random Forest model, while the testing data was reserved to assess the model's performance using evaluation metrics.

3.4. Modelling

To predict the presence of comorbidities in hypertensive patients, this study utilized a Random Forest algorithm. Because a single patient may have multiple concurrent comorbidities, a multi-label classification approach was implemented using a MultiOutputClassifier. This approach simultaneously predicts each comorbidity label as a binary outcome (0/1) within a unified pipeline.

a. Data Preprocessing and Pipeline Construction

A machine learning pipeline was constructed to integrate data preprocessing and model training consistently. The preprocessing stage utilized a ColumnTransformer to select key clinical features: Age, Gender, Systolic Blood Pressure (TDS), Diastolic Blood Pressure (TDD), Body Mass Index (IMT), and Heart Rate. Missing values within these features were handled using a median-based SimpleImputer, while unused columns were discarded. This preprocessing step was then combined with the Random Forest classifier to form a unified modeling pipeline.

b. Hyperparameter Tuning and Custom Thresholding

To optimize performance and address class imbalance during training, the Random Forest model was configured with fixed hyperparameters, notably setting n_estimators. Furthermore, to accommodate the distinct distribution characteristics of each comorbidity, custom decision thresholds were established for each label. The model first

generates prediction probabilities for the positive class, which are subsequently converted into exact binary predictions (0 or 1) based on these label-specific thresholds.

c. Model Training

In the final step, the integrated pipeline comprising the preprocessing rules, optimized hyperparameters, and the multi-label classifier was trained using the training dataset (X_train, y_train). This finalized model was then prepared for the evaluation phase on the testing data.

3.5. Evaluation

The model evaluation was conducted using several quantitative metrics on the test dataset to measure its performance in the multi-label classification task. To provide a clearer overview of the results, the evaluation metrics are summarized in Table 3.

Table 3. Model Evaluation Results

Metric	Score
Subset Accuracy	0.8684
Hamming Loss	0.0227
Micro Precision	0.9032
Micro Recall	0.9749
Micro F1-Score	0.9377

As shown in Table 3, the model achieved a Subset Accuracy of 0.8684, indicating that most predictions matched the complete set of labels correctly. The Hamming Loss of 0.0227 suggests a low prediction error across labels, meaning that only a small portion of labels were incorrectly predicted. In addition, the model obtained a Micro Precision of 0.9032 and a Micro Recall of 0.9749, showing that it was able to identify relevant labels effectively while maintaining relatively accurate predictions. The Micro F1-Score of 0.9377 further indicates a good balance between precision and recall, reflecting strong overall performance in the multi-label classification task.

3.6. Deployment

The deployment phase aimed to implement the evaluated hypertension comorbidity prediction model into a practical application for healthcare professionals. A web-based application was developed using the Streamlit framework. The interface consists of two main sections: Patient Data Input and Prediction Results. Healthcare personnel enter patient identification information, including name and medical record number, along with clinical variables required by the model, namely age, gender, systolic blood pressure, diastolic blood pressure, BMI, and heart rate. Predictions are generated when users click the Predict button, while a Reset feature is provided to clear all input fields for subsequent use. After data submission, the system processes the input using the trained prediction model and displays the output as a list of predicted comorbidities with their corresponding probability values. These probabilities indicate the likelihood of specific comorbid conditions based on patient examination data and can support early screening and clinical decision-making. In addition, the application includes a Download PDF feature that generates a summary report containing patient information, examination data, and prediction results.

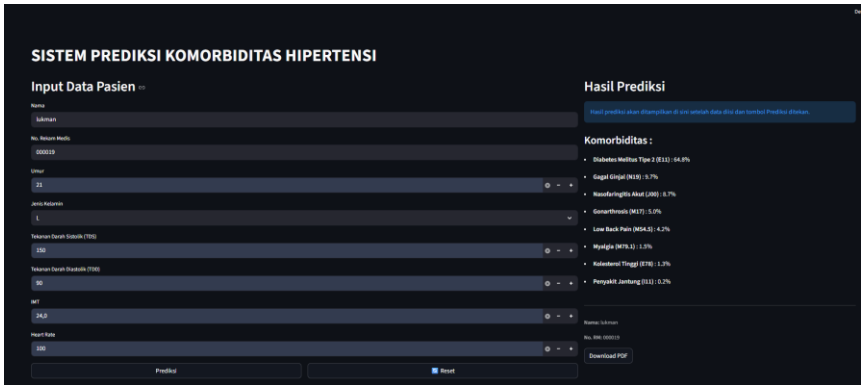

Figure 6. Website Interface After Prediction Execution

Figure 6 presents the interface of the developed hypertension comorbidity prediction system. The application is divided into two main sections: Patient Data Input on the left and Prediction Results on the right. The input section allows healthcare personnel to enter patient identification information, including name and medical record number, followed by clinical variables used as model inputs, namely age, gender, systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), and heart rate. The interface also provides Predict and Reset buttons

to facilitate prediction execution and rapid data re-entry. On the right side, the system displays prediction outputs in the form of comorbidity labels and their corresponding probability values, providing healthcare professionals with supporting information for early screening and clinical assessment. Additionally, a Download PDF feature is available to export the prediction results for documentation purposes.

3.7. Discussion

The results of this study demonstrate that the Random Forest algorithm combined with a multi-label classification approach is effective for predicting hypertension comorbidities using patient medical record data. The model achieved a Subset Accuracy of 0.8684 and a Hamming Loss of 0.0227, indicating that the majority of patient records were classified correctly with only a small proportion of label prediction errors. These findings suggest that the proposed model is capable of identifying multiple comorbidities simultaneously with relatively high reliability.

The high Micro Precision (0.9032), Micro Recall (0.9749), and Micro F1-Score (0.9377) further confirm the robustness of the developed model. The high recall value indicates that the model successfully identified most actual comorbidity cases, which is important in healthcare screening systems because missed predictions may lead to delayed treatment or complications. Meanwhile, the precision score demonstrates that most predicted comorbidities were relevant and accurate, minimizing false positive predictions. The balanced F1-Score shows that the model maintains good performance between sensitivity and prediction accuracy.

The use of the Random Forest algorithm proved suitable for this study because it can handle complex relationships among clinical variables and reduce overfitting through ensemble learning. In addition, the implementation of MultiOutputClassifier enabled the simultaneous prediction of multiple comorbidity labels within one integrated framework. This approach is highly relevant in hypertension cases, where patients frequently experience more than one comorbid condition at the same time. Compared with previous studies that focused only on hypertension classification or single-disease prediction, this research extends the application of machine learning toward multi-label comorbidity prediction, providing more comprehensive clinical insights.

The application of SMOTE also contributed significantly to model performance improvement. Before oversampling, several comorbidity labels had very limited sample sizes, which could bias the model toward majority classes. After applying SMOTE, the dataset became more balanced, allowing the model to learn minority label patterns more effectively. This is reflected in the low Hamming Loss and high Recall obtained during evaluation. However, the increase in synthetic data may also introduce the risk of overgeneralization, particularly for rare comorbidity combinations.

The correlation analysis revealed that most attributes had weak correlations, except for the moderate positive relationship between systolic and diastolic blood pressure. This indicates that the selected features provide relatively independent information for the prediction process, helping reduce multicollinearity issues within the model. Clinical variables such as blood pressure, BMI, heart rate, age, and gender were therefore considered appropriate predictors for identifying hypertension-related comorbidities.

The deployment of the model into a Streamlit-based web application demonstrates the practical applicability of this research. The developed system enables healthcare personnel to input patient examination data and obtain real-time comorbidity predictions quickly. This can support early screening, improve clinical decision-making, and reduce dependence on fully manual assessments. The inclusion of a PDF export feature further enhances documentation and reporting efficiency within healthcare services.

Despite the promising results, this study has several limitations. First, the dataset was collected from only one primary healthcare center, which may limit the generalizability of the model to other healthcare institutions or broader populations. Second, the number of records for certain comorbidity labels remained relatively limited, even after applying SMOTE. Third, the study only utilized a limited number of demographic and clinical features, while additional variables such as laboratory results, lifestyle factors, medication history, and family history could potentially improve predictive performance.

Future research is recommended to involve larger and more diverse datasets from multiple healthcare facilities to improve model generalization. Additional machine learning methods, such as XGBoost, LightGBM, or deep learning approaches, may also be explored and compared with Random Forest performance. Furthermore, integrating explainable artificial intelligence (XAI) techniques could help healthcare professionals better understand the reasoning behind prediction results, thereby increasing trust and interpretability in clinical implementation.

4. CONCLUSION

This study successfully developed a multi-label machine learning model for predicting hypertension comorbidities using the Random Forest algorithm. By utilizing patient demographic and clinical data from Puskesmas Kasihan 2, the model was able to predict multiple comorbid conditions simultaneously, including diabetes mellitus, high cholesterol, heart disease, kidney failure, and several musculoskeletal disorders. The evaluation results demonstrate that the proposed model achieved strong predictive performance, with a Subset Accuracy of 0.8684, Hamming Loss of 0.0227, Micro Precision of 0.9032, Micro Recall of 0.9749, and Micro F1-Score of 0.9377. These

findings indicate that the model can effectively identify comorbidity patterns in hypertensive patients while maintaining low prediction error rates. The implementation of SMOTE also contributed to improving model performance by addressing class imbalance issues within the dataset. Furthermore, the deployment of the model into a Streamlit-based web application demonstrates the practical applicability of this research in supporting healthcare services. The developed system can assist healthcare professionals in performing early screening and identifying potential comorbidities more efficiently and consistently using patient examination data. Overall, this research shows that the Random Forest multi-label classification approach has strong potential as a decision-support tool for hypertension comorbidity prediction in primary healthcare settings. Future studies are recommended to expand the dataset, incorporate additional clinical variables, and compare other machine learning algorithms to further improve predictive accuracy and model generalization.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Kasihan 2 Primary Healthcare Center (Puskesmas Kasihan 2), Yogyakarta, for providing access to the medical record data and supporting the implementation of this study

REFERENCES

[1] Kementerian Kesehatan Republik Indonesia, "Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/4634/2021 tentang Pedoman Nasional Pelayanan Kedokteran Tata Laksana Hipertensi Dewasa," *Keputusan Menteri Kesehatan Republik Indonesia*, no. HK.01.07/MENKES/4634/2021, pp. 1–85, May 2021.

[2] C. A. Dewati, A. R. Natavany, Z. M. Putri, A. Nurfaizi, S. O. Rumbrawer, and D. S. Sri Rejeki, "Literature Review: Faktor Risiko Hipertensi di Indonesia," *Jurnal Kesehatan Masyarakat*, vol. 11, no. 3, pp. 290–307, May 2023, doi: 10.14710/jkm.v11i3.34514.

[3] World Health Organization, "Hypertension."

[4] Dinas Kesehatan Daerah Istimewa Yogyakarta, "Profil Kesehatan Daerah Istimewa Yogyakarta Tahun 2023," p. 223, Jun. 2024.

[5] Dinas Kesehatan Daerah Istimewa Yogyakarta, "Laporan Kinerja Instansi Pemerintah," Yogyakarta, Feb. 2025.

[6] "Website Dinas Kesehatan Kota Yogyakarta." Accessed: Jan. 14, 2026. [Online]. Available: https://kesehatan.jogjakota.go.id/berita/id/619?

[7] S. A. Marendengi and S. A. Palloge, "Gambaran Pasien Hipertensi dengan Penyakit Komorbid di Puskesmas Layang Makassar pada Bulan Juli 2024," Jun. 2025. Accessed: Mar. 05, 2026. [Online]. Available: http://citracendekiacelebes.org/index.php/INAJOH

[8] M. M. Alsaleh *et al.*, "Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review," Jul. 01, 2023, *Elsevier Ireland Ltd*. doi: 10.1016/j.ijmedinf.2023.105088.

[9] T. Suprapti and S. Anwar, "Perbandingan Akurasi Algoritma Random Forest Dan Naïve Bayes Dalam Memprediksi Risiko Hipertensi," *BULLET : Jurnal Multidisiplin Ilmu*, no. 02, pp. 568–573, Apr. 2023, Accessed: Mar. 05, 2026. [Online]. Available: https://journal.mediapublikasi.id/index.php/bullet

[10] Novianti, S. Putri Agustini Alkadri, and I. Fakhruzi, "Klasifikasi Penyakit Hipertensi Menggunakan Metode Random Forest," *Progresif: Jurnal Ilmiah Komputer*, vol. 20, no. 1, pp. 380–392, Feb. 2024.

[11] S. Alia Azhaar, T. Al Mudzakir, H. Yulia Novita, and S. Faisal, "Implementasi Algoritma Support Vector Machine (SVM) dan Random Forest Untuk Klasifikasi Penyakit Hipertensi Berdasarkan Data Kesehatan," *Jurnal Riset Komputer)*, vol. 12, no. 4, pp. 2407–389, 2025, doi: 10.30865/jurikom.v12i4.8744.

[12] N. Hidayati, J. Suntoro, and G. G. Setiaji, "Perbandingan Algoritma Klasifikasi untuk Prediksi Cacat Software dengan Pendekatan CRISP-DM," *Jurnal Sains dan Informatika*, vol. 7, no. 2, pp. 117–126, Nov. 2021, doi: 10.34128/jsi.v7i2.313.

[13] R. F. N. Iskandar, D. H. Gutama, D. P. Wijaya, and D. Danianti, "Klasifikasi Menggunakan Metode Random Forest untuk Awal Deteksi Diabetes Melitus Tipe 2," *Jurnal Teknik Industri Terintegrasi*, vol. 7, no. 3, pp. 1620–1626, Jul. 2024, doi: 10.31004/jutin.v7i3.26916.

[14] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, Dec. 2022, doi: 10.35580/variansiunm31.

[15] B. Fathur Rochman and S. Harits Suryawan, "DASHBOARD OF PRODUCTION, UTILIZATION, FUEL AT PT CIPTA KRIDATAMA USING PYTHON AND STREAMLIT," *Pengabdian Kepada Masyarakat)*, vol. 1, no. 6, 2023.

[16] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 1033–1042, Oct. 2024, doi: 10.25126/jtiik.2024117989.