

A Predictive Model for Type 2 Diabetes Using A Wrapper-Based Feature Selection Method

Khairunisa Hilyati ^{1*}, Nuciko Abdul Halim ¹, Wendi Usino ²

* Correspondence Author: e-mail: 2311601625@student.budiluhur.ac.id

¹ Magister of Computer Science; Budi Luhur University; Jl. Ciledug Raya No.99, Petukangan Utara, Kec. Pesanggrahan, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12260; e-mail: 2311601625@student.budiluhur.ac.id, 2311601021@student.budiluhur.ac.id

² Faculty of Computer Science; Budi Luhur University; Jl. Ciledug Raya No.99, Petukangan Utara, Kec. Pesanggrahan, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12260 e-mail: wendi.usino@budiluhur.ac.id

Submitted : **06/03/2026**
Revised : **13/03/2026**
Accepted : **20/03/2026**
Published : **31/03/2026**

Abstract

Diabetes mellitus continues to show a rising global prevalence, making early detection of diabetes risk essential to prevent serious complications. This research aims to evaluate the effectiveness of a wrapper-based feature selection technique in improving the performance of classification models for early-stage diabetes risk prediction. The feature selection method employed is Recursive Feature Elimination (RFE), which is combined with three classification algorithms: Random Forest, Support Vector Machine (SVM), and Logistic Regression. The dataset used in this research was obtained from RSUD Pemangkat, Sambas Regency, West Kalimantan. The implementation of RFE is expected to identify and eliminate less relevant features, thereby simplifying the model, enhancing interpretability, and improving efficiency without compromising accuracy. This approach is particularly important in medical data analysis, where datasets are often complex and contain numerous clinical variables. Model performance is evaluated using accuracy, F1-score, and Area Under the Curve (AUC) to ensure a comprehensive assessment of classification capability. A comparative analysis is conducted to determine the optimal combination of feature selection method and classification algorithm that yields the best performance. In the scenario of applying the model with all features (baseline), Random Forest showed the best performance compared to other algorithms with an accuracy value of 0.9909, F1-Score of 0.9927, AUC of 0.9995, and sensitivity (recall) of 1.0000, which indicates that all cases of diabetes in the test data were successfully detected without false negative errors. SVM and Logistic Regression produced accuracies of 0.9545 and 0.9273, respectively. Despite having good classification capabilities, SVM tends to produce higher false positives, while Logistic Regression excels in the aspect of model interpretability. With an optimized model, the system has the potential to assist healthcare professionals in screening processes and clinical decision-making more quickly and effectively.

Keywords: feature selection, wrapper method, diabetes mellitus, Random Forest, Logistic Regression.

1. Introduction

Advances in information technology and artificial intelligence have brought about significant transformations across various sectors of life, including the healthcare field. In today's digital age, the use of data and intelligent algorithms has become a strategic solution for addressing increasingly complex global health challenges (Rajula et al., 2020). The integration of machine learning technology into healthcare systems enables large-scale analysis of medical data with high levels of accuracy and efficiency, thereby supporting more precise and rapid clinical decision-making (Beam &

Kohane, 2018). This digital transformation in healthcare not only improves the quality of diagnosis and treatment but also opens opportunities for the implementation of early disease detection systems that can reduce the global public health burden (Topol, 2019). However, the application of this technology requires an appropriate approach to processing medical data, particularly in the context of clinical data with specific characteristics and limitations.

Type 2 diabetes mellitus (T2D) is currently one of the major challenges in global health, with prevalence rates continuing to rise and projected to reach 783 million cases by 2045 (Jafar et al., 2025). This rise in prevalence is driven by several factors, such as urbanization, an aging population, unhealthy lifestyles, lack of physical activity, and poor dietary habits (Panda et al., 2022). T2D is the most prevalent type of diabetes globally and is closely linked to rising obesity rates and sedentary lifestyles, with a high body mass index being one of the primary risk factors (Sun et al., 2022). In Indonesia, this disease is one of the public health issues whose prevalence continues to rise across various regions. Sambas Regency, West Kalimantan, is among the areas with a relatively high disease burden, with the number of patients reaching 3,024, placing it third highest in West Kalimantan Province (Widhawati et al., 2024). This situation indicates that diabetes mellitus in Sambas Regency is a health issue requiring special attention and necessitating comprehensive prevention, early detection, and management strategies that consider social, environmental, and genetic predisposition factors (Khan et al., 2020).

Advances in artificial intelligence (AI) technology, particularly in the form of machine learning (ML), have opened up significant opportunities for developing efficient and accurate prediction systems based on medical data (Alvin et al., 2019). Machine learning is capable of identifying hidden patterns in complex data to assist in the diagnosis and clinical decision-making processes with a high degree of accuracy (Rami et al., 2019). In the context of diabetes prediction, ML algorithms can analyze various clinical and non-clinical parameters to identify individuals at high risk of developing T2D, thereby enabling earlier and more effective preventive interventions (Kopitar et al., 2020). ML's ability to process multidimensional data and capture complex interactions between variables makes it a highly promising tool for early diabetes detection systems, particularly in the era of big data in healthcare (Kavakiotis et al., 2017). However, the effectiveness of ML models heavily depends on the quality of input data and appropriate feature selection, necessitating a comprehensive methodological approach in the development of predictive models.

In the development of diabetes prediction models using machine learning, the feature selection process plays a crucial role in the machine learning pipeline (Christo et al., 2022). Feature selection aims to identify the subset of features that is most relevant and informative for the target variable, while eliminating features that are insignificant or redundant (Zeniarta et al., 2022). One widely used feature selection approach is the wrapper-based method, which evaluates the quality of a feature subset based on the performance of the classification model used (Tarkhaneh et al., 2021). This approach has the advantage of considering the interaction between features and the characteristics of the classification algorithm, thereby potentially yielding predictive models with better performance compared to filter-based feature selection methods (Hidayat & Putra, 2019). Optimal feature selection not only improves prediction accuracy but also reduces model complexity, enhances interpretability, and accelerates computation time, which is crucial for implementing prediction systems in clinical settings (Chandrashekar & Sahin, 2014).

Various machine learning algorithms have been applied to predict diabetes, with Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) being the most popular and proven effective methods (Sutanto et al., 2024). Random Forest is an ensemble learning algorithm that builds multiple decision trees

and generates predictions based on majority voting; it excels at handling non-linear data and is resistant to overfitting. Support Vector Machine uses an optimal hyperplane search approach for classification and is effective in high-dimensional spaces, although it is sensitive to kernel and parameter selection. Meanwhile, Logistic Regression is a simple yet powerful classical method for binary classification, with the advantage of high interpretability (Chen et al., 2025). In the context of feature selection, these models can be implemented using three approaches: (1) full-feature, which uses all available features; (2) selected-feature, which uses only selected features; and (3) a hybrid approach that combines both using weighted feature importance (Talpur et al., 2022). Comparing these three approaches is crucial for identifying the optimal strategy that balances predictive performance, computational efficiency, and model interpretability in the context of clinical applications.

The use of artificial intelligence for disease prediction has seen rapid advancements in recent years, with various successful implementations across medical fields (Esteva et al., 2021). AI systems for diabetes prediction have been developed in various countries and demonstrate high accuracy in identifying high-risk individuals based on demographic, clinical, and laboratory data (Dinh et al., 2019). The implementation of AI in healthcare is not limited to diagnosis but also encompasses the prediction of complications, personalized treatment, and patient risk stratification (Alvin et al., 2019). In developed countries, the integration of AI into electronic health records (EHRs) has enabled automated early warning systems capable of identifying high-risk patients for preventive interventions (Miotto et al., 2017). However, the adoption of AI technology for disease prediction in Indonesia, particularly in regional hospitals, remains very limited due to various technical and infrastructure constraints, as well as a shortage of competent human resources in the fields of data science and AI.

Although various studies have demonstrated the effectiveness of machine learning for diabetes prediction, there remains a theoretical gap in the comprehensive understanding of how wrapper-based feature selection influences the trade-off between model performance, computational complexity, and interpretability (Agustina et al., 2024). Most previous studies have focused on achieving high accuracy without in-depth analysis of the impact of feature selection on model stability, computational complexity, and clinical feature relevance (Tarkhaneh et al., 2021). There is currently no clear theoretical framework on how wrapper-based methods interact with various classification algorithms (RF, SVM, LR) in the context of full-feature, selected-feature, and hybrid approaches. The lack of theoretical understanding regarding the mechanisms of feature selection in addressing redundancy and irrelevance in clinical data with limited sample sizes also represents a gap that needs to be addressed. This theoretical gap is important to address in order to develop a more robust framework for ML implementation in diabetes prediction, particularly in the context of clinical data with unique characteristics.

2. Research Method

This research was conducted using the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which is a standard methodology in the data analysis process. This process consists of five main, integrated stages: business understanding, data understanding, data preparation, modeling, and evaluation of the resulting model's performance. Each stage in this framework is designed to ensure that the data analysis process proceeds systematically, in a structured manner, and in alignment with the research objectives. The research steps are presented in Figure 1 below:

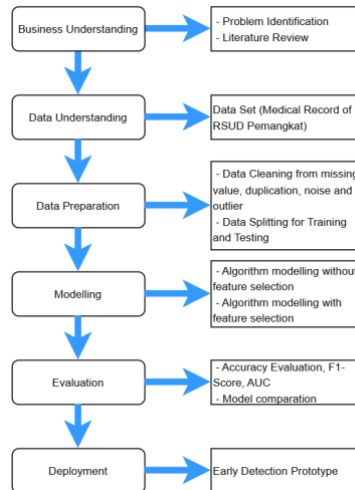


Figure 1. Research Steps

Figure 1 shows the Research Steps starting from Business Understanding Phase, Data Understanding Phase, Data Preparation Phase, Modelling Phase, Evaluation Phase and Deployment Phase with different scope for 1 phase to another phase. The data used in this research were obtained through primary data collection methods. The data were collected from patient medical records at the Pemangkat Regional General Hospital in Sambas, West Kalimantan. This research focused on two main components: hardware and software. Hardware includes all physical components used in the research. Software, on the other hand, includes computer programs, algorithms, and applications for data analysis, information processing, and the presentation of research results.

The data analysis techniques used in this research aim to evaluate the impact of applying a wrapper-based feature selection method on the performance of classification models in predicting the risk of early-stage diabetes. The analysis was conducted quantitatively, by measuring and comparing model performance based on several evaluation metrics commonly used in machine learning. The analytical steps include comparative experiments, model validation, evaluation metrics, statistical analysis, and visualization of results.

In each scenario, the classification process was performed using three predefined algorithms Random Forest, Support Vector Machine, and Logistic Regression resulting in several model combinations that were systematically evaluated. Model performance evaluation is conducted using comprehensive performance metrics, including accuracy, precision, recall, and F1-score, to provide a comprehensive overview of the effectiveness of each approach in terms of both prediction and class balance. To ensure the validity, stability, and generalization ability of the models, the training and testing processes were conducted using stratified k-fold cross-validation. This approach was chosen because it minimizes bias caused by uneven data partitioning and provides more robust and consistent model performance estimates compared to the single-fold hold-out method.

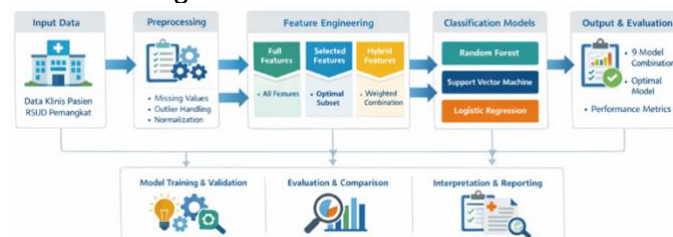


Figure 2. Diabetes Type 2 Model Design Prediction

The model design in this research was systematically developed to create an optimal prediction system for Type 2 Diabetes Mellitus by comparing various feature engineering approaches and classification algorithms. The developed system architecture consists of several key components that are integrated with one another, starting from the clinical data preprocessing stage, the implementation of three feature engineering approaches (full feature, selected feature, and hybrid feature), the application of three classification algorithms (Random Forest, Support Vector Machine, and Logistic Regression), to a comprehensive evaluation to determine the optimal combination. Each component was designed with consideration for the characteristics of Pemangkat Regional General Hospital's clinical data, which has limitations in sample size, potential missing values, and the need for model interpretability for practical implementation in a clinical setting.

3. Results and Analysis

3.1. Data Understanding

The initial analysis phase of this research focused on gaining a comprehensive understanding of the data characteristics before proceeding with further processing. Activities conducted included analyzing data distribution to determine the range of values for each variable, identifying common patterns that emerge, and detecting potential anomalies, inconsistencies, or data deviating from normal conditions. This process aimed to provide an initial overview of the quality and structure of the data used. Each variable in the dataset was systematically examined to ensure the validity and completeness of the data. This examination includes identifying missing values, duplicate data, and mismatches in data types or value ranges. Additionally, mapping and standardization of values for each variable are performed to ensure consistency in data format, thereby minimizing the potential for errors in subsequent processing stages. This effort is crucial, as inconsistent data can affect the accuracy of analysis and modeling results.

3.2. Data Collection

Data collection for this research was conducted using secondary data sourced from patient medical records at Pemangkat Regional General Hospital. The researcher obtained official access to the hospital's medical records information system after receiving written approval and permission from the Head of the Medical Records Unit. This access was granted on a limited basis and was used solely for research purposes. A complete description of each attribute used is provided in Table 1 below:

Table 1. Attribute Data Variable

No	Variable	Definition	Relevance	Notes
1	Age	The patient's age, calculated based on the date of birth up to the time of the visit.	This is an important risk factor because the prevalence of this disease tends to increase with age.	Numerical data (years).
2	Gender	Gender Biological characteristics of patients that distinguish between males and females.	Gender influences differences in risk and physiological responses to diabetes.	Categorical data (Male/Female).
3	Complaints	The primary complaints reported by patients during a healthcare visit.	Clinical complaints can serve as early indicators of metabolic disorders associated with type 2 diabetes.	Categorical data / text

No	Variable	Definition	Relevance	Notes
4	Diagnosis	Classification of the patient's condition based on recorded medical diagnoses.	Diagnosis is used as the target variable in the modeling process.	Target variable (Diabetes / Non-Diabetes)

The data collected resulted from a query of patient records that had been filtered based on time criteria, specifically the period from September to November 2025. Throughout the data collection process, the researchers consistently applied research ethics principles, with the protection of patient privacy and confidentiality being the top priority. Thus, all stages of data collection were conducted systematically, responsibly, and in accordance with applicable health research ethics guidelines. The symptoms described in the medical records were manually entered into an Excel format, resulting in data as shown in Table 2.

Table 2. Attribute Data Description

No.	Medical Record No	Age	Gender	Maternity	Poliuria	Polidipsia	Genetik Diabetes	*****	Diagnosis
1	94527183	21	Male	No	Yes	No	No	*****	Diabetes
2	54629817	38	Male	No	No	No	Yes	*****	Non-Diabet
3	63829475	42	Male	No	No	No	No	*****	Non-Diabet
4	59738462	46	Male	No	No	No	Yes	*****	Non-Diabet
5	21458639	55	Male	No	Yes	Yes	Yes	*****	Diabetes
6	95472836	58	Male	No	No	No	No	*****	Diabetes
7	51392468	67	Male	No	No	Yes	Yes	*****	Diabetes
8	84329576	22	Male	No	No	No	No	*****	Diabetes
*	*****	**	*****	*****	*****	*****	*****	*****	*****
*	*****	**	*****	*****	*****	*****	*****	*****	*****
544	49678253	38	Male	No	No	No	Yes	*****	Non-Diabet
545	86142795	41	Male	No	Yes	Yes	Yes	*****	Diabetes
546	36492578	46	Male	No	No	No	No	*****	Non-Diabet
547	24798531	50	Female	No	Yes	Yes	Yes	*****	Diabetes
548	83572641	54	Male	No	No	No	Yes	*****	Non-Diabet
549	93582746	58	Male	No	No	No	No	*****	Non-Diabet
550	48295763	67	Male	No	Yes	Yes	Yes	*****	Diabetes

3.3. Analysis Technique

The data analysis techniques used in this research aim to evaluate the impact of applying a wrapper-based feature selection method on the performance of classification models in predicting the risk of early-stage diabetes. The analysis was conducted quantitatively, by measuring and comparing model performance based on several evaluation metrics commonly used in machine learning. The analysis steps are outlined as follows:

3.3.1. Comparative Experiment

This research consists of two main scenarios:

- a. Baseline Model (without feature selection): A classification model is built using all features in the dataset, without undergoing a feature selection process. The results from this scenario are used as a baseline.
- b. Model with Feature Selection: The model is developed after applying the Recursive Feature Elimination (RFE) method, and the selection results are then used to build the classification model.

Each scenario will be tested using three algorithms: Random Forest, Support Vector Machine (SVM), and Logistic Regression..

3.3.2. Model Validation

Each model combination will be validated using the k-fold cross-validation technique with $k = 5$. This approach was chosen to reduce the likelihood of bias from single-fold data splitting and to provide a more accurate and stable performance estimate.

3.3.3. Metrix Evaluation

At this stage, the model will be tested to assess its accuracy following the feature selection process. The evaluation and validation processes are conducted using the Confusion Matrix methods.

3.3.4. Statistic Analysis

To determine the statistical significance of the differences between models with and without feature selection, inferential statistical analyses were conducted using nonparametric tests such as the Wilcoxon signed-rank test or ANOVA, depending on the data distribution. The purpose of this was to test whether the observed improvement in performance was truly statistically significant and not merely a result of data variability.

3.3.5. Visualization Result

The model evaluation results are presented in the following formats:

- a. Comparative table: displaying accuracy, F1-score, and AUC values for each model.
- b. Bar charts and heatmaps: to visualize comparisons between combinations of selection methods and classification algorithms.

3.4 Design and Data Testing

This research was designed to evaluate the performance of a predictive classification model for Type 2 Diabetes Mellitus in two main scenarios, as previously described in the system architecture design and feature engineering approach.

- a. The first scenario is a classification model without feature selection (full feature), in which all clinical features of patients at Pemangkat Regional General Hospital including demographic, clinical, and laboratory data are used in their entirety in the model training process. This approach serves as the baseline model, aiming to measure the initial performance of the classification algorithm without any feature selection intervention, while also representing the maximum complexity of the data.
- b. The second scenario is a classification model with feature selection (selected features), where the feature selection process is performed prior to the model training stage using the Recursive Feature Elimination (RFE) method. The RFE method works iteratively by evaluating the contribution of each feature to the model's performance,

then gradually eliminating the features with the lowest impact until an optimal subset of features is obtained. This approach aims to improve computational efficiency, reduce the risk of overfitting, and maintain or improve prediction accuracy.

3.5 Feature Selection

The feature selection process is a critical step in developing a model for the early detection of Type 2 Diabetes Mellitus (T2D), aimed at ensuring that only features with high relevance and significance for disease diagnosis are retained in the analysis. Attribute selection is conducted by considering clinical relevance as well as findings from various literature supporting the relationship between these variables and the risk of developing diabetes. The retained attributes include demographic data such as gender and age, as well as various clinical symptoms and medical history that play a crucial role in the diagnostic process. These symptoms include pregnancy, polyuria, polydipsia, weight loss, fatigue, polyphagia, infections, blurred vision, itching, irritability, slow wound healing, tingling, muscle stiffness, hair loss, a family history of diabetes, and obesity. All of these attributes are retained because they have been empirically proven to contribute significantly to the early identification of diabetes, as supported by various previous studies.

From the data collection process, a total of 550 records were obtained, with the unit of analysis being the individual, and containing a number of attributes representing demographic characteristics, medical history, and clinical symptoms related to Type 2 Diabetes. These attributes include age, gender, pregnancy history, polyuria, polydipsia, sudden weight loss, fatigue, polyphagia, blurred vision, itching, irritability, tingling, slow-healing wounds, obesity, family history of diabetes, and one target attribute, namely "class," which consists of two categories: Positive and Negative.

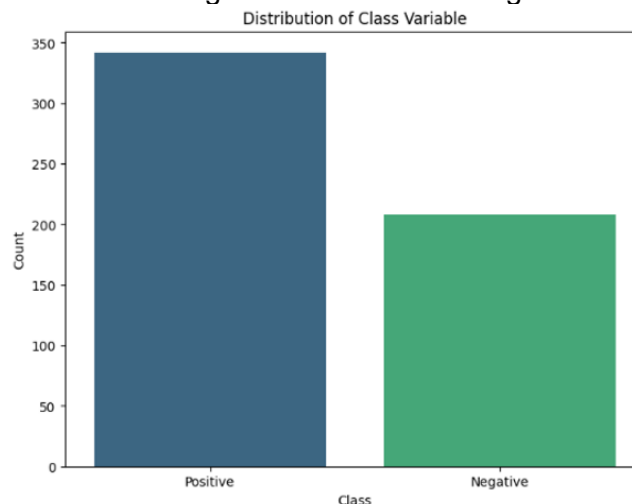


Figure 3. Dataset Composition

Based on Figure 3, it is observed that the Positive class has a larger number of data points, approximately 340 records, while the Negative class has approximately 210 records. This indicates that the dataset has a moderately imbalanced class distribution, yet it remains within acceptable limits for classification research. This information serves as a crucial foundation in the modeling phase, particularly in selecting the appropriate algorithm and evaluation metrics so that the model's performance can fairly represent its predictive capability for both classes.

3.5 Analysis Result

In the results visualization, the model's performance using all features is typically displayed as a "Baseline" bar chart. This allows users to see directly whether

using all the data yields better or worse results compared to more advanced feature selection methods. This analysis is crucial for testing the hypothesis that not all collected medical data is relevant for diagnosing diabetes. Sometimes, using fewer but more targeted features can significantly improve the model's generalization.

3.5.1. Random Forest

Based on the modeling results, the Random Forest achieved an accuracy of 0.9909, indicating that the model was able to classify diabetes status with a very high degree of accuracy on the test data. This accuracy score indicates that the model effectively captured the patterns of relationships between clinical variables and diabetes status. With a very low error rate, Random Forest can be considered to have good generalization capabilities, making it a strong candidate for a decision support system for predicting type 2 diabetes.

Table 3. Confusion Matrix model Random Forest

Index	Predicted Negative	Predicted Positive
True Negative	41	1
True Positive	0	68

Table 3 presents the performance metrics of the classification model following the application of the feature selection process, presented in the form of a confusion matrix. Based on these results, the model successfully identified 41 data points as True Negatives—that is, negative-class data correctly predicted—and 68 data points as True Positives, demonstrating the model's ability to accurately classify positive data. Furthermore, only 1 data point was classified as a False Positive, while no False Negative cases were found. These findings indicate that the model has excellent capability in recognizing the positive class with minimal error rates, leading to the conclusion that the attribute selection process contributes significantly to improved classification performance.

3.5.2. Support Vector Machine (SVM)

The SVM model was then built using an RBF kernel with parameters $C = 1$ and $\gamma = \text{'scale'}$, as well as $\text{class_weight} = \text{'balanced'}$ to address potential class imbalance between diabetic and non-diabetic patients. The data was split using a stratified train-test split with an 80% training set and a 20% test set, ensuring that the class distribution remained preserved. Based on the test results, the SVM achieved an accuracy of 0.9545, demonstrating the model's high capability in classifying type 2 diabetes status. Although slightly lower than that of the Random Forest, this result confirms that the SVM is a reliable model, particularly in handling high-dimensional data with nonlinear relationships among clinical features.

Table 4. Confusion Matrix model SVM

Index	Predicted Negative	Predicted Positive
True Negative	0	42
True Positive	0	68

Table 4 presents the performance evaluation results of the Support Vector Machine (SVM) model, represented in the form of a confusion matrix. Based on the table, all data with the negative class were predicted as positive, so there were no True Negative values, while 42 data points fell into the False Positive category. Additionally, the model successfully classified 68 positive data points correctly as True Positive, and no classification errors in the form of False Negatives were found. These results indicate that the SVM model exhibits very high sensitivity toward the positive class but struggles to distinguish the negative class, leading to a significant error rate in

predicting the negative class. This suggests the model is biased toward the positive class and requires parameter tuning or more optimal data handling strategies.

3.5.3. Logistic Regression

Based on the evaluation results, Logistic Regression achieved an accuracy score of 0.9273, indicating that the model is capable of classifying data with a high degree of accuracy. Although its accuracy is lower than that of Random Forest and SVM, Logistic Regression remains valuable because it provides clear insights into the contribution of each feature to the risk of type 2 diabetes, making it highly relevant for clinical analysis and medical decision-making.

Table 5. Confusion Matrix model SVM

Index	Predicted Negative	Predicted Positive
True Negative	37	5
True Positive	1	67

Table 5 presents the performance evaluation results of the Logistic Regression (LR) model, displayed in the form of a confusion matrix. Based on the table, the model successfully classified 37 data points as True Negatives—that is, negative data correctly predicted—and 67 data points as True Positives, demonstrating the model's ability to accurately identify the positive class. In addition, there are 5 data points classified as False Positives and 1 data point classified as a False Negative. These results indicate that the LR model achieves a fairly good balance in predicting both classes, although a small number of classification errors still occur. Overall, the model's performance can be considered quite optimal and stable in distinguishing between positive and negative classes.

3.5 Prototype

The resulting system prototype is a tangible manifestation of the integration of all the technical components discussed, packaged as a ready-to-use web application. This prototype features a structured workflow, ranging from the Data Upload page to the Model Training page and the Live Prediction page. The user interface design prioritizes clear navigation, featuring a persistent sidebar menu and informative process status indicators. Users are guided step by step through the model training process, from the file upload button to the save model button, with a progress bar visualization that provides visual feedback on the progress of the background process.

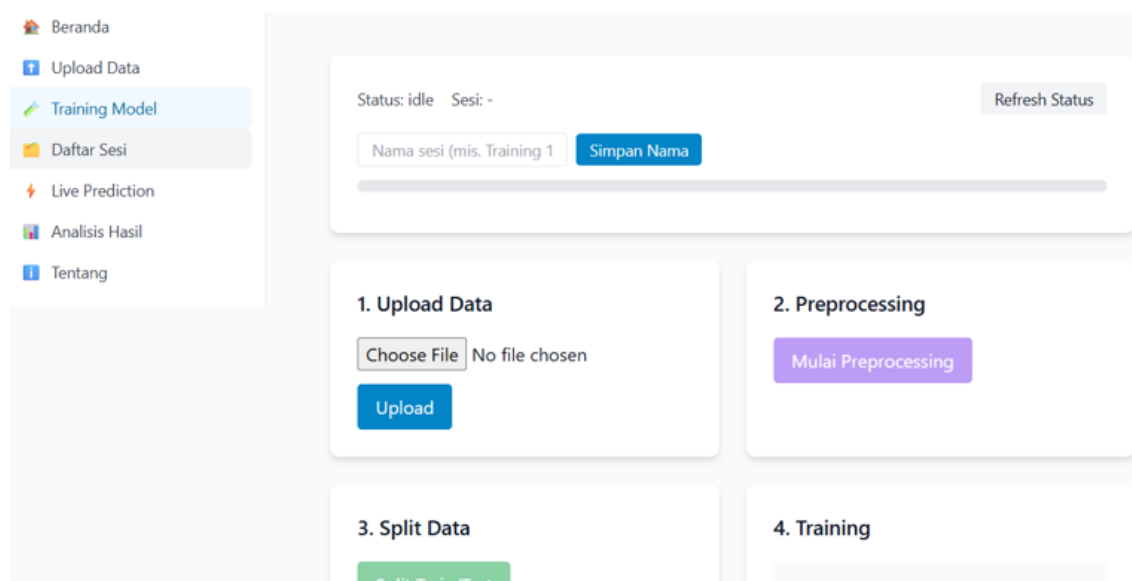


Figure 4. Training Model Page

The Training page in this prototype features an interactive terminal log, which represents a breakthrough in transparency within machine learning applications. Users don't just see a loading animation; they can view technical details of the ongoing training process, such as the number of epochs and the loss value of the neural network. This feature gives users a greater sense of control and a deeper understanding of what the system is doing. Additionally, the training parameter configuration form allows even novice users to experience what it's like to be a data scientist by directly adjusting model variables.

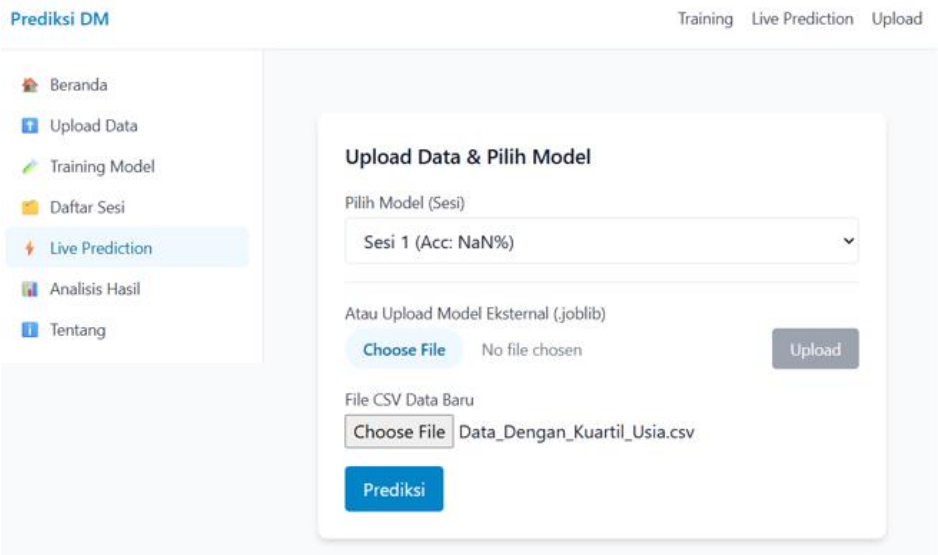


Figure 5. Model Upload Page

The Live Prediction section of the prototype demonstrates the system's practical utility, allowing users to apply the model they have just trained to predict data for new patients. The model selection feature via a dropdown menu highlights the system's flexibility in managing various versions of historical models. For more details on the prediction result will be shown in the Figure 6 below.

NO	PREDIKSI	KEYAKINAN (PROBABILITAS)
1	Positif Diabetes	95.00%
2	Negatif	0.00%
3	Negatif	0.00%
4	Negatif	0.00%
5	Positif Diabetes	96.00%

Figure 6. Prediction Result

The prediction results are displayed in a neat tabular format, complete with the model's confidence probabilities where available. This prototype also includes a Results Analysis page that visualizes performance comparisons across training

sessions using bar and radar charts, facilitating data-driven decision-making. Overall, this prototype has met the functionality standards for a basic clinical decision support system. Input validation, error handling, and result visualization have functioned as per the test scenarios. This prototype serves not only as a prediction tool but also as an educational resource regarding the machine learning process itself. The clean, modern interface with medical-themed colors (blue/green) enhances user comfort, while the system's response speed demonstrates the efficiency of the implemented backend architecture. This prototype is ready for further testing by end-users in a controlled environment.

4. Conclusion

Based on the study findings, the following conclusions are drawn: This study successfully developed type 2 diabetes prediction models using Random Forest, SVM, and Logistic Regression following the CRISP-DM methodology. Using all features, Random Forest outperformed the other models, achieving the highest accuracy (0.9909), F1-Score (0.9927), and AUC (0.9995), with perfect sensitivity (1.0000). SVM and Logistic Regression achieved lower accuracies (0.9545 and 0.9273, respectively), with SVM showing higher false positives and Logistic Regression offering better interpretability. The RFE method identified eight key features (Gender, Pregnancy, Polyuria, Polydipsia, Tingling, Irritability, Itching, Infection). After feature selection, SVM (RFE) achieved the most balanced performance (accuracy: 0.9727, F1-Score: 0.9778), while Logistic Regression (RFE) maintained perfect sensitivity but increased false positives.

Acknowledgements

We would like to express our deepest gratitude to the Computer Science of Budi Luhur University lecturer, friends and family for their support and trust to support this research. And for those who involved in helping us complete this research.

Author Contributions

Khairunisa Hilyati proposed the topic; Khairunisa Hilyati conceived models and designed the experiments; Nuciko Abdul Halim conceived the optimisation algorithms; Khairunisa Hilyati and Nuciko Abdul Halim analysed the result for the research in consultation with Mr. Wendi Usino.

Conflicts of Interest

The author declare no conflict of interest

References

- Agustina, T., Masrizal, M., & Irmayanti, I. (2024). Performance Analysis of Random Forest Algorithm for Network Anomaly Detection using Feature Selection. *Sinkron*, 8(2), 1116–1124. <https://doi.org/10.33395/sinkron.v8i2.13625>
- Alvin, R., Jeffrey, D., & Isaac, K. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMr1814259>
- Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA*, 319(13), 1317–1318. <https://doi.org/10.1001/jama.2017.18391>
- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28. <https://doi.org/https://doi.org/10.1016/j.compeleceng.2013.11.024>
- Chen, H., Song, H., Huang, H., Fang, X., Chen, H., Yang, Q., Zhang, J., Ding, W.,

- Gong, Z., & Ke, J. (2025). Machine learning prediction and interpretability analysis of high-risk chest pain: a study from the MIMIC-IV database. *Frontiers in Physiology*, 16(June), 1–17. <https://doi.org/10.3389/fphys.2025.1594277>
- Christo, V. R. E., Nehemiah, H. K., Brighty, J., & Kannan, A. (2022). Feature Selection and Instance Selection from Clinical Datasets Using Co-operative Co-evolution and Classification Using Random Forest. *IETE Journal of Research*, 68(4), 2508–2521. <https://doi.org/10.1080/03772063.2020.1713917>
- Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(1), 1–15. <https://doi.org/10.1186/s12911-019-0918-5>
- Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2021). Deep learning-enabled medical computer vision. *Npj Digital Medicine*, 4(1), 1–9. <https://doi.org/10.1038/s41746-020-00376-2>
- Hidayat, B. M. H., & Putra, R. E. (2019). Penerapan CNN dengan Filter Gabor sebagai feature extractor untuk Content-Based Image Retrieval. *Journal of Informatics and Computer Science (JINACS)*, 1(01), 16–25. <https://doi.org/10.26740/jinacs.v1n01.p16-25>
- Jafar, N., Setyawati, A., Malasari, S., & Grestin Sandy R. (2025). Motivational interviewing-based health coaching to improve diabetes knowledge and manage blood glucose: A quasi-experimental study. *Jurnal Kedokteran Dan Kesehatan Indonesia*, 70–79. <https://doi.org/10.20885/jkki.vol16.iss1.art8>
- Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104–116. <https://doi.org/10.1016/j.csbj.2016.12.005>
- Khan, M. A. B., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., & Kaabi, J. Al. (2020). Epidemiology of Type 2 diabetes - Global burden of disease and forecasted trends. *Journal of Epidemiology and Global Health*, 10(1), 107–111. <https://doi.org/10.2991/JEGH.K.191028.001>
- Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, 10(1), 1–12. <https://doi.org/10.1038/s41598-020-68771-z>
- Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236–1246. <https://doi.org/10.1093/bib/bbx044>
- Panda, M., Mishra, D. P., Patro, S. M., & Salkuti, S. R. (2022). Prediction of diabetes disease using machine learning algorithms. *IAES International Journal of Artificial Intelligence*, 11(1), 284–290. <https://doi.org/10.11591/ijai.v11.i1.pp284-290>
- Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina (Lithuania)*, 56(9), 1–10. <https://doi.org/10.3390/medicina56090455>
- Rami, V., Chirra, R., Uyyala, S. R., Krishna, V., & Kolli, K. (2019). *Revue d ' Intelligence Artificielle Deep CNN: A Machine Learning Approach for Driver Drowsiness Detection Based on Eye State*. 33(6), 461–466.
- Sun, J. Y., Huang, W. J., Hua, Y., Qu, Q., Cheng, C., Liu, H. L., Kong, X. Q., Ma, Y. X., & Sun, W. (2022). Trends in general and abdominal obesity in US adults: Evidence from the National Health and Nutrition Examination Survey (2001–2018). *Frontiers in Public Health*, 10. <https://doi.org/10.3389/fpubh.2022.925293>
- Sutanto, T., Aditya, M. R., Budiman, H., Noor Ridha, M. R., Syapotro, U., & Azijah, N. (2024). Comparison of Logistic Regression, Random Forest, SVM, KNN Algorithm for Water Quality Classification Based on Contaminant Parameters. *INTI Journal*,

- 2022(1). <https://doi.org/10.61453/jods.v2023no48>
- Talpur, S., Azim, F., Rashid, M., Syed, S. A., Talpur, B. A., & Khan, S. J. (2022). Uses of Different Machine Learning Algorithms for Diagnosis of Dental Caries. *Journal of Healthcare Engineering*, 2022(Cdc). <https://doi.org/10.1155/2022/5032435>
- Tarkhaneh, O., Nguyen, T. T., & Mazaheri, S. (2021). A novel wrapper-based feature subset selection method using modified binary differential evolution algorithm. *Information Sciences*, 565, 278–305. <https://doi.org/https://doi.org/10.1016/j.ins.2021.02.061>
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Widhawati, R., Lubis, V. H., & Komalasari, O. (2024). Implementasi Protokol Pengkajian Dan Perawatan Kaki Diabetik Berbasis Video Kharisma. *Jurnal Pengabdian Kepada Masyarakat (JPKM) - Aphelion*, 4, 171–178. <https://jurnal.globalhealthsciencegroup.com/index.php/JPM/article/view/2494>
- Zeniarja, J., Salam, A., & Ma'ruf, F. A. (2022). Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa. *Jurnal Rekayasa Elekrika*, 18(2), 102–108. <https://doi.org/10.17529/jre.v18i2.24047>