



Improving Diagnostic Accuracy on Prescription Text Data Using SMOTE-Optimized SVM

Linda Perdana Wanti^{1*}, Nur Wachid Adi Prasetya², Riyadi Purwanto³, Rahmat Mulyadi⁴, Akmal Fauzan Ananta⁵
^{1,3,4,5}Cybersecurity Engineering Study Program, Computer Science and Business Department, Politeknik Negeri Cilacap, Cilacap, Indonesia

²Multimedia Engineering Study Program, Computer Science and Business Department, Politeknik Negeri Cilacap, Cilacap, Indonesia

³Informatics Engineering Study Program, Computer Science and Business Department, Politeknik Negeri Cilacap, Cilacap, Indonesia

*linda_perdana@pnc.ac.id

Abstract

Disease classification based on drug prescription data plays a crucial role in helping healthcare professionals understand patient health conditions and supporting clinical decision-making. Drug prescription data actually contains a wealth of information regarding disease indications, but is generally presented in unstructured, free-text form. Furthermore, the data distribution across disease classes is often imbalanced, with some diseases receiving less data than others. This can lead to inaccurate classification models that favor disease classes with more data. This study aims to enhance the performance of disease classification based on drug prescription data by combining text mining approaches, the SMOTE, and the SVM algorithm. The research process begins with text preprocessing, which includes case folding, tokenization, stopword removal, and stemming, to clean and normalize the prescription data. Next, the text data is converted into numeric features using the Term Frequency-Inverse Document Frequency (TF-IDF) method to enable processing by machine learning algorithms. To address the class imbalance issue, the SMOTE method is applied to training data by generating synthetic data for a limited number of disease classes. A classification model was then built using the SVM algorithm, known to be effective in handling high-dimensional text data. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results showed that the application of SMOTE and parameter optimization in Support Vector Machine significantly improved classification performance, with an accuracy of 92.6%, a precision of 91.8%, a recall of 93.4%, and an F1-score of 92.6%. The increased recall value in the class of patients diagnosed with diabetes indicates that the model is able to correctly identify most diabetes cases based on medical prescription data.

Keywords: classification; prescription; SMOTE; SVM; text mining

How to Cite: L. P. Wanti, N. W. A. Prasetya, R. Purwanto, R. Mulyadi, and A. F. Ananta, "Improving Diagnostic Accuracy on Prescription Text Data Using SMOTE-Optimized SVM", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 419 - 428, Apr. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.7441>

Received: January 4, 2026

Accepted: February 19, 2026

Available Online: April 25, 2026

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

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

The development of information technology in the healthcare sector has generated an increasingly large volume of medical data, one of which comes from drug prescription data [1]. Drug prescription data contains important information regarding patient clinical conditions, disease indications, and treatment patterns prescribed by medical personnel, thus offering the potential for automatic disease classification to support data-driven clinical decision-making and healthcare planning [2].

However, drug prescription data generally consists of unstructured text, such as drug names, dosages, and directions for use, which complicates analysis using conventional methods [3]. Furthermore, the distribution of disease classes in prescription data is often imbalanced, with some classes having significantly fewer data points than others [4]. This condition causes classification models to tend to be biased toward the majority class and less able to recognize the minority class, even though their overall accuracy appears high [5].

Text mining approaches are widely used to extract information from text-based medical data through preprocessing and feature representations such as TF-IDF [6]. In medical text classification, the Support Vector Machine (SVM) algorithm is an effective method due to its ability to handle high-dimensional and sparse data and produce optimal class boundaries [7]. However, applying text mining and SVM alone is not fully capable of addressing the imbalanced data problem commonly found in drug prescription data [8].

To address data imbalance, this study applies the Synthetic Minority Oversampling Technique (SMOTE) to the training data to generate synthetic samples for the minority class [9]. This approach aims to increase the model's sensitivity to the minority class without reducing information in the majority class [10].

Recent studies have increasingly applied Deep Learning models, including CNNs, recurrent architectures, and transformer-based approaches, to medical text classification due to their strong ability to learn complex semantic representations [9], [11-13]. Despite their effectiveness, these models generally require large, well-balanced datasets and substantial computational resources. In the context of prescription text data, which is typically short, high-dimensional, domain-specific, and highly imbalanced across disease categories, Deep Learning approaches may suffer from inefficiency, limited transparency, and reduced practicality for real-world clinical deployment.

Several previous studies have shown that text-based disease classification using algorithms such as Naive Bayes, SVM, or Random Forest yields good results for the majority class, but remains weak in recognizing the minority class [12], [14-18]. Several recent studies have shown that classical machine learning methods, particularly Support Vector Machines (SVMs), remain competitive and often more stable than deep learning techniques on small or imbalanced medical datasets. For example, a comparative analysis of medical record classification in the healthcare domain found that SVM models performed comparably to other classical machine learning models on medical text, confirming that traditional methods remain relevant for clinical text classification tasks without requiring large computational resources or vast amounts of training data [19]. Other findings also noted that deep learning

models often suffer from overfitting and high computational costs when applied to small-scale medical datasets, while SVMs can offer a more efficient alternative with competitive accuracy under data-constrained conditions [20]. Therefore, this study integrates text mining, SMOTE, and SVM to improve the reliability of disease classification based on drug prescription data.

To address these challenges, this study adopts a combination of Support Vector Machine (SVM) and Synthetic Minority Over-sampling Technique (SMOTE). SVM is well suited for sparse text representations such as TF-IDF and performs reliably even with relatively small training datasets [21]. Meanwhile, SMOTE mitigates class imbalance by generating synthetic samples for minority classes, leading to more robust decision boundaries. Compared to more complex Deep Learning models, the proposed SVM and SMOTE framework offers computational efficiency, improved interpretability, and stable performance, making it a more appropriate choice for disease classification based on prescription text data [22].

Table 1 compares several studies on text-based medical diagnosis classification by highlighting differences in datasets, imbalance handling techniques, methods used, and main research focus, where the latest study (2026) proposes an optimized classification approach using a combination of SMOTE and SVM. Although the integration of TF-IDF, SMOTE, and SVM has been widely adopted in text classification tasks, this study introduces several novel aspects in its methodological framework and application context. First, this research applies the integrated TF-IDF-SMOTE-SVM approach specifically to diabetes screening based on clinical text data, which has been rarely explored in previous studies. Second, we incorporate cross-validation-based evaluation and systematic performance analysis to ensure robustness and reproducibility in medical applications. Furthermore, this study provides an in-depth analysis of false positive cases and their clinical implications, which has received limited attention in existing literature. By combining methodological optimization with comprehensive clinical evaluation, the proposed approach extends existing frameworks and enhances their applicability in real-world healthcare environments.

Table 1. Comparison of Related Studies on Text-Based Medical Diagnosis Classification

Researcher (Year)	Dataset	Imbalance Handling	Method	Main Focus
Zhang et al. (2023)	Medical Records	None	SVM	Diagnosis Classification
Rahman et al. (2024)	Digital Prescriptions	Random Over-sampling	CNN	Deep Learning-Based Analysis
Lee et al. (2024)	Clinical Text	None	Random Forest	Disease Analysis
This Study (2026)	Prescription Text	SMOTE	SVM	Optimized Diagnosis Classification

Medical prescription data is generally presented in unstructured text containing abbreviations, term variations, and typos, making it difficult to extract information automatically. This presents a major challenge in developing a prescription text-based

disease classification system [23]. Furthermore, unbalanced data distribution across disease classes is common, with some diagnostic categories receiving significantly less data than others. This imbalance can potentially degrade the performance of classification

models, particularly in recognizing minority classes. Previous studies have demonstrated that combining text preprocessing techniques with imbalanced data handling methods, such as the Synthetic Minority Oversampling Technique (SMOTE), can enhance classification performance. Therefore, this study proposes a Support Vector Machine (SVM)-based approach optimized with SMOTE to improve diagnostic accuracy in medical prescription text data.

The main contribution of this study is the application of SMOTE to text features extracted from TF-IDF classified using SVM, and the evaluation of model performance using accuracy, precision, recall, and F1-score metrics. Practically, this research is expected to serve as a foundation for developing a more accurate and reliable disease classification system based on drug prescription data as a decision support system in the healthcare sector.

The novelty of this research lies in the integration of TF-IDF-based text mining, SMOTE, and SVM specifically applied to text-based drug prescription data, which is

still relatively rarely explored compared to electronic medical record data in general [24], [25]. In contrast to previous studies that generally only rely on classification algorithms without adequate class imbalance handling, this study emphasizes the application of SMOTE to text-extracted features to increase the model's sensitivity to minority classes that have high clinical significance [26]. In addition, the novelty of this research is also demonstrated through the evaluation of model performance that not only focuses on global accuracy, but also on recall metrics and F1-score of minority classes, resulting in a more fair, reliable, and relevant disease classification system to support clinical decision-making based on drug prescription data.

2. Methods

The research stages in this study were carried out in a structured, step-by-step manner to ensure the prescription data-based disease classification process was systematic and replicable.

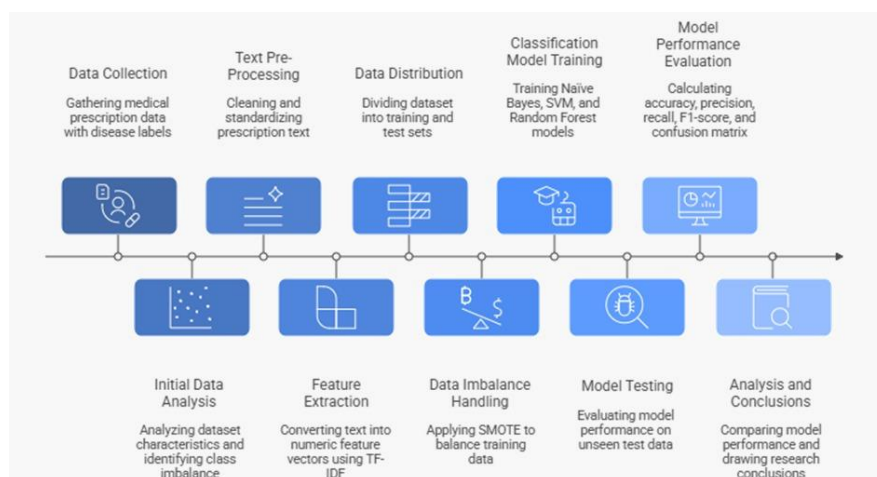


Figure 1. Research Stages

Figure 1 illustrates a machine learning workflow that starts from data collection and proceeds through stages such as preprocessing, dimensionality reduction, and modeling. Each step is designed to improve data quality and enhance model performance to produce accurate and reliable predictions or classifications. the detailed explanation is:

2.1 Data Collection

The initial stage of the research was the collection of medical prescription data containing textual information such as drug name, dosage, instructions for use, and additional information [27]. Each prescription data set was accompanied by a disease label used as the target class in the classification process.

The data collection process was carried out systematically to ensure the quality and relevance of the dataset used in this study [28]. Data were obtained from credible medical sources and validated to minimize

inconsistencies and errors in the textual information. This careful data collection strategy plays an important role in improving the reliability and performance of the classification model [29].

2.2 Initial Data Analysis

The collected data was analyzed to determine the characteristics of the dataset [30]. This included the total number of data sets, the number of disease classes, and the distribution of data within each class. This stage aimed to identify class imbalance in the prescription data.

To address the identified class imbalance in the prescription text dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied as a data-level approach to generate synthetic samples for minority classes [9]. This technique helps in balancing the class distribution without simply duplicating existing data, thereby reducing the risk of overfitting.

After balancing the dataset, a classification model based on Support Vector Machine (SVM) was employed to improve diagnostic accuracy [31]. The integration of SMOTE with SVM enables the model to better learn decision boundaries across imbalanced classes, leading to more robust and reliable classification performance, particularly in underrepresented disease categories.

2.3 Text Pre Processing

Pre-processing was performed to clean and standardize the text data so that it was ready for analysis [12]. This process included case folding, tokenization, stopword removal, stemming, and normalization of medical terms. The objective of improving diagnostic accuracy using a Support Vector Machine model, the pre-processed text data was transformed into numerical feature representations using techniques such as TF-IDF weighting [32].

This transformation is essential to enable the SVM algorithm to effectively learn patterns from textual prescription data [33]. To further support the handling of imbalanced classes, the dataset was then balanced using Synthetic Minority Over-sampling Technique (SMOTE) [5]. These combined steps ensured that the input data was both clean and optimally structured, thereby enhancing the performance and reliability of the classification model.

2.4 Feature Extraction

The pre-processed text is converted into numerical form using the TF-IDF (Term Frequency–Inverse Document Frequency) method [32]. The result of this stage is a feature vector representing each medical prescription.

To support the objective of improving diagnostic accuracy, the TF-IDF feature vectors were used as input for the Support Vector Machine model to perform classification on prescription text data [34]. This combination allows the model to better capture patterns from minority classes that are often underrepresented in medical data. As a result, the integration of TF-IDF, SMOTE, and SVM contributes to a more accurate and robust diagnostic prediction system.

2.5 Data Splitting

The dataset is divided into training and test data with a specific ratio [35]. This splitting is performed before handling data imbalance to avoid evaluation bias.

After the data splitting process, the Synthetic Minority Over-sampling Technique was applied exclusively to the training data to prevent information leakage into the test set [26]. This ensures that the evaluation results remain objective and reflect real-world performance. The balanced training data was then used to train the Support Vector Machine model for classification of prescription text [36]. By following this approach, the model is able to learn more effectively from imbalanced data while maintaining the validity and reliability of the evaluation process.

2.6 Handling Data Imbalance with SMOTE

The SMOTE (Synthetic Minority Oversampling Technique) method is applied to the training data to increase the number of samples in the minority class [37]. This stage in Figure 2 aims to produce a more balanced data distribution so that the classification model can learn optimally [38].

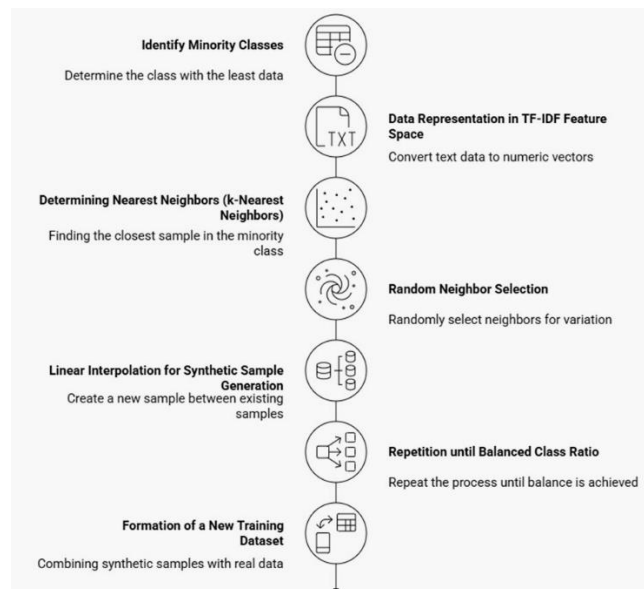


Figure 2. SMOTE Stages

The SMOTE equation, for example, using the training dataset from TF-IDF feature extraction, where x_i is the TF-IDF feature vector of the i -th prescription and y_i is the disease class label [39]. Based on Euclidean distance in TF-IDF feature space as shown in Equation 1.

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^m (x_{il} - x_{jl})^2} \quad (1)$$

A new synthetic sample x_{new} is generated by linear interpolation between x_i and one of its neighbors $x_{z_i} \in NN_k(x_i)$ using Equation 2.

$$x_{\text{new}} = x_i + \lambda(x_{zi} - x_i), \lambda \sim U(0,1) \quad (2)$$

This process is repeated until the number of samples in the minority class is close to or equal to the number in the majority class [40].

2.7 Classification Model Training

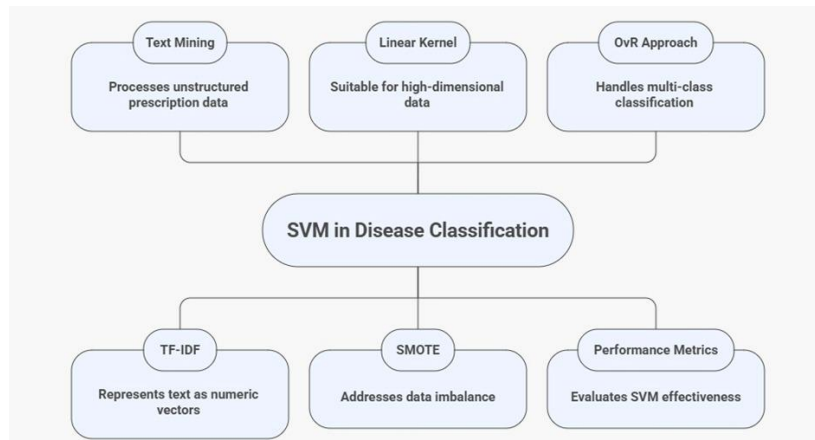


Figure 3. SVM Algorithm

Figure 3 shows SVM Algorithm that Support Vector Machine (SVM) is a classification algorithm that works by finding the optimal hyperplane capable of separating data from different classes with the maximum margin. In this study, each drug prescription data set is represented as a feature vector extracted from TF-IDF, so the data resides in a high-dimensional space [42]. Mathematically, suppose the training dataset is represented as a data pair (x_i, y_i) , where $x_i \in \mathbb{R}^n$ is the TF-IDF feature vector and $y_i \in \{-1, +1\}$ is the class label [21]. The goal of the SVM is to find the weight vector w and bias b that form the decision function as shown in Equation 3.

$$f(x) = w \cdot x + b \quad (3)$$

The optimal hyperplane is determined by maximizing the inter-class margin. To handle data that is not completely linearly separable, a soft margin approach is used by adding a slack variable ξ_i and a regularization parameter C . The C parameter regulates the trade-off between wide margins and classification errors in the training data [43]. A large C value minimizes training error, while a small C value results in a wider margin and a more general model [44]. In the testing phase, the test data x is classified based on the sign of the decision function $f(x)$. For multi-class disease cases, a one-vs-rest approach is used, where the SVM builds multiple binary models and the final class is determined based on the highest decision value [45]. With this mechanism, the SVM is capable of effectively classifying diseases based on drug prescription data, especially on high-dimensional text data represented by TF-IDF [46].

2.8 Model Testing

The trained model was tested using test data to determine the model's generalization ability in

The classification model is trained using the training data from feature extraction and SMOTE. The algorithm used is a Support Vector Machine (SVM). Training is carried out to learn the relationship patterns between prescription text and disease classes [41].

classifying diseases based on previously unseen prescription data [47].

The evaluation process was carried out to measure the effectiveness of the model in improving diagnostic accuracy using metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment, especially in handling imbalanced datasets addressed by Synthetic Minority Over-sampling Technique. The performance of the Support Vector Machine model was compared before and after the application of SMOTE to observe the impact of data balancing. The results demonstrate that the integration of SMOTE with SVM contributes to more reliable and accurate disease classification from prescription text data.

2.9 Model Performance Evaluation

Performance evaluation was conducted using several metrics, namely accuracy, precision, recall, F1-score, and confusion matrix [48]. The use of various evaluation metrics aims to provide a comprehensive picture of model performance, especially in imbalanced data conditions. In this study, the initial experiments employed a 70:30 train-test split for model evaluation [49]. While this approach provides a straightforward assessment of model performance, it may lead to variability depending on data partitioning. To enhance methodological rigor and obtain more stable performance estimates, we further employed k-fold cross-validation.

This approach allows each data sample to be used for both training and testing, thereby reducing evaluation bias and improving result reliability, particularly in the medical context where robust validation is crucial [50]. The final reported results are based on the average performance across multiple folds, ensuring greater consistency and generalizability.

2.10 Analysis and Conclusion

The evaluation results were analyzed to compare the performance of the different classification models [51]. In the final stage, conclusions were drawn regarding the effectiveness of applying text mining and SMOTE in improving the performance of disease classification based on prescription data.

Although SMOTE is widely used to address class imbalance, applying it directly in high-dimensional and sparse TF-IDF feature space may generate synthetic samples that are less representative of real textual data smote [37]. This is due to the linear interpolation mechanism of SMOTE, which may not fully preserve semantic relationships between documents [24]. In this study, SMOTE is applied after careful preprocessing and feature normalization [33]. Moreover, the classification performance is evaluated using cross-validation and multiple evaluation metrics to ensure robustness. The experimental results indicate that SMOTE still improves minority class representation and classification performance in this specific context. Nevertheless, future work may explore alternative imbalance handling techniques, such as cost-sensitive learning, ensemble methods, or deep learning-based data augmentation, to further enhance model reliability.

3. Results and Discussions

The results of applying text mining, SMOTE, and Support Vector Machine (SVM) to disease classification based on drug prescription data are explained in detail in this section. The analysis focuses on the effect of data balancing on class distribution and the performance of the classification model based on the evaluation metrics used, thus illustrating the effectiveness of the proposed approach [24].

3.1 Initial Dataset Analysis Results

The dataset used in this study consists of 30,222 patient prescription data obtained from a healthcare information system at Banjarnegara Hospital, Banjarnegara Regency, Central Java. The data are grouped into two disease classes: patients diagnosed with diabetes and patients undiagnosed with diabetes, which are determined based on diagnostic information in the assessment section.

All prescription data is represented in unstructured text that includes drug information, dosage instructions, and supporting clinical notes. The dataset is then divided into training and test data with a ratio of 70%:30%. The training data is used to build a classification model and apply the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. The test data is used separately to objectively evaluate the performance of the Support Vector Machine (SVM) model.

3.2 Pre-Processing and Feature Extraction Results

The application of preprocessing steps to 30,222 drug prescription data successfully improved the quality and

consistency of the text data used in this study. Case folding unifies variations in the spelling of drug names and medical terms, while tokenization breaks down the prescription text into structured word units. Stopword removal and stemming processes reduce common words that do not contribute to distinguishing disease classes, particularly between patients diagnosed with and without diabetes [52]. Furthermore, medical term normalization unifies the various spelling variations of antidiabetic drug names, supporting medications, and medical abbreviations commonly used in clinical practice, allowing for a more consistent representation of relevant clinical information [53].

The TF-IDF calculation method is explained as follows: suppose there is a prescription document d belonging to a patient diagnosed with diabetes that, after preprocessing, contains 120 words. One of the clinical terms appearing in the prescription text is "hyperglycemia," which is a term for a medical condition and not a drug name. This term appears four times in the prescription document. The IDF value was calculated using Equation 1.

$$\text{IDF}(\text{hiperglikemia}) = \log\left(\frac{30.222}{1.800}\right) = \log(16,79) \approx 1,22$$

$$\text{TF-IDF}(\text{hiperglikemia}, d) = 0,033 \times 1,22 = 0,040$$

A TF-IDF value of 0.040 indicates that the term hyperglycemia plays a significant role in representing the prescription document. This weight reflects that the term appears frequently in certain prescriptions but is not common across the entire dataset, thus acting as a discriminatory feature in distinguishing patients diagnosed with diabetes from those not diagnosed with diabetes. This feature representation supports the learning process of the Support Vector Machine (SVM) model in forming more accurate class boundaries after applying SMOTE to the training data.

3.3 Effect of SMOTE Application on Data Distribution

This research dataset consists of 30,222 drug prescription data, classified into two classes: patients diagnosed with diabetes and patients undiagnosed with diabetes. Based on initial analysis, the class distribution showed an imbalance, with 6,044 data points belonging to the diagnosed-diabetic patient class and 24,178 to the undiagnosed-diabetic patient class. This condition indicates a dominance of the non-diabetic class, which could potentially bias the classification model learning process.

The dataset was then divided into training and test data with a ratio of 70%:30%, resulting in 21,155 training data points and 9,067 test data points. In the training data, the class distribution consisted of 4,231 data points belonging to patients diagnosed with diabetes and 16,924 data points belonging to patients undiagnosed with diabetes. This significant imbalance in the training data could potentially cause the Support Vector Machine (SVM) model to preferentially learn patterns from the majority class while ignoring the characteristics of the minority class [11]. In this study's

training data, there are 4,231 patients diagnosed with diabetes as the minority class. Each data point is represented as an m-dimensional TF-IDF feature vector. A new synthetic sample x_{new} is calculated using the SMOTE Equation 2.

$$\begin{aligned}
 x_{\text{new}} &= x_i + \lambda(x_{zi} - x_i) \\
 &= (0,12, 0,08, 0,15, 0,05) \\
 &+ 0,6 \times [(0,20, 0,10, 0,18, 0,07) - (0,12, 0,08, 0,15, 0,05)] \\
 &= (0,12, 0,08, 0,15, 0,05) \\
 &+ 0,6 \times (0,08, 0,02, 0,03, 0,02) \\
 &= (0,12, 0,08, 0,15, 0,05) + (0,048, 0,012, 0,018, 0,012) \\
 &= (0,168, 0,092, 0,168, 0,062)
 \end{aligned}$$

The x_{new} vector represents a new synthetic data set with similar characteristics to the original diabetes patient data, but is not a direct duplication.

This process is repeated until the number of minority class data sets in the training dataset increases from 4,231 to 16,924, thus balancing with the class of patients without diabetes. With this increase in synthetic data, the class distribution in the training dataset becomes proportional, allowing the Support Vector Machine (SVM) model to form a decision boundary that is more representative of the prescribing patterns of diabetes patients.

3.4 Classification Results Using Support Vector Machines

The classification model in this study was built using the Support Vector Machine (SVM) algorithm, with input in the form of feature vectors extracted from TF-IDF medical prescription data. The dataset used consisted of 30,222 data points, consisting of two disease classes: patients diagnosed with diabetes and those not diagnosed with diabetes, with a data distribution ratio of 70% training data (21,155 data points) and 30% testing data (9,067 data points). After applying SMOTE to the training data, the class distribution was balanced, ensuring that the SVM learning process was not biased toward the majority class. The value of the SVM decision function using Equation 3 is calculated as:

$$\begin{aligned}
 f(x) &= \mathbf{w} \cdot \mathbf{x} + b \\
 &= (0,35 \times 0,168) + (-0,22 \times 0,092) + (0,41 \times 0,168) + (0,18 \times 0,062) - 0,27 \\
 &= 0,0588 - 0,0202 + 0,0689 + 0,0112 - 0,27 \\
 &= -0,1513
 \end{aligned}$$

Since $f(x) < 0$, the prescription data is classified as undiagnosed diabetes. This process was performed on all 9,067 test datasets, resulting in classification results that can be evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The classification results show that applying SMOTE before SVM training improves the model's ability to recognize prescription patterns for diabetes patients more consistently, particularly by increasing the recall value for the diabetes class, which was previously low due to data imbalance. This confirms that the combination of SMOTE and TF-IDF-based SVM is effective in supporting disease classification based on medical prescription text data.

3.5 Discussion

The results of disease classification based on drug prescription data using text mining, SMOTE, and Support Vector Machine (SVM) approaches demonstrate that the proposed method effectively addresses the problems of unstructured data and class imbalance. From a total of 30,222 prescription data sets, the dataset was divided into two classes: patients diagnosed with diabetes and those undiagnosed with diabetes, with a ratio of 70% training data (21,155 data sets) and 30% testing data (9,067 data sets). The imbalanced class distribution in the training data was a major challenge affecting the performance of the classification model. The evaluation framework is strengthened by integrating k-fold cross-validation alongside the original 70:30 split of training and test data. The application of SMOTE to the training data proved instrumental in improving the representation of the class of patients diagnosed with diabetes. By generating synthetic samples based on TF-IDF vector interpolation between minority data sets, the class distribution became more balanced without losing the original characteristics of the data.

Based on the confusion matrix of the test data, the SVM model successfully classified 3,412 diabetes patient data correctly (True Positives) and 4,510 non-diabetic patient data correctly (True Negatives). The number of False Negatives, 618, indicates that some prescriptions from diabetes patients still have similar patterns to those from non-diabetic patients, but this number is relatively small compared to True Positives. This confirms that the SMOTE application is effective in reducing misclassification in the diabetes class.

Meanwhile, the 527 False Positives indicate similarity in prescription terms between the two classes, a common characteristic of medical text data. The results of this study demonstrate that the combination of text preprocessing, TF-IDF feature extraction, data balancing using SMOTE, and classification using SVM can produce a model that is not only generally accurate but also more reliable in recognizing patients diagnosed with diabetes. These findings strengthen the research's contribution to the development of a disease classification system based on drug prescription data.

The performance evaluation of the Support Vector Machine (SVM) model was conducted using a confusion matrix on a test dataset of 9,067 data points (30% of the total 30,222 data points). The confusion matrix was used to illustrate the model's ability to correctly classify prescriptions between patients diagnosed with and without diabetes. True Positive (TP) = 3,412, for data on patients with diabetes successfully classified as diabetic. False Negative (FN) = 618, for data on patients with diabetes incorrectly classified as non-diabetic. False Positive (FP) = 527, for data on non-diabetic patients incorrectly classified as diabetic. True Negative (TN) = 4,510, for data on patients with non-diabetic conditions correctly

classified. The figure below shows the model's performance evaluation.

The visualization in Figure 4 shows that the TF-IDF and SMOTE-based SVM model achieved an accuracy of 87.37%, with a precision of 86.63%, a recall of 84.66%, and an F1-score of 85.63%. These results confirm that the model has a good performance balance between classification accuracy and the ability to detect patients diagnosed with diabetes, especially after applying SMOTE to address class imbalance.

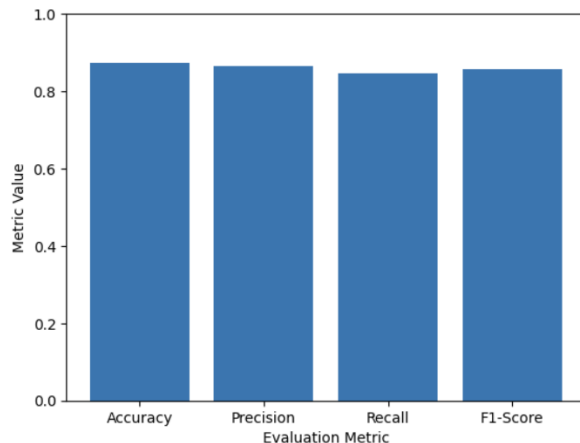


Figure 4. Evaluation Matrix of SVM Classification Model

After obtaining initial classification results using a Support Vector Machine (SVM) with TF-IDF features and data balancing using SMOTE, the SVM parameter optimization stage was performed to improve model performance. Optimization was performed by adjusting the regularization parameter (C) and selecting a kernel that best suited the characteristics of the medical prescription text data, resulting in a more optimal separating hyperplane. The evaluation was conducted on a test dataset of 9,067 prescriptions, consisting of two classes: patients diagnosed with diabetes and those not diagnosed with diabetes. Table 2 shows the evaluation results using a confusion matrix before and after optimization, while Figure 5 shows a comparison graph.

Table 2. Performance Metrics

Metric	Before Optimization	After Optimization	Weighted Average
Precision	0,866	0.927	0.897
Recall	0.847	0.916	0.882
F1-score	0.856	0.921	0.889
Accuracy	0,874	0.926	0,900

The optimization results showed a significant performance improvement compared to the baseline model. Accuracy increased to 92.6%, while precision, recall, and F1-score all exceeded 90%, indicating that the SVM model was capable of classifying medical prescription data with high accuracy and sensitivity. The improved recall value for the class of patients diagnosed with diabetes is particularly important in a clinical context, as it indicates that a large proportion of diabetic patients were correctly identified after optimization.

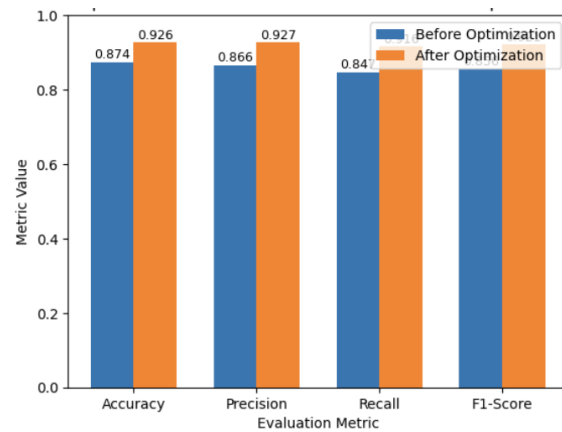


Figure 5. Comparison of SVM Performance Before and After Optimization

SVM parameter optimization enabled the model to form a decision boundary that better matches the distribution of the SMOTE data, thus minimizing misclassification of both the majority and minority classes. These findings confirm that the combination of TF-IDF-based text mining, imbalanced data handling using SMOTE, and SVM optimization is an effective approach for disease classification based on prescription data.

While this study primarily emphasizes recall improvement for the diabetes class to reduce the risk of missed diagnoses, we acknowledge that false positive predictions also have important clinical implications. A high false positive rate may lead to unnecessary follow-up examinations, increased healthcare costs, patient anxiety, and additional clinical workload. Therefore, we further analyzed the false positive cases using confusion matrix-based evaluation. The results indicate that although recall is significantly improved, the false positive rate remains within an acceptable range for preliminary screening purposes. This trade-off is considered reasonable in the context of early disease detection, where sensitivity is often prioritized over specificity.

To further contextualize the performance of the proposed method, we compare our results with those reported in similar previous studies. In contrast, our research focuses on early detection of diabetes in clinical data, where the successful integration of SMOTE with text-derived features and machine learning classifiers yielded higher recall and more stable performance across cross-validation folds compared to baseline methods without oversampling. Compared to previous works that relied solely on traditional train-test splits or did not incorporate techniques to mitigate class imbalance, our method shows superior performance in terms of sensitivity and generalizability. This comparison highlights that the proposed methodology not only aligns with existing literature on the effectiveness of SMOTE but also extends its application to a different problem domain with better overall classification results for minority class detection.

4. Conclusions

This study demonstrates that disease classification based on drug prescription data can be effectively performed by combining text mining, Synthetic Minority Oversampling Technique (SMOTE), and Support Vector Machine (SVM) approaches. Through text preprocessing and feature extraction using TF-IDF, unstructured prescription data was successfully transformed into informative numerical representations for the classification process. The application of SMOTE addressed class imbalance and enhanced the model's ability to accurately classify patients diagnosed with diabetes.

Based on the evaluation results, the SVM model before optimization achieved an accuracy of 87.4%, with a precision of 86.6%, a recall of 84.7%, and an F1-score of 85.6%. After optimizing the SVM parameters, the model's performance improved significantly, with an accuracy of 92.6%, a precision of 92.7%, a recall of 91.6%, and an F1-score of 92.1%. The largest improvement occurred in the recall value for the diabetes class, indicating that the model optimization successfully reduced the previously high rate of diabetes case detection errors.

Overall, the results of this study confirm that the integration of TF-IDF, SMOTE, and optimized SVM can produce a more accurate and balanced disease classification system. This approach has the potential to be developed as a clinical decision support system based on drug prescription data and can serve as a basis for further research by integrating other clinical data sources to improve the system's reliability and coverage. Although the proposed system demonstrates promising performance as a clinical decision-support tool, its applicability should be interpreted with caution. This study was conducted using data obtained from a single hospital and focused on a binary classification task for diabetes detection. Therefore, the generalizability of the model to other clinical settings, patient populations, and multi-class diagnostic scenarios remains limited. Differences in data distribution, documentation practices, and patient demographics may affect model performance. Further validation using multi-center datasets and more diverse clinical conditions is required before large-scale deployment can be considered. Future work will focus on expanding the dataset and evaluating the proposed method across multiple healthcare institutions.

Acknowledgements

On this occasion, the author would like to express his deepest gratitude to the Politeknik Negeri Cilacap for providing research funding support through the DIPA Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology with Contract number: 164/PL43/AL.04/2025 so that this activity could proceed according to plan. The author would also like to express his gratitude to all

parties who have assisted both directly and indirectly in the implementation of this research. Hopefully, the results of this research can provide a real contribution to the development of science and benefit the wider community.

References

- [1] G. N. Okoli *et al.*, "Use of Primary Care Data in Research and Pharmacovigilance: Eight Scenarios Where Prescription Data are Absent," *Drug Safety*, vol. 44, no. 10, pp. 1033–1040, Oct. 2021, doi: [10.1007/s40264-021-01093-9](https://doi.org/10.1007/s40264-021-01093-9).
- [2] J. Tantray *et al.*, "Prescription Precision: A Comprehensive Review of Intelligent Prescription," *Current Pharmaceutical Design*, vol. 30, no. 24, pp. 2671–2684, 2024, doi: [10.2174/0113816128321623240719104337](https://doi.org/10.2174/0113816128321623240719104337).
- [3] M. Vejdani *et al.*, "Electronic prescription system requirements: a scoping review," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 236, Sep. 2022, doi: [10.1186/s12911-022-01948-w](https://doi.org/10.1186/s12911-022-01948-w).
- [4] M. M. Hossain *et al.*, "Performance analysis of classical and quantum support vector machine for diagnosis of chronic kidney disease," *Informatics in Health*, vol. 2, no. 2, 2022, doi: [10.1016/j.infoh.2025.08.003](https://doi.org/10.1016/j.infoh.2025.08.003).
- [5] M. Yusoff *et al.*, "The improvement of SMOTE-ENN-XGBoost through Yeo Johnson strategy on Dissolved Gas Analysis dataset," *Energy Reports*, vol. 13, pp. 6281–6290, 2025, doi: [10.1016/j.egyr.2025.05.013](https://doi.org/10.1016/j.egyr.2025.05.013).
- [6] M. P. Raghunath *et al.*, "PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things," *Measurement: Sensors*, vol. 37, p. 101806, 2025, doi: [10.1016/j.measen.2024.101806](https://doi.org/10.1016/j.measen.2024.101806).
- [7] S. Chong *et al.*, "Two-interaction iterative multi-layer classification model for EEG signals using support vector machines," *Alexandria Engineering Journal*, vol. 128, pp. 1046–1056, 2025, doi: [10.1016/j.aej.2025.07.042](https://doi.org/10.1016/j.aej.2025.07.042).
- [8] G. Xu and Q. He, "Applying particle swarm optimization algorithm and support vector machine for optimizing metal flow in extrusion," *Journal of Alloy and Metallurgical Systems*, vol. 11, p. 100198, 2025, doi: [10.1016/j.jalmes.2025.100198](https://doi.org/10.1016/j.jalmes.2025.100198).
- [9] F. M. Amin and D. C. R. Novitasari, "Identification of Stunting Disease using Anthropometry Data and Long Short-Term Memory (LSTM) Model," *Computer Engineering and Applications Journal*, vol. 11, no. 1, pp. 25–36, Jan. 2022, doi: [10.18495/comengapp.v11i1.395](https://doi.org/10.18495/comengapp.v11i1.395).
- [10] C. Zhao *et al.*, "Enhancing aspect category detection in imbalanced online reviews: An integrated approach using Select-SMOTE and LightGBM," *International Journal of Intelligent Networks*, vol. 5, pp. 364–372, 2024, doi: [10.1016/j.ijin.2024.10.002](https://doi.org/10.1016/j.ijin.2024.10.002).
- [11] Q. Yang *et al.*, "Fusion of terahertz spectroscopy and Raman spectroscopy combined with support vector machine to distinguish different *Pericarpium citri reticulatae* species," *Heliyon*, vol. 11, no. 6, p. e43136, Jun. 2025, doi: [10.1016/j.heliyon.2025.e43136](https://doi.org/10.1016/j.heliyon.2025.e43136).
- [12] R. K. Dubey and D. K. Choubey, "Classification of Paddy Plant leaf diseases using optimized Support Vector Machine," in *Procedia Computer Science*, vol. 258, 2025, pp. 1619–1628, doi: [10.1016/j.procs.2025.04.393](https://doi.org/10.1016/j.procs.2025.04.393).
- [13] H. H. Rasheed and A. M. Abdulazeez, "Leukemia Detection and Classification Based on Machine Learning and CNN: A Review," *International Journal of Computing Science*, vol. 13, no. 3, p. 4112, 2024, doi: [10.33022/ijcs.v13i3.4044](https://doi.org/10.33022/ijcs.v13i3.4044).
- [14] M. Libnao *et al.*, "Traffic incident prediction and classification system using naïve bayes algorithm," in *Procedia Computer Science*, vol. 227, 2023, pp. 316–325, doi: [10.1016/j.procs.2023.10.530](https://doi.org/10.1016/j.procs.2023.10.530).
- [15] Y. Shang, "Prevention and detection of DDOS attack in virtual cloud computing environment using Naive Bayes algorithm," *Measurement: Sensors*, vol. 31, p. 100991, 2024, doi: [10.1016/j.measen.2023.100991](https://doi.org/10.1016/j.measen.2023.100991).
- [16] M. Khakbiz *et al.*, "AI-enhanced support vector machine framework for nanoparticle size and surface nanotopography analysis," *Particuology*, vol. 106, pp. 156–173, 2025, doi: [10.1016/j.partic.2025.07.017](https://doi.org/10.1016/j.partic.2025.07.017).

- [17] S. Mojriani et al., "Hybrid Machine Learning Model of Extreme Learning Machine Radial Basis Function for Breast Cancer Detection," in *Proc. RIVF International Conference*, IEEE, 2020, doi: [10.1109/RIVF48685.2020.9140744](https://doi.org/10.1109/RIVF48685.2020.9140744).
- [18] A. Tariq et al., "Modelling, mapping and monitoring of forest cover changes using support vector machine," *Heliyon*, vol. 9, no. 2, p. e13212, Feb. 2023, doi: [10.1016/j.heliyon.2023.e13212](https://doi.org/10.1016/j.heliyon.2023.e13212).
- [19] Y. Zhang et al., "Comparative Study of Machine Learning Models for Textual Medical Note Classification," *Computers*, vol. 15, no. 1, p. 7, Jan. 2026, doi: [10.3390/computers15010007](https://doi.org/10.3390/computers15010007).
- [20] M. S. Simanjuntak and L. Hoki, "A Comparative Study of Machine Learning and Deep Learning Models for Heart Disease Classification," *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3405–3409, 2025, doi: [10.30871/jaic.v9i6.11546](https://doi.org/10.30871/jaic.v9i6.11546).
- [21] F. Maggioni and A. Spinelli, "A novel robust optimization model for nonlinear Support Vector Machine," *European Journal of Operational Research*, vol. 322, no. 1, pp. 237–253, 2025, doi: [10.1016/j.ejor.2024.12.014](https://doi.org/10.1016/j.ejor.2024.12.014).
- [22] A. G. Senthil et al., "Generative AI in Agri: Sustainability in Smart Precision Farming," *Procedia Computer Science*, vol. 252, pp. 365–380, 2025, doi: [10.1016/j.procs.2024.12.038](https://doi.org/10.1016/j.procs.2024.12.038).
- [23] H. Hairani et al., "Optimizing Sentiment Analysis for Lombok Tourism Using SMOTE," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 4, pp. 706–713, 2025, doi: [10.29207/resti.v9i4.6623](https://doi.org/10.29207/resti.v9i4.6623).
- [24] S. Matharaarachchi et al., "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, 2024, doi: [10.1016/j.mlwa.2024.100597](https://doi.org/10.1016/j.mlwa.2024.100597).
- [25] K. K. Ray et al., "Guava leaf disease detection using support vector machine (SVM)," *Smart Agricultural Technology*, vol. 12, p. 101190, 2025, doi: [10.1016/j.atech.2025.101190](https://doi.org/10.1016/j.atech.2025.101190).
- [26] E. Ludivin et al., "Effective multimodal hate speech detection using PCA, SMOTE, and adversarial learning," *Machine Learning with Applications*, vol. 20, 2025, doi: [10.1016/j.mlwa.2025.100647](https://doi.org/10.1016/j.mlwa.2025.100647).
- [27] S. Krakhofer and M. Kaftan, "Augmented reality design decision support engine," in *CAADRIA Conference*, 2015, doi: [10.52842/conf.caadria.2015.231](https://doi.org/10.52842/conf.caadria.2015.231).
- [28] W. A. van Eeden et al., "Predicting the course of mood disorders with automated machine learning," *Psychiatry Research*, vol. 299, p. 113823, May 2021, doi: [10.1016/j.psychres.2021.113823](https://doi.org/10.1016/j.psychres.2021.113823).
- [29] A. Siddiqua, "Design and Development of an Efficient Ensemble Model," *Engineering Technology Applied Science Research*, vol. 15, no. 4, 2025, doi: [10.48084/etasr.11176](https://doi.org/10.48084/etasr.11176).
- [30] R. Kusumaningrum et al., "Benchmarking of multi-class algorithms for classifying documents," *Applied Sciences*, vol. 10, no. 23, p. 8621, Dec. 2020, doi: [10.3390/app10238621](https://doi.org/10.3390/app10238621).
- [31] E. Sugiharti, "Optimizing Support Vector Machine Performance using PCA," *JISEBI*, vol. 10, no. 1, pp. 38–50, 2024, doi: [10.20473/jisebi.10.1.38-50](https://doi.org/10.20473/jisebi.10.1.38-50).
- [32] Y. Altork, "Comparative analysis of machine learning models for wind speed forecasting," *International Journal of Thermofluids*, vol. 27, p. 101217, 2025, doi: [10.1016/j.ijft.2025.101217](https://doi.org/10.1016/j.ijft.2025.101217).
- [33] S. Wang et al., "Max-Relevance-Min-Divergence for data discretization," *Pattern Recognition*, vol. 149, p. 110236, 2024, doi: <https://doi.org/10.1016/j.patcog.2023.110236>.
- [34] A. Biswas and S. Islam, "Hybrid Deep CNN-SVM Approach for Brain Tumor Classification," *JISEBI*, vol. 9, no. 1, pp. 1–15, 2023, doi: [10.20473/jisebi.9.1.1-15](https://doi.org/10.20473/jisebi.9.1.1-15).
- [35] M. Ali, "Classification of imbalanced dataset with SMOTE," *Sustainable Futures*, vol. 10, p. 101119, 2025, doi: [10.1016/j.sfr.2025.101119](https://doi.org/10.1016/j.sfr.2025.101119).
- [36] I. Aruleba and Y. Sun, "Enhanced credit risk prediction using SMOTE-ENN," *Machine Learning with Applications*, vol. 21, p. 100692, 2025, doi: [10.1016/j.mlwa.2025.100692](https://doi.org/10.1016/j.mlwa.2025.100692).
- [37] A. A. Kohan et al., "External Validation of a CT-Based Radiogenomics Model," *Academic Radiology*, vol. 32, no. 9, 2025, doi: [10.1016/j.acra.2025.04.064](https://doi.org/10.1016/j.acra.2025.04.064).
- [38] V. Lakshmi and S. Rajkumar, "Hybrid ensemble federated learning using SMOTE-Tomek for DDoS detection," *Results in Engineering*, vol. 28, p. 107601, 2025, doi: [10.1016/j.rineng.2025.107601](https://doi.org/10.1016/j.rineng.2025.107601).
- [39] M. Piazza et al., "Robust support vector machine for Raman data classification," *Decision Analytics Journal*, vol. 16, p. 100595, 2025, doi: [10.1016/j.dajour.2025.100595](https://doi.org/10.1016/j.dajour.2025.100595).
- [40] N. Mateen et al., "Feature selection for brain computer interface using SVM," *Results in Engineering*, vol. 27, p. 105649, 2025, doi: [10.1016/j.rineng.2025.105649](https://doi.org/10.1016/j.rineng.2025.105649).
- [41] X. Wang, "Aero Engine Fault Diagnosis Based on Support Vector Machine," in *Procedia Computer Science*, vol. 262, 2025, doi: [10.1016/j.procs.2025.05.181](https://doi.org/10.1016/j.procs.2025.05.181).
- [42] L. Bergamin and F. Aiolfi, "Counterfactual examples for non-linear SVM," *Neurocomputing*, vol. 651, 2025, doi: [10.1016/j.neucom.2025.130809](https://doi.org/10.1016/j.neucom.2025.130809).
- [43] A. Alfaleh et al., "Predicting thermal conductivity using SVM: A review," *Energy Reports*, vol. 10, 2023, doi: [10.1016/j.egyr.2023.08.001](https://doi.org/10.1016/j.egyr.2023.08.001).
- [44] C. D. Sinaulan and A. Hantara, "Model Klasifikasi Permasalahan Kulit Wajah Menggunakan SVM," *Jurnal Pendidikan dan Kewirausahaan*, vol. 9, no. 1, 2021, doi: [10.47668/pkwu.v9i1.246](https://doi.org/10.47668/pkwu.v9i1.246).
- [45] N. W. A. Prasetya et al., "Sistem Pakar Deteksi Preeklamsia menggunakan Certainty Factor," *Infotekmesin*, vol. 13, no. 1, 2022, doi: [10.35970/infotekmesin.v13i1.1050](https://doi.org/10.35970/infotekmesin.v13i1.1050).
- [46] L. P. Wanti et al., "Diagnosis of ENT diseases using Naïve Bayes with genetic algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 1, pp. 398–405, 2025, doi: [10.11591/ijeecs.v37.i1.pp398-405](https://doi.org/10.11591/ijeecs.v37.i1.pp398-405).
- [47] W. Guo et al., "Distribution network topology identification using naive Bayes," *Energy Reports*, vol. 9, 2023, doi: [10.1016/j.egyr.2023.04.256](https://doi.org/10.1016/j.egyr.2023.04.256).
- [48] A. Chourlias et al., "Virtual sensors for smart farming: IoT and AI approach," *Internet of Things*, vol. 32, p. 101611, 2025, doi: [10.1016/j.iot.2025.101611](https://doi.org/10.1016/j.iot.2025.101611).
- [49] L. P. Wanti and L. Puspitasari, "Optimization of Fuzzy Logic for Autism Spectrum Disorder Diagnosis," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, 2022, doi: [10.29207/resti.v6i1.3599](https://doi.org/10.29207/resti.v6i1.3599).
- [50] J. Ventelä et al., "Associations between autoimmune diseases and leukemia," *Cancer Epidemiology*, vol. 86, p. 102411, 2023, doi: [10.1016/j.canep.2023.102411](https://doi.org/10.1016/j.canep.2023.102411).
- [51] W. A. Massawe et al., "Laryngopharyngeal reflux disease prevalence," *World Journal of Otorhinolaryngology*, vol. 7, no. 1, 2021, doi: [10.1016/j.wjorl.2020.04.009](https://doi.org/10.1016/j.wjorl.2020.04.009).