Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 5, No. 2, May 2026, pp. 198-206
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI : https://doi.org/10.69916/jkbti.v5i2.473

COMPARATIVE ANALYSIS OF PERFORMANCE OF MACHINE LEARNING FEATURE SELECTION IN EARLY DETECTION OF DIABETES

Lilik Joko Susanto*[1], Agus Wantoro[2]

[1]Department of Education Informatics, STKIP Rosalia Metro, Indonesia
[2]Department of Informatics Engineering, Faculty of Technology and Informatics, Universitas Aisyah Pringsewu, Pringsewu, Indonesia
Email: [1]lilikjoko09@gmail.com, [2]aguswantoro@aisyahuniversity.ac.id

(Received: April 30, 2026; Revised: May 5, 2026; Published: May 9, 2026)

Abstract

Diabetes is one of the most serious global health problems and continues to increase significantly worldwide. Early detection is essential to reduce complications and improve patient survival rates. Recently, Machine Learning (ML) has shown great potential in supporting early diabetes prediction through data-driven analysis. However, the presence of irrelevant and redundant features may decrease model performance and increase computational complexity. Therefore, this study aims to evaluate the effectiveness of feature selection techniques and ML algorithms for early diabetes detection using the PIMA Indians Diabetes Dataset. The dataset consists of 768 records, 8 features, and two classes. Data preprocessing was conducted to handle missing values and outliers using mean imputation and data cleaning techniques. Three feature selection methods were applied, namely Information Gain (IG), Gain Ratio (GR), and ANOVA, to identify the most relevant features. Furthermore, several ML algorithms, including k-Nearest Neighbor (k-NN), Random Forest, Support Vector Machine (SVM), Naive Bayes, and Neural Network, were evaluated using 10-fold cross-validation. The results showed that feature selection techniques improved classification performance compared to using all features. Glucose, BMI, Age, and Insulin were identified as the most influential features in diabetes prediction. Among all evaluated models, Random Forest combined with ANOVA achieved the best performance with an accuracy of 0.753. In general, the application of feature selection techniques increased model accuracy by up to 3.82%. These findings demonstrate that combining effective feature selection methods with robust ML algorithms can significantly enhance the performance of early diabetes detection systems.

Keywords: diabates; machine learning; feature selection.

1. INTRODUCTION

Diabetes is a serious global health problem, ranking among the leading causes of internal disease and diabetes-related deaths among women worldwide [1]. Diabetes is one of the main causes of diabetes deaths among women globally [2]. Recent data shows that the incidence of Diabetes continues to increase, making it the most frequently detected type of diabetes. The main cause of death in diabetes is blood sugar complications and the presence of other accompanying diseases, especially cardiovascular disease [3]. Early detection is a fundamental key in increasing survival rates and treatment success. When early detection is carried out, the patient's chance of recovery is significantly higher compared to cases detected in type 2 diabetes[4].

In recent years, rapid advances in the field of Machine Learning (ML) have paved the way for the development of more accurate and efficient early diabetes detection systems. [5]. ML algorithms can analyse medical data, including genetic data, medical images, and clinical history, to identify complex patterns that may not be visible to the human eye. The potential of ML is able to increase the accuracy of detection and predict disease risk in various studies [6].

However, the application of ML in early diabetes detection faces two main challenges, namely feature selection. Pima Indians Diabetes Dataset is often used as a basis for evaluating models for early detection of diabetes mellitus [7]. This dataset has eight features, where all features need to be analysed, whether they are relevant or not to model accuracy. The presence of irrelevant features can lead to decreased model performance, increased computational complexity, and overfitting [8]. Therefore, effective feature selection techniques are essential to identify the most informative subset of features that have an influence on model performance and reduce noise [9]. Integration between effective feature selection and ML models is capable of early detection of diabetes [10]. However, there is still a need for a comprehensive evaluation of the combination of various feature selection techniques and ML models to identify the best approach in this context

This research aims to carry out performance-based analysis of various feature selection methods and class imbalance handling techniques in the context of early detection of diabetes using ML such as k-NN, Tree, Naive Bayes, and Random Forest. By comparing the performance of various combinations of these methods, it is hoped that the most optimal approach can be found to build an accurate diabetes early detection model, thus contributing to increasing patient survival.

2. RESEARCH METHODS

This stage outlines the procedures followed when conducting research. The research stage begins with data collection. Before the data is entered into the model, it is necessary to pre-process the data. At this stage, the data is cleaned and adjusted for processing to the next step of feature selection using Information Gain (IG), Gain Ratio (GR), and ANOVA techniques. Next, the selected features are used for Diabetes diabetes classification using ML algorithms such as k-NN, Tree, Naive Bayes, Random Forest. The framework design is illustrated in Figure 1.

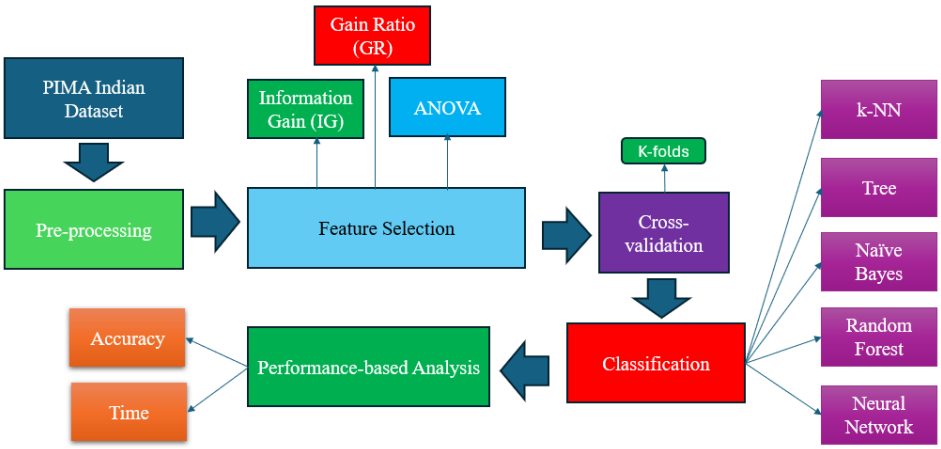


Figure 1. Framework design

After classifying diabetes, a comparative analysis is then carried out by calculating the accuracy value and the computing time required for the ML algorithm to build the model

3.1. Diabetes Dataset

This study uses the PIMA Indians dataset taken from www.kaggle.com/uciml/pima-indians-diabetes-database. The dataset has 768 records, 8 features, and 2 classes [11]. The amount of data in each class is shown in Table 1.

Table 1. Number of classes

Diabetes	Non-diabetes
268	500

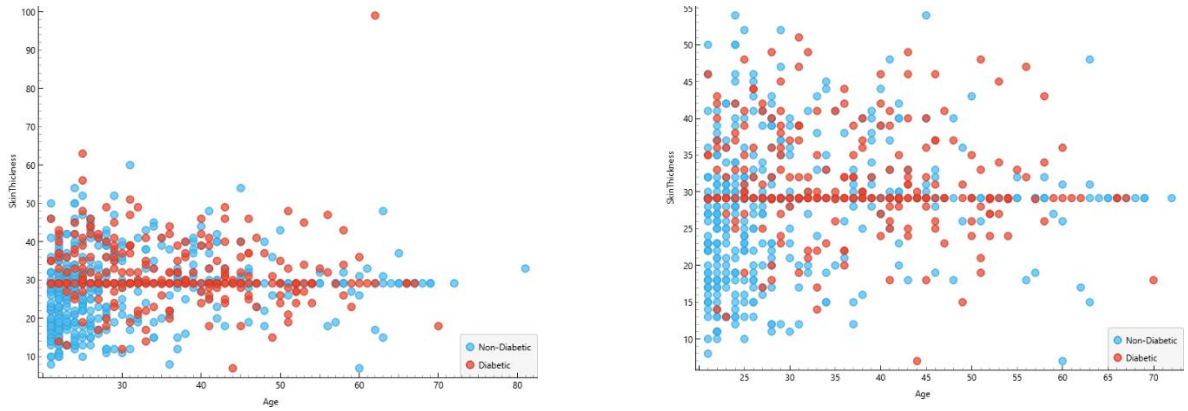
The features used in this dataset are presented in Table 2.

Table 2. Feature, and attribute dataset

ID	Feature	Information
f1	Pregnancies	Number of pregnancies
f2	Glucose Plasma	Plasma glucose levels two hours after consuming glucose
f3	Blood Pressure	Diastolic blood pressure (mm Hg)
f4	Skin	Skinfold thickness of upper arm triceps (mm)
f5	Insulin	Serum insulin levels in the blood two hours after the glucose test
f6	BMI	Body mass index kg (Height in m)
f7	Pedigree	A value that measures genetic risk factors based on a family history of diabetes
f8	Age	Patient's age
f9	Class	Class

3.2. Pre-processing

The crucial process includes data cleaning to deal with missing values or outliers. This dataset has 10.6% missing values, so improvements need to be made. We apply charging using the technique of using average values. Apart from that, we make improvements to data that has outliers. This can improve the accuracy, efficiency and effectiveness of the model. The results of the improved dataset are shown in Figure 2.



(a) Data outlier (b) Data inliers

Figure 2. Data outlier dan inliers

3.3. Feature Selection

One of the most important things in classification is determining features to get the best performance results [13]. Datasets used in ML processes usually contain redundant and irrelevant features, and do not improve accuracy [14], has no effect on the learning model, and may even degrade the performance of the learning model [15], therefore it is important to perform relevant feature analysis. In this case we apply the Information Gain, Gain Ratio, and ANOVA feature selection algorithms. These three techniques have good capabilities in selecting features by giving weight to each feature and displaying them in ranking[16].

a. Information Gain (IG)

IG measures the reduction in entropy (uncertainty) in a data set after splitting based on certain attributes. The higher the GI, the better the attribute is at classifying data [17]. Decision tree induction is the foundation of this method. Information gain is used as an attribute selection criterion. The information acquisition method has a faster time in the feature selection process compared to other methods. Features with the most information will be ranked high in this method; otherwise, features with the most information will be ranked high [18]

b. Gain-Ratio (GR)

GR is a modification of IG designed to overcome this bias. GR normalizes IG by dividing the value of the attribute in question. Split Information measures how evenly or unevenly the attribute divides the data[13]. When selecting features, GR considers the number and size of datasets [19]. GR modifies information acquisition which can reduce its bias. GR selects attributes based on the number and size of branches. With this normalization, GR tends to select attributes that produce more balanced and meaningful separations, not just those that have many unique categories. This leads to better generalization of the model (reduces overfitting).

3.4. ANOVA

ANOVA feature selection is a method in machine learning to select the most relevant features by analysing the variance between features and target classes [20]. This method compares variability between classes with variability within each class to determine which features have the greatest discriminating power. Features with higher statistical (F) values are considered more effective and more likely to be included in the model [21].

3.5. Evaluation Performance Machine Learning

This study examines how the Confusion Matrix can be used to measure accuracy and error rates. Confusion matrix is a table used to evaluate the performance of classification models in ML. This table compares the model's predicted results with the actual (actual) values from the data, so it can provide an in-depth picture of the advantages and disadvantages of the mode [22]. The confusion matrix variable is displayed in Table 3.

Table 3. Confusion Matrix

Class	Positive Prediction	Negative Prediction
Positive Actual	Number of True Positive (TP)	Number False Negative (FN)
Negative Actual	Number of False Positive (FP)	Number True Negative (TN)

Accuracy is a method for evaluating the performance of ML algorithms. These variables can be obtained from the Confusion Matrix in Table 3 and calculated using equation (1).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

3.6. Cross-validation (k-fold)

Classification is a form of data mining technique that is currently popular [23]. This strategy uses various methods to assess available data to produce diabetes predictions [13]. The classification model will be validated using k-fold cross-validation. The cross-validation method is generally used for training sets [24]. Figure 3 displays the cross-validation procedure

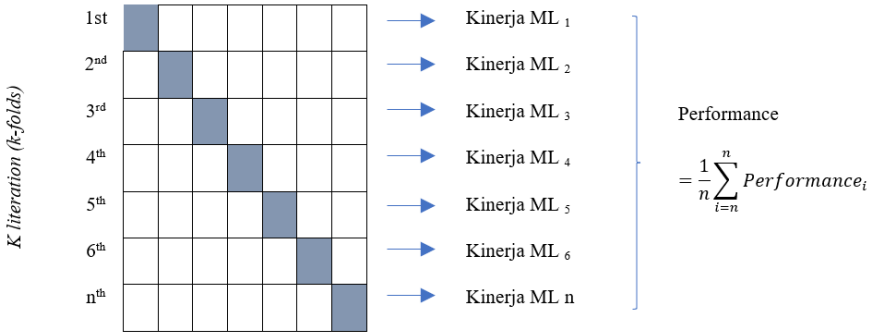


Figure 3. Procedure of k-fold Validation

3. RESULTS AND DISCUSSION

3.1. Feature Selection Evaluation

The process of evaluating feature selection techniques and testing model performance was carried out using software (Orange 3.39.0). The test design is shown in Figure 4.

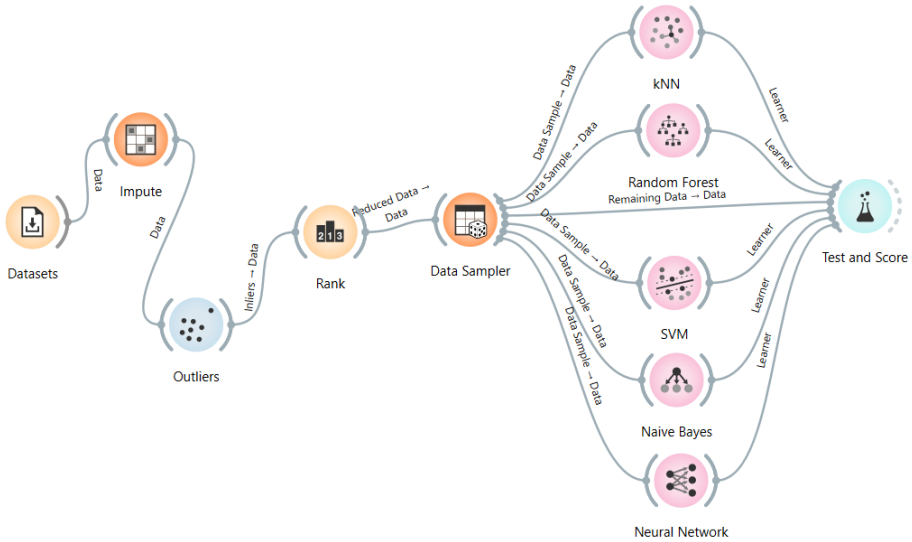


Figure 4. Design for testing feature selection techniques and ML models

Based on this figure, the features that have the most influence on the class and have the highest weight are obtained. The feature selection results are shown in Table 4.

Table 4. Ranking weighting (w) feature selection techniques (f)

Information Gain (IG)		Gain Ratio (GR)		ANOVA	
f	**w**	**f**	**w**	**f**	**w**
Glucose	0.176	Glucose	0.088	Glucose	239.09
Age	0.089	Age	0.044	BMI	70.8
BMI	0.067	BMI	0.034	Age	53.39
Insulin	0.057	Insulin	0.033	Insulin	43.53

The selection results using three feature selection techniques in Table 6 produce features with different rankings. The three feature selection techniques produce the same features, but the IG and GR techniques produce features with the same weighting, while ANOVA produces features with different weights. Next, we carry out a performance comparison using all features and feature selection. We applied k-fold=10, because of the test results, we found that this value yielded the best performance against the model. The comparison test results are shown in Table 5.

Table 5. Performance comparison of ML algorithms using feature selection

ID	Model	All Fitur	IG	GR	ANOVA
1	k-NN	0.632	0.665	0.665	0.665
2	Random Forest	0.709	0.742	0.720	**0.753**
3	SVM	0.604	0.621	0.621	0.621
4	Naive Bayes	0.670	0.687	0.687	0.687
5	Neural Network	0.687	0.714	0.714	0.703
Average		**0.660**	**0.681**	**0.686**	**0.686**

We found differences in the accuracy of each ML algorithm's performance. In this case, the algorithm performance is more optimal using the IG technique. This is different from other research [13] which uses the same approach. Based on the average accuracy, although not significant, there was an increase in accuracy after feature selection, except for the Neural Network (NN) algorithm. The application of the IG technique increased accuracy by 3.82%, GR by 3.16%, and ANOVA by 3.82%. In addition, we found that the Random Forest algorithm with ANOVA technique is the best combination. The performance of the algorithm based on the feature selection technique is shown in Figure 5.

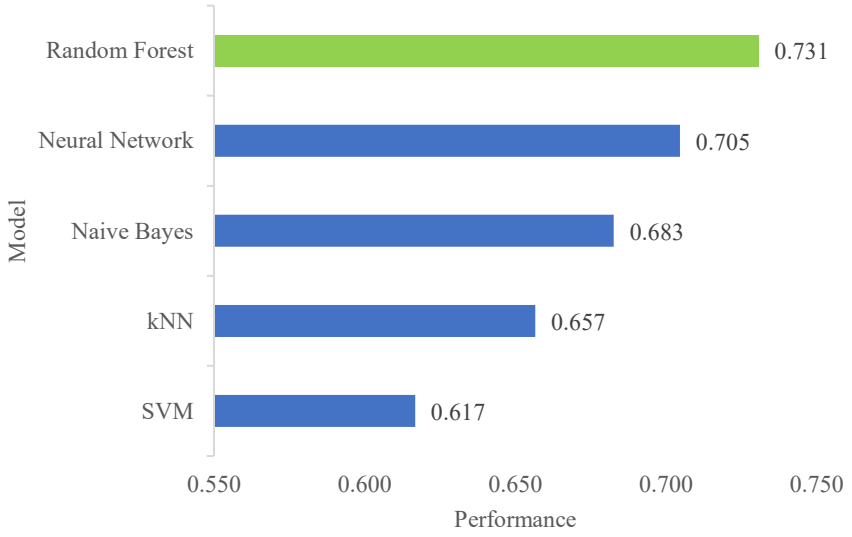


Figure 5. Performance (accuracy) of ML algorithms based on feature selection

Figure 5 shows that the Random Forest model shows the best performance, followed by the Neural Network model. Meanwhile, SVM and k-NN models show the worst performance. This is because the Random Forest model is resistant to overfitting and its ability to handle complex data. By combining many decision trees, this model is stable, strong against outliers, and able to work well on non-linear data and missing data. Next, we carried out a comparison test based on the precision variable shown in Figure 6.

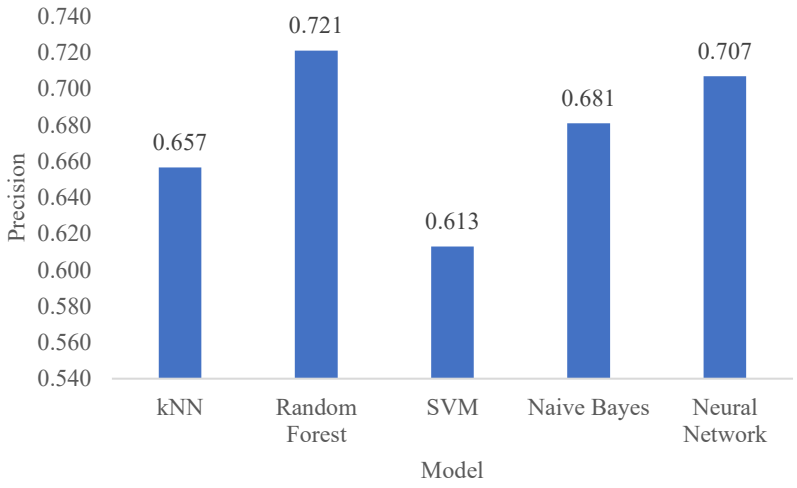


Figure 6. Comparison of model performance based on precision

Figure 6 shows that the Random Forest model shows the best performance, followed by the Neural Network (NN) model based on variable precision. Meanwhile, SVM and k-NN models show the worst performance. Next, we carried out a comparison test based on the recall variable shown in Figure 7.

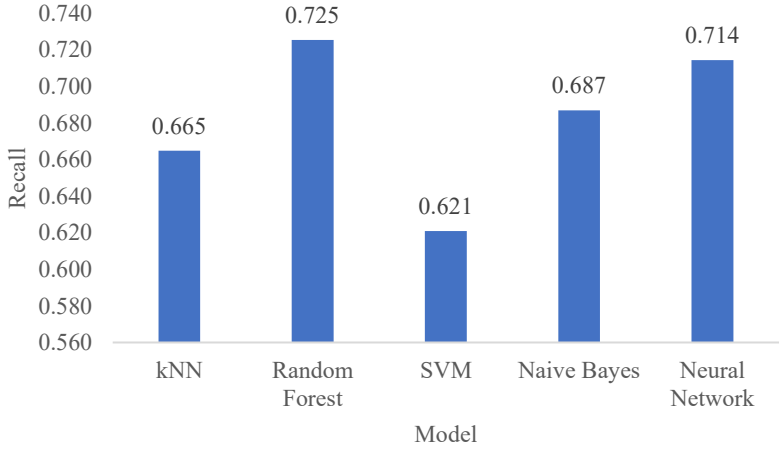


Figure 7. Comparison of model performance based on recall

Next, we conducted a comparative analysis of the performance of the three feature selection techniques using the average accuracy values shown in Figure 8.

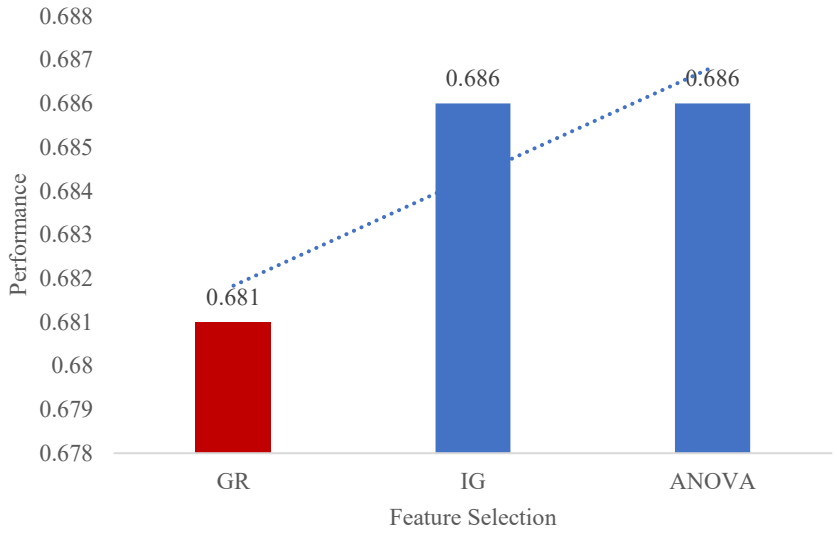


Figure 8. Feature selection performance comparison

Figure 6 shows the performance of different selection techniques. ANOVA and IG techniques performed best in this case, whereas GR performed poorly. These findings indicate that the choice of feature selection techniques and ML models is an important step in the case of medical classification.

3.2. Discussion

The discussion of this study demonstrates that the implementation of feature selection techniques significantly influences the performance of ML algorithms in diabetes prediction. The results indicate that feature selection methods, namely IG, GR, and ANOVA, were able to improve classification accuracy compared to the use of all features. This finding confirms that not all features in the PIMA Indians Diabetes Dataset contribute equally to the prediction process. Irrelevant and redundant features may reduce model effectiveness, increase computational complexity, and potentially cause overfitting. Among the evaluated features, Glucose consistently appeared as the most influential feature across all feature selection methods, indicating that blood glucose level is the strongest indicator for diabetes classification. Other important features such as BMI, Age, and Insulin also showed significant contributions to model performance, which aligns with medical understanding that obesity, aging, and insulin abnormalities are closely related to diabetes risk.

Furthermore, the Random Forest algorithm achieved the best overall performance, particularly when combined with the ANOVA feature selection technique. This result highlights the robustness of Random Forest in handling non-linear relationships, noisy data, and complex feature interactions. The ensemble mechanism in Random Forest enables the model to reduce overfitting and improve generalization performance compared to single classifiers such as k-NN or SVM. In contrast, SVM and k-NN produced lower accuracy values, indicating that these algorithms may be less suitable for datasets with imbalanced distributions and complex patterns such as diabetes data. Although Neural Network also showed competitive performance, the improvement after feature selection was not as significant as Random Forest, suggesting that NN may require larger datasets and more optimized parameter tuning to achieve maximum performance.

The comparison among feature selection techniques also revealed interesting findings. IG and ANOVA provided the highest average accuracy improvements, while GR produced slightly lower performance. This suggests that entropy-based and variance-based feature evaluation methods are more effective for identifying discriminative features in diabetes datasets. The superior performance of ANOVA may be attributed to its capability to statistically measure differences between classes, enabling the model to focus on features with strong discriminative power. Overall, the findings of this study emphasize that the combination of appropriate feature selection techniques and robust ML algorithms is essential for developing accurate and reliable early diabetes detection systems. These results can contribute to the development of intelligent medical decision-support systems that assist healthcare professionals in identifying diabetes risk more effectively and efficiently.

4. CONCLUSION

Based on the results of the comparative analysis that has been carried out, it can be concluded that the IG, GR, and ANOVA feature selection techniques in the PIMA Indians Diabates Dataset have an impact on the performance of the classification model. The IG and GR techniques produce four features that have the highest weight, namely

Glucose, Age, BMI, and Insulin. This is different from selection using the ANOVA technique which produces features namely Glucose, BMI, Age and Insulin. Based on the selected features, we evaluate the ML algorithm to obtain comprehensive information on the algorithm's performance. Results of the performance comparison test of the three FS techniques, we found that the IG and ANOVA techniques showed the best performance.

Apart from that, the results of the evaluation of the ML model showed that the performance of the model was different. In this study, the Random Forest model showed optimal performance, followed by Neural Network (NN), while the SVM model showed the worst performance. This is because Random Forest is very reliable and flexible, excels in high accuracy, in handling large data with complex features, and is resistant to overfitting. Its main advantages include prediction stability because it uses the average of many decision trees, the ability to handle missing values, and robustness to outliers

Class imbalance handling strategies and feature selection must be an integral part of the medical classification system development pipeline, because they can affect the overall interpretability and detection accuracy. Further research is recommended to evaluate the effectiveness of other imbalance handling techniques such as ensemble-based sampling or cost-sensitive learning, as well as applying these approaches to larger and more complex clinical datasets.

5. ACKNOWLEDGMENTS

The author would like to express his deepest gratitude to the Faculty of Health, Aisyah Pringsewu University (UAP) and STMIK Rosalia Metro, for helping validate the anonymous breast diabetes patient data used in this research. We also thank the Machine Learning (ML) Research Group at UAP and STKIP Metro for their valuable input during the model development and evaluation phase. Special awards are given to doctors whose clinical insight contributes significantly to the interpretation of breast diabetes prediction results

REFERENCES

[1] E. Decroli, *Diabetes Melitus Tipe 2*, Pertama. Padang, Sumatera Barat: Fakultas Kedokteran Universitas Andalas, 2019.

[2] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Appl. Intell.*, vol. 52, no. 3, pp. 2411–2422, 2022, doi: 10.1007/s10489-021-02533-w.

[3] M. Al Switi, B. Alshraideh, A. Alshraideh, A. Massad, and M. Alshraideh, "Treatment of diabetes type II using genetic algorithm," *Int. J. online Biomed. Eng.*, vol. 15, no. 11, pp. 53–68, 2019, doi: 10.3991/ijoe.v15i11.10751.

[4] J. Lindstrom and J. Tuomilehto, "International Diabetes Federation - IDF Complications Congress 2020," International Diabetes Federation. Accessed: Apr. 03, 2020. [Online]. Available: https://www.idf.org/our-activities/congress/idf-complications-congress-2020.html?gclid=EAIaIQobChMIzvimsLDL6AIVzBErCh3qnQb7EAAYASAAEgL6lfD_BwE

[5] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," *IEEE Access*, vol. 9, pp. 103737–103757, 2021, doi: 10.1109/ACCESS.2021.3098691.

[6] B. Mahesh, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/art20203995.

[7] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," 2023. doi: 10.3390/diagnostics13142383.

[8] S. S. Rautaray, S. Dey, M. Pandey, and M. K. Gourisaria, "Nuclei segmentation in cell images using fully convolutional neural networks," *Int. J. Emerg. Technol.*, vol. 11, no. 3, pp. 731–737, 2020, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087307723&partnerID=40&md5=4356c05335e3af18806af217ccfe5d55

[9] X. Song *et al.*, "Evolutionary computation for feature selection in classification: A comprehensive survey of solutions, applications and challenges," *Swarm Evol. Comput.*, vol. 90, p. 101661, 2024, doi: https://doi.org/10.1016/j.swevo.2024.101661.

[10] L. K. Singh, M. Khanna, and R. Singh, "Efficient feature selection for breast diabetes classification using soft computing approach: A novel clinical decision support system," *Multimed. Tools Appl.*, vol. 83, no. 14, pp. 43223–43276, 2024, doi: 10.1007/s11042-023-17044-8.

[11] C. Sharma and A. Singla, "Advanced PTSVM Based Breast Diabetes Classification with Weighted Feature Selection," *SN Comput. Sci.*, vol. 6, no. 1, p. 50, 2024, doi: 10.1007/s42979-024-03590-x.

[12] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 4, pp. 530–535, 2019, doi: 10.1016/j.cegh.2018.12.004.

[13] H. Sulistiani, A. Syarif, K. Muludi, and Warsito, "Performance evaluation of feature selections on some ML approaches for diagnosing the narcissistic personality disorder," *Bull. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 1383–1391, 2024, doi: 10.11591/eei.v13i2.6717.

[14] J. Wang, S. Zhou, Y. Yi, and J. Kong, "An improved feature selection based on effective range for classification," *Sci. World J.*, vol. 2014, 2014, doi: 10.1155/2014/972125.

[15] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2019, pp. 619–623. doi: 10.1109/IBCAST.2019.8667106.

[16] W. Van Casteren, "The Waterfall Model And The Agile Methodologies : A Comparison By Project Characteristics-Short The Waterfall Model and Agile Methodologies," *Acad. Competences Bachelor*, no. February, pp. 10–13, 2017, [Online]. Available: https://www.researchgate.net/publication/313768860

[17] J. Gao, Z. Wang, T. Jin, J. Cheng, Z. Lei, and S. Gao, "Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection," *Knowledge-Based Syst.*, vol. 286, p. 111380, 2024, doi: https://doi.org/10.1016/j.knosys.2024.111380.

[18] P. Bhat and K. Dutta, "A multi-tiered feature selection model for android malware detection based on Feature discrimination and Information Gain," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, Part B, pp. 9464–9477, 2022, doi: https://doi.org/10.1016/j.jksuci.2021.11.004.

[19] M. Trabelsi, N. Meddouri, and M. Maddouri, "A New Feature Selection Method for Nominal Classifier based on Formal Concept Analysis," *Procedia Comput. Sci.*, vol. 112, pp. 186–194, 2017, doi: 10.1016/j.procs.2017.08.227.

[20] E. Taghizadeh, S. Heydarheydari, A. Saberi, S. JafarpoorNesheli, and S. M. Rezaeijo, "Breast diabetes prediction with transcriptome profiling using feature selection and machine learning methods," *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–9, 2022, doi: 10.1186/s12859-022-04965-8.

[21] V. Vijayasarveswari *et al.*, "Development of Statistically Modelled Feature Selection Method for Microwave Breast Diabetes Detection," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 50, no. 1, pp. 250–263, 2025, doi: 10.37934/araset.50.1.250263.

[22] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1806–1819, 2017, doi: 10.1109/TKDE.2017.2682249.

[23] F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, *Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques*. London, 2019. doi: 10.1007/979-981-13-8798-2-12.

[24] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1292–1303, 2022, doi: https://doi.org/10.1016/j.jrmge.2022.03.002.