ISSN 2579-7069

A Comparative Study of Naive Bayes, SVM, and Decision Tree Algorithms for Diabetes Detection Based on Health Datasets

Satria Dwi Nurwicaksana[1,*], Lee Kyung Oh[2], Husni Teja Sukmana[3]

[1]*Program of Information System,Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia*
[2]*Sun Moon University Asan, Republic of Korea,*
[3]*Informatics Department, Faculty of Science and Engineering, Universitas Islam Negeri Syarif Hidayatullah, Jakarta, Indonesia*

(Received February 25, 2024; Revised May 30, 2024; Accepted September 5, 2024; Available online December 1, 2024)

Abstract

Diabetes is a chronic, progressive condition whose global prevalence continues to rise, creating substantial public health and economic burdens. Early diagnosis and timely intervention are critical to preventing severe complications and improving long-term patient outcomes. In recent years, artificial intelligence (AI) particularly machine learning (ML) has emerged as a powerful tool in medical diagnostics, offering capabilities in automated pattern recognition and disease classification. This study aims to evaluate and compare the predictive performance of three supervised ML algorithms such as Naïve Bayes, Support Vector Machine (SVM), and Decision Tree for classifying and predicting diabetes based on two primary physiological indicators: glucose level and blood pressure. The dataset employed was sourced from Kaggle, comprising 995 patient records containing relevant clinical attributes. The research methodology involved several stages, including data preprocessing to ensure quality and consistency, data partitioning into training and testing subsets using an 80:20 split ratio, model training, and performance evaluation. Each algorithm's effectiveness was measured using accuracy, precision, recall, and F1-score metrics. The experimental findings demonstrate that the Decision Tree algorithm achieved the highest classification accuracy (94.47%), outperforming SVM and Naïve Bayes, both of which recorded 92.96% accuracy. Moreover, the Decision Tree exhibited balanced precision and recall values, underscoring its robustness in identifying both diabetic and non-diabetic cases with minimal misclassification. These outcomes indicate that the Decision Tree model provides an optimal balance between predictive accuracy and interpretability, making it particularly suitable for clinical decision-support applications.

Keywords: Machine Learning, Decision Tree, Naive Bayes, SVM, Classification, Health, Diabetes

1. Introduction

Diabetes is a disease that is becoming a health problem with numbers increasing every time and experiencing a consistent increase [1]. It is estimated that from 1990, diabetes has recorded an increase until the number of diabetes cases reaches 829 million by 2022 [2]. This disease is characterized by high levels of glucose in the blood caused by impaired insulin production or action. Diabetes must be treated quickly so as not to cause serious complications. Early diagnosis of diabetes is needed to determine whether someone has a symptom based on checking glucose levels and insulin action. Therefore, the utilization of technology in health, especially artificial intelligence technology, is a good choice to do a very early diagnosis for diabetes detection. Machine learning is a part of artificial intelligence that can be used. Doctors are already using the help of deep learning to find medical problems and search for genetic information from patients to detect diseases [3].Machine learning can do this with the help of machine learning algorithms in it.

Machine learning algorithm is one of the important parts in machine learning technology. In its application in the medical world for the detection of diabetes diagnosis, algorithms can be used to train patient data to be used to identify certain features related to a particular disease. By using a classification model, an algorithm can predict whether a person is likely to suffer from diabetes based on glucose levels and blood pressure.

This research aims to compare three machine learning algorithms, namely Naive Bayes, Support Vector Machine (SVM) and Decision Tree, in the classification of diabetes based on a health dataset. Basically, naïve bayes has almost the same use as decision tree and SVM. Naïve Bayes It is used to predict the probability of each class in the data as the probability that a given data can fall into a particular class [4]. Then Support Vector Machine (SVM) is a classification

***Corresponding author: Satria Dwi Nurwicaksana (satriamapeldmi@gmail.com)**
DOI: https://doi.org/10.47738/ijiis.v7i4.230

that divides data into groups using hyperlane inter-class distances [5]. Furthermore, Decision tree is a classification algorithm that uses internal nodes that describe feature relationships with leaf nodes that show the potential for the results to be understood [6]. In this research, the data used in the dataset is 995 rows and 2 main features, namely glucose levels and blood pressure. The main focus of this research is to evaluate the accuracy of each algorithm in its application for early detection of diabetes. Through this approach, it is expected to produce research results that can help as a fast and efficient diabetes early detection system based on previous data experience so that medical personnel can be helped to diagnose the disease.

2. Literature Review

2.1. Overview of Machine Learning Algorithms in Classification

Machine learning (ML) algorithms have shown remarkable success in various data classification tasks across different domains. Techniques such as Naïve Bayes, Support Vector Machine (SVM), and Decision Tree are widely used because of their ability to learn complex patterns and make accurate predictions from structured datasets. Each algorithm offers distinctive advantages—Naïve Bayes is simple and probabilistic, SVM provides high accuracy with optimal margins, and Decision Tree models are intuitive and interpretable for non-linear relationships. Previous studies have highlighted these algorithms' diverse applications. For example, [7] compared the accuracy of SVM, K-Nearest Neighbor (KNN), Decision Tree, and Naïve Bayes in classifying obesity data. The study revealed that the Decision Tree algorithm achieved the highest accuracy, reaching 84.98%, demonstrating the ability of ML models to process health-related datasets effectively. Similarly, [8] evaluated the performance of Decision Tree, SVM, and Naïve Bayes for lung cancer prediction using evaluation metrics such as accuracy, precision, recall, and F1-score. The findings indicated that Decision Tree and Naïve Bayes yielded the best predictive performance, emphasizing the reliability of ML in medical diagnostics.

2.2. Algorithm Performance in Health-Related Studies

Further research by [9] focused on kidney disease diagnosis, using both training and testing data to evaluate model performance. The results showed that SVM achieved the highest accuracy of 97.75%, followed by Decision Tree (97.50%) and Naïve Bayes (95.75%). This indicates that SVM excels in classification problems where data patterns are complex and multidimensional. Collectively, these findings demonstrate that ML algorithms can efficiently analyze patient health data, support early detection, and enhance diagnostic decision-making. In broader health informatics, ML has been instrumental in identifying chronic diseases such as diabetes and hypertension by analyzing diverse variables like glucose level, blood pressure, and body mass index. The consistent results across studies affirm that these algorithms can handle both linear and non-linear relationships, making them valuable for predictive analytics in healthcare. Furthermore, the combination of precision, recall, and F1-score as evaluation metrics provides a more holistic understanding of model performance beyond accuracy alone.

2.3. Applications Beyond Healthcare and Research Implications

Beyond the medical domain, ML algorithms are also effective in diverse fields such as sentiment analysis, cybersecurity, and social media analytics. For instance, [10] utilized SVM and Naïve Bayes to classify movie reviews on IMDb, achieving accuracy scores of 88% and 85%, respectively. These findings show that ML can effectively capture linguistic nuances and classify textual sentiments. Similarly, [11] compared SVM with Logistic Regression, Naïve Bayes, KNN, and Decision Tree algorithms in detecting cyber-attacks in automotive systems. The study concluded that SVM demonstrated the highest accuracy, sensitivity, and specificity, proving its robustness for intrusion detection and data security applications. Based on these studies, it is evident that machine learning algorithms—particularly Naïve Bayes, SVM, and Decision Tree—consistently deliver strong classification performance across domains. This serves as the foundation for the present study, which aims to compare these three algorithms in classifying diabetes data. The research specifically evaluates accuracy, precision, recall, and F1-score to identify which model is most suitable for detecting diabetes based on glucose and blood pressure variables. The findings are expected to contribute to the development of a machine learning-based diagnostic system capable of classifying health data rapidly and accurately, thereby supporting early disease detection and effective medical decision-making.

 ISSN 2579-7069

3. Method

This research uses an experimental quantitative approach that aims to compare the prediction accuracy of three classification algorithms, namely Decision Tree, Naïve Bayes, and Support Vector Machine (SVM) on diabetes disease datasets. This approach was chosen because it allows an objective and systematic measurement of the performance of each algorithm. By conducting direct experiments on the data, this approach is expected to provide results that can be tested and have a high level of validity in the context of evaluating the performance of machine learning algorithms.

In this study, each algorithm is tested using the same data with a total of 995 data and through the preprocessing stage, so that the resulting accuracy comparison is really able to classify diabetes data properly. The main focus of this approach is to determine the extent to which the three algorithms are able to accurately predict whether a person is at risk of developing diabetes or not, based on the variables of glucose levels, and blood pressure with diabetes being the target variable.

The model evaluation stage is conducted by measuring various performance metrics such as accuracy, precision, recall, and F1-score, which aims to provide a comprehensive overview of the strengths and weaknesses of each algorithm. All evaluation processes are designed systematically and sequentially to ensure consistent and reliable results. The evaluation process of this model can be seen visually in Figure 1.

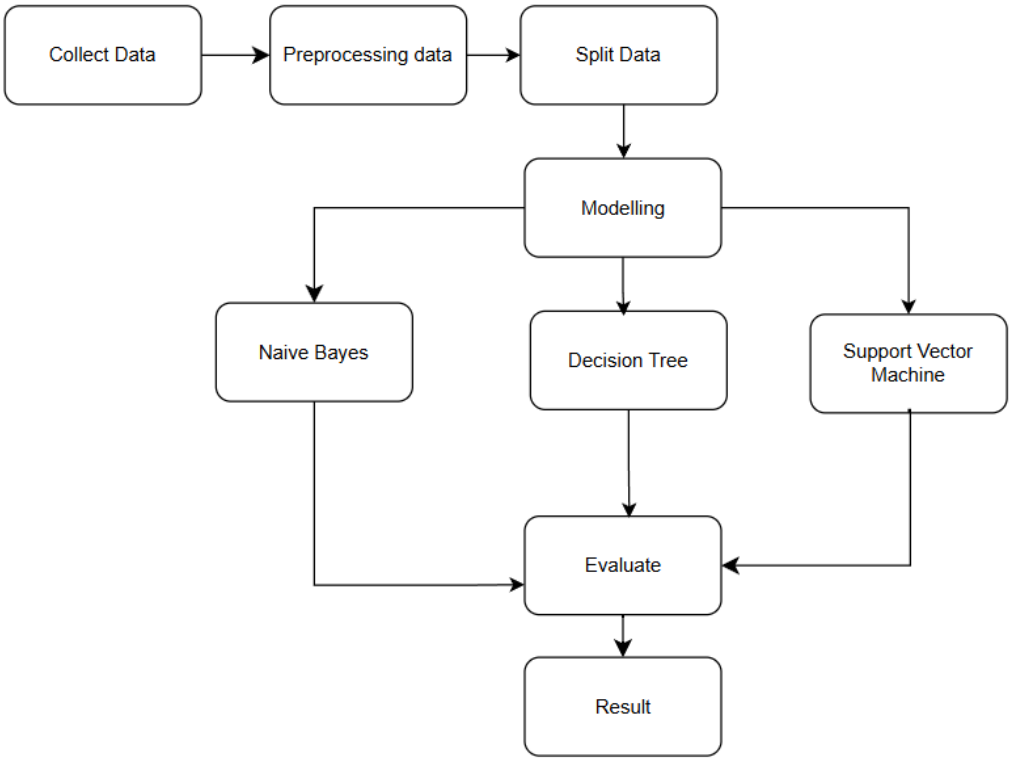

Figure 1. Flow of Research Methods

3.1. Collect Data

In this study using a public dataset derived from the Kaggle site. This dataset consists of 995 rows and 2 feature columns, namely glucose and blood pressure and 1 target column that indicates whether a person has diabetes or not, labeled 0 and 1 (diabetes). This dataset was chosen because it is relevant and can be used for prediction and classification. The collection of this dataset is important because it is a critical component that has a significant impact on the model [12].

 ISSN 2579-7069

3.2. Preprocessing Data

In this stage the data used will be carried out at several stages so that the data is ready to be used according to the training and testing needs. Such as cleaning data, calculating missing values, coding categories and selecting features [13]. The collected dataset is checked to remove or replace empty values on attributes that should have values. This stage ensures and checks such as outliers and data duplication [14]**.** Determines the final quality of the analysis as the data produces an accurate model that can be used for testing.

3.3. Split Data

This research applies the technique of dividing data into two subsets, namely training data and testing data. The size of the dataset becomes the reference for the division of the existing data ratio [15]. The purpose of this division is to ensure that the developed model can learn effectively from the training data, and then be tested with data that was not used in the training process to assess the generalization performance of the model. In this context, the training data acts as the basis for model learning, while the testing data provides an overview of the extent to which the model is able to recognize patterns from new data. The division process is done proportionally, where 80% of the overall data is used as training data and the remaining 20% is used for testing purposes.

3.4. Modeling Algorithm

In this stage, three machine learning algorithms are implemented as models to be built to perform the classification approach and see the diverse perspectives of the data prediction performance. Decision Tree is an algorithm that forms a tree structure in order to make feature-based decisions. It consists of root nodes, splitting nodes, and leaf nodes and at each node, there is a feature and associated threshold [16] Splitting nodes use the Gini Impunity criterion to minimize node heterogeneity.

$$Entropy(S) = -\sum_{i=1}^{c} p \log_2(p)$$

$$Gain\ (S, A) = Entropy(S) - \sum_{Values(A)} \frac{S_v}{S} Entropy(S_v)$$

With S = One set data, c = Number of classes, p = The proportion of data in class i, A = attribute under test, Values(A) = all unique values of attribute A, and Sv: subset of S for value v.

Naive Bayes is method classifies data using probability and statistical methods that predict future opportunities based on previous experience. This classification is likelihood-based by favoring the most likely value [17]. This algorithm is based on the Bayes Theorem calculation to find predictions based on previous data.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

With P(H|X): Probability of hypothesis H after getting data X (posterior), P(X|H) The probability that data X occurs if H is true (likelihood), P(H) Initial probability of hypothesis H (prior), and P(X) The probability of data X as a whole (evidence).

Support Vector Machine (SVM) is method is a hyperlane-based data classification method utilizing the distance between each class. This algorithm ensures dynamic feature extraction, energy efficiency and adaptive decision of the model [18] . SVM has high performance on high-dimensional data and has an optimal hyperplane to separate different classes. SVM finds a hyperplane that has the largest possible fraction of points from the same class on the same plane [19].

3.5. Evaluation

The performance evaluation of the model was conducted to measure the extent to which the algorithm used was able to classify health data accurately and reliably. This evaluation process uses several common metrics in machine learning, namely accuracy, precision, recall (sensitivity), and f1-score. The accuracy value is used to calculate the percentage of correct predictions compared to the overall test data, which reflects the general performance of the model. Precision evaluates the extent to which the model is able to correctly identify positive data, while recall measures the model's ability to capture all truly positive data. This comparison also produces True Positive, True Negative and False Negative values [20].

4. Results

The results in this study are based on the methods that have been described. With the use of systematic calculations and model building on 3 existing machine learning algorithms, namely Decision Tree, SVM, and Naive Bayes, the following results can be presented. The data shown includes the results of recall, precision, accuracy and f1score calculations. Each result of the Decision Tree, SVM and naive bayes algorithms will be used as an indicator of the level of ability of each machine learning algorithm. After the algorithm results are available, the next step is to collect the results of the three algorithms into a comparison table so that they can be analyzed. This analysis looks at the highest accuracy, recall, precision and f1 score points because this research aims to compare algorithm results and classify data based on a feature. This section is important so that the tests that have been carried out can get results that can be used as a comparison to determine the best algorithm when used in diabetes datasets with glucose and blood pressure features

4.1. Naive Bayes

The Naive Bayes model provides performance with an accuracy of 92.96%. With the confusion matrix it produces 86 True Negative (TN), 99 True Positive (TP), 7 False Positive (FP), and 7 False Negative (FN). In the assessment of precision, recall and f1 score gets a value that is balanced with an order of 0.93 in each value. A high precision indicates that most of the positive predictions made by the model are indeed positive. Similarly, the recall reflects that the model is able to capture most of the positive cases in the data. The high F1-score value shows that the model managed to maintain a balance between precision and recall, which is important in the context of disease diagnosis, where misclassification can have serious consequences. The results in Naive Bayes show that the model is quite balanced in diagnosing diabetes. The naive bayes algorithm found that the confusion matrix assessment resulted in 86 True Negative (TN), 99 True Positive (TP), 7 False Positive (FP), and 7 False Negative (FN). it can be concluded that the model can recognize positive cases and can reduce or even avoid errors in predictions that have been made. But there are FP 7 and FP 7 values that need to be worried because of the possibility that the patient is not detected.

4.2. Support Vector Machine (SVM)

The Support Vector Machine model provides performance with an accuracy of 92.96%. This result is distinguished by the results of the confusion matrix where SVM produces 88 TN, 97 TP, 5 FP, and 9 FN, as well as in the assessment of precision, recall and f1 score gets a value in the order of 0.95, 0.92 and 0.93. The results of this SVM show that the model is also quite balanced in diagnosing diabetes like naive bayes. The SVM algorithm found that the confusion matrix assessment resulted in 88 TN (True Negative), 97 TP (True Positive), 5 FP (False Positive), and 9 FN (False Negative), so it can be concluded that the model can recognize positive cases in the predictions that have been made. But there is a value of FP 5 and FN 9 that needs to be worried because there is a possibility that the patient is not detected.

4.3. Decision Tree

The best performing model was obtained from the Decision Tree algorithm. This model achieved the highest accuracy of 94.47%, with a precision value of 0.96, recall of 0.93, and f1-score of 0.95 for the positive class. Based on the confusion matrix, the model correctly classified 89 non-diabetic and 99 diabetic data, and generated 4 FP and 7 FN. These results show that Decision Tree is more accurate in recognizing both classes and reduces misclassification than the other two models. The Decision Tree algorithm found that the confusion matrix assessment resulted in 89 TN (True

Negative) which is the level of not having diabetes, 99 TP (True Positive) as an indicator of people with diabetes, 4 FP (False Positive), and 7 FN (False Negative) as an indicator of undetected cases. From the results of the confusion matrix, the Decision Tree algorithm is the best algorithm in making predictions in this study. with a small error rate and able to detect cases well.

4.4. Comparasion Table and Matrix

Based on the results of performance testing of three classification algorithms, namely Decision Tree, Naïve Bayes, and Support Vector Machine (SVM), it is found that the Decision Tree algorithm shows the best performance in classifying diabetes disease data, as shown in Table 1. This is proven through the evaluation metrics used, where Decision Tree managed to obtain the highest accuracy and F1-score values compared to the other two algorithms. In addition, the algorithm is also able to maintain a balance between precision and recall, which is very important in the context of medical diagnosis such as diabetes detection. This balance shows that Decision Tree is not only good at identifying patients who actually have diabetes (recall), but also able to minimize errors in classifying patients who do not have diabetes as sufferers (precision). These advantages make Decision Tree the most optimal choice of model in the context of this study, because it is able to provide accurate and balanced classification results. This model also has the advantage that the prediction results can be easily understood. Thus, the results of the overall performance comparison of the classification algorithms tested, as summarized in Table 1, show that Decision Tree is feasible to be used as a tool in the early detection system of diabetes based on historical patient data.

Table 1. Evaluation Result Comparison

Model	*Accuracy*	*Precision*	*Recall*	*F1- Score*
Decision Tree	94,47%	96%	93%	95%
Support Vector Machine	92,96%	95%	92%	93%
Naive Bayes	92,96%	93%	93%	93%

From Table 1, the Decision tree algorithm obtained an accuracy of 94.47%, supported by precision (0.96) recall (0.93) and f1 score (0.95) which illustrates that this algorithm is balanced when used on this dataset. Meanwhile, the Naive Bayes and Support Vector Machine (SVM) algorithms have the same accuracy of 92.96%, but SVM excels in the precision section which is (0.95) while naive bayes is only (0.93). But Naive Bayes excels at the recall value which is (0.93) compared to SVM which gets (0.92). The overall Decision Tree was chosen as a model that provides balanced and accurate results on the dataset.

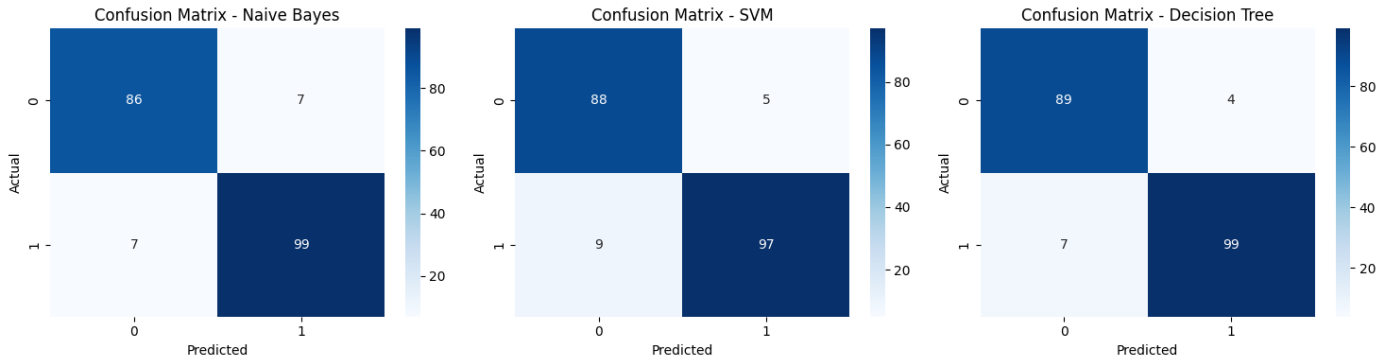

Figure 2 Comparison of Confusion Matrix

After testing, it can be seen in the matrix comparison image (Figure 2) that the Decision Tree Algorithm is superior to other algorithms, namely the Naive Bayes and SVM algorithms. With the FP, FN, TP, TN indicators, the data can be calculated statistically to find the value of Precision, recall and f1 score. So after seeing the comparison picture, it is found that Decision Tree is superior to the calculation of the confusion matrix indicator.

4.5. Discussion

The results of this study indicate that the three machine learning algorithms—Decision Tree, Support Vector Machine (SVM), and Naïve Bayes—demonstrated strong performance in diagnosing diabetes based on glucose and blood pressure data. Among them, the Decision Tree algorithm achieved the highest overall performance with an accuracy of 94.47%, outperforming both SVM and Naïve Bayes, which each reached 92.96%. This finding is consistent with previous research such as [7] and [9], which also found that Decision Tree models often perform well on structured medical datasets due to their ability to handle non-linear feature interactions and to produce interpretable results. The Decision Tree model also yielded the highest F1-score (0.95), reflecting a good balance between precision and recall. This balance is particularly important in the context of disease prediction, where minimizing false negatives (patients who actually have diabetes but are classified as non-diabetic) is crucial to avoid undetected cases and ensure early treatment.

The confusion matrix results further confirm the Decision Tree's superior performance, with the lowest number of misclassifications (4 false positives and 7 false negatives) compared to SVM (5 false positives and 9 false negatives) and Naïve Bayes (7 false positives and 7 false negatives). This means that the Decision Tree model was more capable of correctly identifying both diabetic and non-diabetic cases. Although SVM exhibited slightly higher precision (0.95) than Naïve Bayes (0.93), its recall (0.92) was slightly lower, showing that it was more precise but less sensitive in capturing all true diabetic cases. Naïve Bayes, on the other hand, displayed stable and balanced results with an accuracy of 92.96%, a precision of 0.93, and a recall of 0.93, suggesting that even a simple probabilistic model can still perform reliably in this context. These results collectively indicate that while all three models are suitable for diabetes prediction, the Decision Tree offers the best trade-off between interpretability, sensitivity, and predictive accuracy. This finding reinforces the practicality of using Decision Tree models in medical diagnostic applications, where transparency and interpretability are as essential as accuracy.

4.6. Limitation

Despite the encouraging results, several limitations should be acknowledged in this study. The first limitation lies in the dataset used, which contained only two key features: glucose and blood pressure. In actual clinical scenarios, diabetes diagnosis depends on a much wider range of variables, including age, body mass index (BMI), insulin levels, family medical history, and lifestyle factors. Limiting the model to only two variables restricts its ability to capture the complex patterns and interactions that exist in real-world diabetes prediction. Moreover, the dataset size used for training and testing was relatively small, which may limit the model's generalizability and increase the risk of overfitting.

Another limitation is the use of default hyperparameters for all three algorithms without detailed optimization. Techniques such as Grid Search or Random Search could have been employed to find the most effective parameter combinations, especially for SVM, which is known to be sensitive to its kernel type and regularization parameters. Additionally, the dataset used in this study was relatively clean and balanced, while real-world health data often contain missing values, noise, and class imbalances. Therefore, the results presented here may represent an optimistic estimate of model performance compared to actual medical datasets. Finally, the study focused solely on traditional machine learning models and did not include ensemble or deep learning methods, which could potentially enhance predictive performance in more complex classification tasks.

4.7. Future Research Suggestions

For future research, several directions are suggested to enhance both the robustness and applicability of machine learning models for diabetes prediction. First, future studies should expand the feature set to include more physiological and behavioral indicators such as BMI, insulin levels, cholesterol, dietary habits, and family medical history. Incorporating more diverse datasets from various demographic groups would also improve model generalization and ensure that the model can perform well in broader clinical contexts. Second, researchers should consider implementing hyperparameter optimization methods such as Grid Search, Random Search, or Bayesian Optimization to fine-tune algorithm parameters. This approach can significantly improve performance, particularly for SVM and Decision Tree, by identifying the most efficient configuration for classification tasks.

 ISSN 2579-7069

In addition, future studies should explore the use of advanced machine learning and deep learning techniques such as Random Forest, Gradient Boosting, XGBoost, and Convolutional Neural Networks (CNNs). These models may yield higher accuracy and better generalization in handling non-linear and high-dimensional medical data. Furthermore, integrating these predictive models into real clinical decision-support systems would allow for real-world validation and testing of their effectiveness in assisting healthcare professionals. Finally, future research should also focus on explainable AI (XAI) approaches such as SHAP or LIME to improve model transparency and interpretability, which are vital for ethical and trustworthy use in healthcare. Incorporating fairness, privacy, and bias assessment into future studies would also ensure that machine learning applications for medical diagnosis remain responsible, transparent, and equitable.

5. Conclusion

This study was conducted to compare the performance of three machine learning algorithms—Naïve Bayes, Support Vector Machine (SVM), and Decision Tree—in classifying diabetes based on glucose and blood pressure data. The research employed systematic data processing and evaluation using several performance metrics, including accuracy, precision, recall, and F1-score. The results revealed that the Decision Tree algorithm achieved the highest accuracy of 94.47%, outperforming Naïve Bayes and SVM, which each obtained 92.96%. These findings demonstrate that the Decision Tree model is more effective in identifying patterns within the dataset, successfully distinguishing between diabetic and non-diabetic cases with higher precision and fewer classification errors.

The superior performance of the Decision Tree algorithm can be attributed to its ability to handle non-linear relationships and interpret hierarchical decision rules efficiently. The confusion matrix analysis showed that the Decision Tree produced the lowest number of false positives and false negatives, indicating that it performs well in both identifying positive cases and avoiding misclassification of healthy individuals. The balanced precision and recall values also suggest that the model is not biased toward a specific class, making it reliable for medical classification tasks. The strong and consistent results of all three algorithms further validate the applicability of machine learning techniques in medical diagnosis, where they can assist healthcare professionals in processing complex patient data and generating data-driven insights for early disease detection.

Based on these results, the Decision Tree algorithm can be considered a robust and interpretable model for predicting diabetes using clinical parameters such as glucose and blood pressure. Its high performance makes it a suitable foundation for developing automated diabetes detection systems that can support medical decision-making and improve diagnostic accuracy. The findings of this study are expected to contribute to future research and practical applications of artificial intelligence in the healthcare sector, particularly in disease prediction, prevention, and personalized treatment strategies. Furthermore, this study reinforces the importance of integrating machine learning methods into clinical systems to enhance early diagnosis, reduce diagnostic errors, and optimize patient care in the era of digital health.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, S.D.N., L.K.O., and H.T.S.; Methodology, S.D.N. and L.K.O.; Software, L.K.O. and H.T.S.; Validation, L.K.O. and H.T.S.; Formal Analysis, S.D.N.; Investigation, L.K.O. and H.T.S.; Resources, L.K.O. and H.T.S.; Data Curation, L.K.O.; Writing—Original Draft Preparation, S.D.N.; Writing—Review and Editing, L.K.O. and H.T.S.; Visualization, H.T.S. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement
Not applicable.

6.5. Informed Consent Statement
Not applicable.

6.6. Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Leonard Baatiema et all, "Contextual awareness, response and evaluation (CARE) of diabetes in poorurban communities in Ghana: the CARE diabetes project qualitative studyprotocol," *Global Health Action,* 2024.

[2] A. R. B. N. Bin Zhou, "Worldwide trends in diabetes prevalence and treatment," *The Lancet 2024; 404: 2077–93,* pp. 2077-2093, 2024.

[3] A. Timbadiya, "Machine learning algorithms for healthcare," *World Journal of Advanced Research and Reviews, 2025, 25(02),,* pp. 1139-1143, 2025.

[4] B. J. T. K. M. K. P. R. Shikha Agarwal, "Hybrid of Naive Bayes and Gaussian Naive Bayes for Classification: A Map," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* pp. 266-268, 2019.

[5] A. G. G. A. K. I. S. Preeti Voditel, "Foetal Health Prediction using Random Forest and Support Vector Machine," *Journal of Neonatal Surgery,* pp. 574-582, 2025.

[6] V. H. S. D. S. H. N. S. R. S. D. M. E. D. S. R. S. S. Beatriz N. C. Silveira, "Advancing Test Data Selection by Leveraging Decision TreeStructures: An Investigation into Decision Tree Coverage and Mutation Analysis," *Journal of Software Engineering Research and Development,* 2025.

[7] N. A. H. N. M. C. ,. M. A. A. N. R. A. C. H. P. A. S. Amanda Iksanul Putri, "Implementation of K-Nearest Neighbors, NaïveBayes Classifier, Support Vector Machineand Decision Tree Algorithms for Obesity Risk Prediction," *Institute ofResearch and Publication Indonesia, Public Research Journal of Engineering, Data Technology and Computer Science,* pp. 26-33, 2024.

[8] A. F. P. L. L. B. Dewi Widyawati, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indonesian Journal of Data and Science,* pp. 78-87, 2023.

[9] O. D. R. D. A. S. A. J. Admi Syarif, "A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification: Decision Tree, Support Vector Machine, and Naive Bayes," *International Journal of Advanced Computer Science and Applications,* pp. 597-603, 2023.

[10] N. L. N. S. Dhurba Subedi, "Sentiment Analysis of IMDb Movie Reviews Using SVM and Naive Bayes Classifier," *Journal of Engineering and Sciences,* pp. 56-68, 2025.

[11] D. S. K. Vaishali Mishra, "The performance of Logistic Regression, Decision Tree, KNN, Naive Bayes and SVM for identifying Automotive Cybersecurity Attack and Prevention: An Experimental Study," *Journal of Electrical Systems,* pp. 687-699, 2024.

[12] L. W. R. I. Sabda Norman Hayat, "Skin Cancer Detection Approach Using Convolutional Neural Network Artificial Intelligence," *International Journal of Informatics and Information Systems,* pp. 46-54, 2024.

[13] D. P. Yasodha, "Data Preprocessing Methods for Machine Learning: An Empirical Comparison," *International Journal for Multidisciplinary Research,* 2025.

[14] Z. H. E. W. Shuo Zhang, "Data Cleaning Using Large Language Models," *Cornell University Computer Science,* 2024.

[15] I. Olaniyi, "Ideal Dataset Splitting Ratios In Machine Learning Algorithms: General Concerns For Data Scientists And Data Analysts," *International Mardin Artuklu Scientific ,* 2022.

[16] M. H. A. C. L. M. W. V. Karim Hossny, "Decision tree insights analytics (DTIA) tool: an analytic framework to identify insights from large data records across fields of science," *Machine Learning: Science and Technology,* 2024.

[17] B. K. G. A. S. M. S. B. V. M. H. A. A. S. A. A. S. G. Ajith Abraham, "Naïve Bayes Approach for Word Sense Disambiguation System With a Focus on Parts-of-Speech Ambiguity Resolution," *IEEE Access, vol. 12,,* 2024.

[18] P. R. S Shanmugapriya, "Adaptive Rhea Optimization Enhanced CNN SVM Framework (ROC-SVM) for Precise MRI Based Brain Tumor Classification," *INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY,* p. 1939–1952, 2025.

[19] A. I. H. F. R. U. Hovi Sohibul Wafa, "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)," *INFORMATICS AND DIGITAL EXPERT,* pp. 40-45, 2022.

[20] K. B. S. S. Subhajeet Das, "Performance Analysis of Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest and SVM on Hate Speech Detection from Twitter," *International Research Journal of Innovations in Engineering and Technology,* 2023.