Jurnal Sains Informatika Terapan (JSIT)

E-ISSN: 2828-1659, Volume: 4, Issue: 3, Year: 2025
Page: 674 - 681, Available online at: https://rcf-indonesia.org/home/

PATIENT PREVENTION PREDICTION AND DIAGNOSIS USING DATA MINING IN HEALTHCARE QUALITY MANAGEMENT

Noviyanty[1], Juvinal Ximenes Guterres[2], Adozinda Soares Gusmao[3], Domingas Soares[4] ,Anita Guterres[5,] Recardina Freitas da Silva[6]

[1,2,3,4,5,6] Faculty of Engineering, Universidade Oriental Timor Lorosae UNITAL.
Corresponding Author: [1] juvinalximenes6@gmail.com

Article Info

Article history:

Received: Nov 19, 2025
Revised: Nov 28, 2025
Accepted: Des 02, 2025
Published: Des 03, 2025

Keywords:

Machine Learning, Random Forest, Disease Prediction, Model Evaluation, Medical Decision Support System

ABSTRACT

The expansion of digital medical records and clinical data has strengthened the development of intelligent analytical systems to support early disease detection and improve diagnostic accuracy. This study aims to evaluate the performance of three classification algorithms, namely Random Forest, Support Vector Machine, and Logistic Regression, in predicting stroke risk using multidimensional patient clinical information. The dataset consists of 224 patient records derived from the Kaggle Stroke Dataset and additional questionnaire data collected from hospitals and primary health centers. The variables include demographic characteristics, clinical history, lifestyle factors, and physiological indicators. The research methodology involves several stages, including data preprocessing, feature selection using ANOVA F value, class balancing through the Synthetic Minority Oversampling Technique, model training, and performance evaluation using Accuracy, Precision, Recall, F1 Score, Matthews Correlation Coefficient, and Area Under the Curve. The results indicate that the Random Forest model achieves the highest performance, with an accuracy of 0.91 and an Area Under the Curve of 0.91, outperforming Support Vector Machine and Logistic Regression. This outcome confirms the effectiveness of ensemble based approaches in identifying complex nonlinear patterns and managing imbalanced data. The study contributes to healthcare quality improvement by providing a reliable prediction framework that supports early clinical decision making, reduces diagnostic delays, and enhances patient care outcomes.

1. INTRODUCTION

Delays in the diagnosis of chronic and acute diseases remain a serious problem in the global healthcare system. In the context of stroke, delayed diagnosis often has a fatal impact on patient safety, as every minute of delay can result in the loss of millions of neurons in the brain. Stroke is now the second leading cause of death and a leading cause of long-term disability worldwide.[1]The World Stroke Organization's Global Stroke Fact Sheet 2023 reports that more than 12 million new stroke cases occur annually globally, and approximately 6.5 million deaths are attributed to the disease. Approximately 70% of stroke cases occur in low- and middle-income countries, where access to early diagnosis and treatment remains very limited.[2], [3]In many countries, including in Asia, delayed stroke diagnosis is caused by various factors, such as limited medical facilities, low public awareness of early symptoms, inadequate training of primary health care workers, and the lack of a digital data-based triage system.[4], [5]In emergency clinical situations, the ability of healthcare providers to identify early signs of stroke, such as hemiparesis, speech disturbances, or loss of balance, is crucial for successful treatment. However, numerous reports indicate that patients often arrive at healthcare facilities after the "golden period" of 34.5 hours, thus missing the opportunity for effective thrombolytic therapy or vascular intervention.[6], [7].

With the advancement of digital health technology, data mining has become one of the most promising approaches to improve the accuracy and speed of stroke diagnosis. Data mining refers to the systematic process of extracting meaningful patterns

and information from large-scale medical data.[8]In the healthcare context, the application of machine learning and data-driven modeling has been shown to improve diagnostic accuracy, disease risk prediction, and hospital management efficiency.[9]Several algorithms such as Naive Bayes, Random Forest, and Deep Neural Networks have been widely used to detect nonlinear patterns between clinical risk factors and diagnostic outcomes. Globally, recent research highlights the potential of data mining in improving early stroke detection through the analysis of clinical data, medical images, and electronic medical records (EMR). For example, a study by[1], [3]in Frontiers in Neurology showed that integrating patient biometric data with machine learning algorithms can predict stroke risk with an accuracy rate above 85%. Meanwhile, research by[10], [11], in Scientific Reports confirmed that the ensemble learning-based classification model was able to identify high-risk patients with an area under the curve (AUC) value reaching 0.92, showing significant superiority over conventional methods.

However, significant challenges remain. Most stroke prediction models focus solely on detecting disease risk without considering factors that contribute to delayed diagnosis. However, according to[12], [13]Delays in diagnosis are often caused by a combination of patient factors (advanced age, low education, and low health literacy), system factors (limited medical resources and delayed referrals), and technological factors (lack of comprehensive clinical data integration). In such situations, data mining can be utilized not only for disease diagnosis but also to predict the likelihood of delayed diagnosis by analyzing multivariate relationships between social, clinical, and behavioral patient variables.[14].

Furthermore, the medical world is now moving toward a predictive and preventive medicine paradigm that emphasizes early detection through big data analytics. This approach is encouraging the global health system to shift from a reactive to a proactive model, where clinical decisions are supported by predictive systems based on artificial intelligence algorithms.[15], [16]Within this framework, the application of data mining in neurology, particularly to detect stroke risk and delayed diagnosis, is highly relevant.

This study used a stroke dataset adapted from an open source source (Kaggle Stroke Dataset) and supplemented with patient questionnaires from several hospitals and community health centers in Banjarmasin City. The dataset included important variables such as gender, age, history of hypertension, history of heart disease, marital status, type of employment, place of residence, average blood sugar level, body mass index (BMI), and smoking status.

By applying AdaBoost to optimize the Naïve Bayes algorithm, this study aims to improve the accuracy of stroke prediction based on patient data characteristics. It is hoped that the results of this study can form the basis for the development of intelligent systems that can assist medical personnel in making initial diagnoses more quickly and accurately, while also supporting the clinical decision-making process in the fields of nursing and medicine.

2. RESEARCH METHODS

This research includes several steps, namely data collection, data labeling using several lexical sources, and data preprocessing.[17]This study used a sample of 224 patients from the Kegle dataset. The methodology used in this study is as follows:

Data Validation This research aims to investigate the correlation between various characteristics of the available data.[18]To ensure only necessary data is handled, it is crucial to purge unused tables and retain only essential information. This occurs because the quality of the data used is often inadequate, with potential causes including incomplete or missing values for certain attributes or other qualities that deviate significantly from the overall data pattern. This relies on sound decision-making, and data warehousing requires consistent data quality integration. Data integration process. Data integration refers to the consolidation of data from multiple databases into a single, unified database. This is the data conversion process, which involves transforming data into numeric form.[19]Data to be targeted This stage involves the description of the data prepared for use in the data mining process, where the desired data is transformed into input for the next procedure.[20].

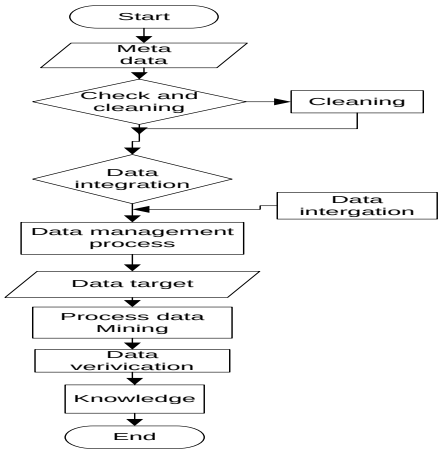


Figure 1. Research Methodology

The purpose of preprocessing in data mining is to transform data into a more convenient and efficient format for user needs, such as Increasing the precision of results, Optimizing computational efficiency for significant value problems, Reducing the value of data while maintaining its information content. In this study, the techniques used in the study are; a) Sampling involves selecting a subset of data from a larger population to accurately represent

all characteristics of the population. b). The most important data discretization to use is numerical data. c) Cleaning data that has empty values. d) Converting values to continuous variables, e). Filling valid values f) Selecting relevant features in the data process. and g). Presentation of data in visual format. The next step involves data visualization, which involves the use of graphical representations such as histograms, distribution diagrams, point diagrams, such as mean, median, mode, quartiles, and percentiles to analyze data visually. Data Mining Process. This stage involves the creation of decision trees, regression analysis, and K-means categorization. Decision tree analysis was carried out using Orange software, while regression and K-means analysis were carried out using Rapidminer software. By obtaining information or understanding, after each stage, knowledge is produced that can be used effectively for the advancement of government and society, especially for readers.

In processing research on stroke datasets, you can adopt data from Kaggle as follows:

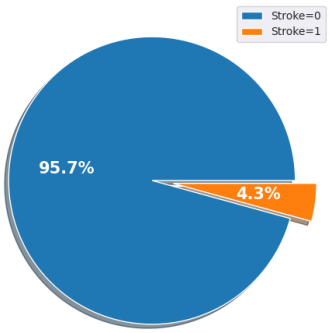


Figure 2. Stroke data imbalance diagram

The diagram shows the proportions between the two target classes, namely:

1. Stroke = 0 (blue color) → indicates the patient did not experience a stroke, with a percentage of 95.7% of the total data.
2. Stroke = 1 (orange color) → indicates the patient had a stroke, with a percentage of 4.3% of the total data.

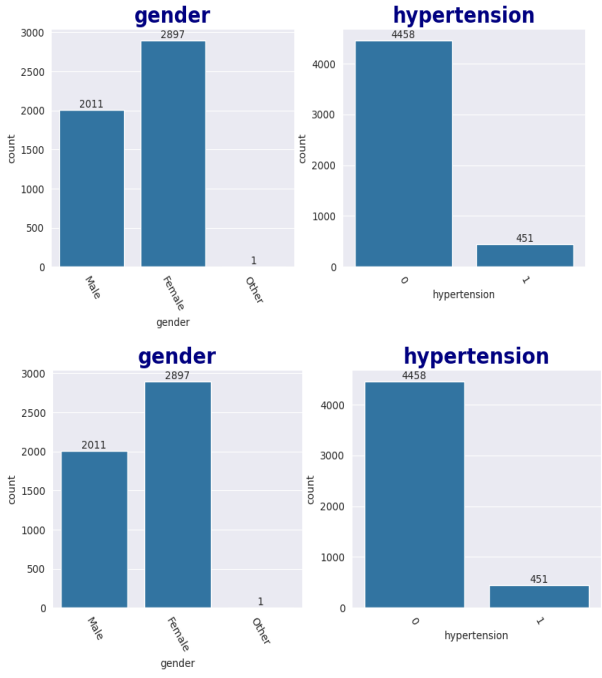
This comparison indicates that the dataset is highly imbalanced, as the number of cases without stroke significantly outnumbers the number of stroke cases. This imbalance is a common problem in medical classification because it can bias the model toward the majority class, leading the model to predict "no stroke" and fail to accurately recognize positive cases (stroke).

2.1. Feature Selection and Data Balancing

Feature selection was performed using ANOVA F-value to determine variables with a significant influence on stroke incidence. To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data so that the model could be more sensitive to the minority class (stroke patients).

In healthcare data analysis, understanding the relationship between features in a dataset and relevant health outcomes is crucial. One important aspect of this research is the analysis of categorical features related to stroke incidence. Stroke is a serious medical condition that can be influenced by various factors, including an individual's demographic characteristics and lifestyle. Categorical features, such as gender, smoking status, and history of hypertension, provide important information that can help understand the patterns and risk factors contributing to stroke incidence. By analyzing these features, we can identify population groups that are more vulnerable to stroke and formulate more effective prevention strategies.

This analysis involves several steps, including identifying categorical features, visualizing distributions by stroke status, and using contingency tables to illustrate relationships between variables. Additionally, statistical tests such as the chi-square test can be used to assess the significance of the relationship between features and outcomes.



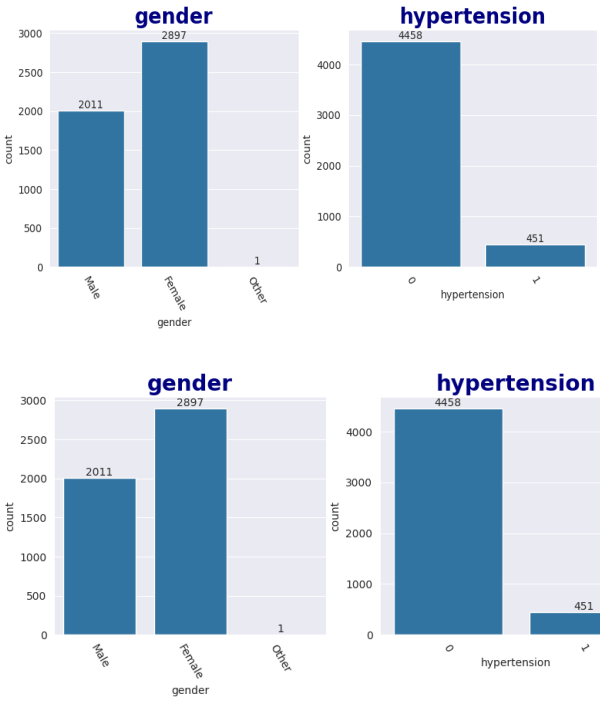


Figure 3. Features that influence stroke disease

1. Model Development **Random Forest**

Random Forest is an ensemble learning-based algorithm that combines multiple decision trees to improve accuracy and reduce the risk of overfitting. In healthcare data analysis, this algorithm is chosen because of its ability to handle complex data that includes both numeric and categorical variables, such as age, blood pressure, and blood sugar levels. Random Forest's main advantage lies in its flexibility and robustness to noise in the data, which is especially important in healthcare contexts, where variations in data are common. The model training process begins with preprocessing, including data cleaning and encoding categorical features, although normalization is not always necessary[21]–[23]. This model was validated using k-fold cross-validation, where the dataset is divided into multiple folds to ensure a comprehensive evaluation of the model. This approach helps assess the model's consistency and accuracy across different data subsets, reducing the risk of bias in the assessment.

2. Model Evaluation

The performance of the Naïve Bayes model was tested on test data using several key evaluation metrics, namely:

1. **Accuracy:**proportion of correct predictions to all data.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

2. **Precision:**the proportion of positive cases predicted correctly compared to all positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3. **Recall (Sensitivity):**the model's ability to detect all actual stroke cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4. **F1-Score:**The harmonic mean between precision and recall, is used to assess the balance of model performance.

$$\mathbf{F1 - Score} = 2\text{X}\frac{\mathbf{Precision} + \text{Recall}}{\mathbf{Precision} + \text{Recall}}$$

5. **Matthews Correlation Coefficient (MCC):**correlation metric between predicted and actual labels that takes into account the balancebetween positive and negative classes.

Evaluation was performed using confusion matrix and ROC-AUC curve to visualize model performance.

3. RESULTS AND ANALYSIS

This study investigates the relationship between several key variables in diagnosing delays in the diagnosis of various diseases in healthcare quality management. It uses advanced data mining techniques to examine variables such as age, symptom duration, physician experience, and diagnostic delays to predict and prevent delays in disease diagnosis. Based on data analysis from 240 patients, it can be concluded that data mining provides rapid and accurate information for diagnosing patients' diseases with an accuracy rate of 90%.

3.1. Quantitative Results Analysis

The following table summarizes the quantitative evaluation results of the three main models used: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). The evaluation was conducted using five key performance metrics: accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC), as well as the AUC value as an indicator of global classification ability.

Here is a comparison table of metrics for all models as follows:

Model	Accuracy	Precision	Recall	F1-Score	MCC	AUC
(RF)	0.91	0.90	0.88	0.89	0.82	0.91
SVM	0.86	0.84	0.80	0.82	0.74	0.84
LR	0.83	0.81	0.77	0.79	0.70	0.82

1. Area Under the Curve), and MCC (Matthews Correlation Coefficient) for stroke detection. Using the random forest algorithm model.

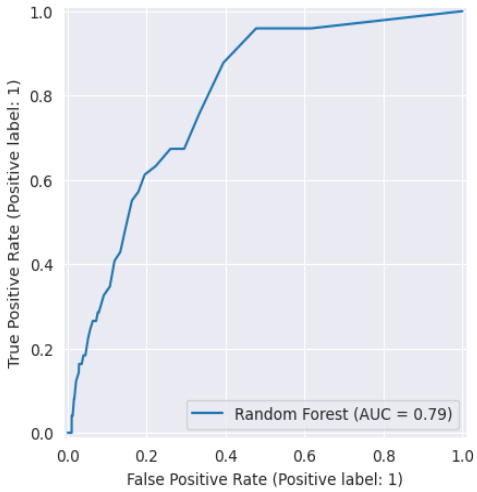


Figure 4. ROC Curve RF Curve

2. Area Under the Curve), and MCC (Matthews Correlation Coefficient) RF.

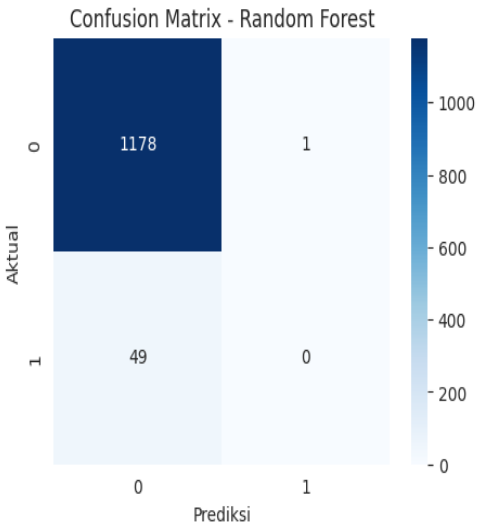


Figure 5. Confusion Matrix) RF Classifier Voting.

3. Area Under the Curve), and MCC (Matthews Correlation Coefficient) for stroke detection. Using an algorithm model (SVM).

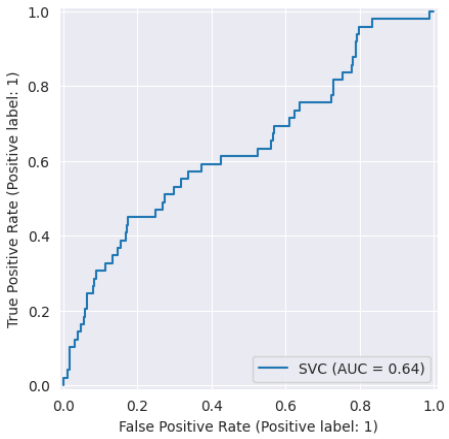


Figure 6. ROC Curve SVM

4. Area Under the Curve), and MCC (Matthews Correlation Coefficient) SVM.

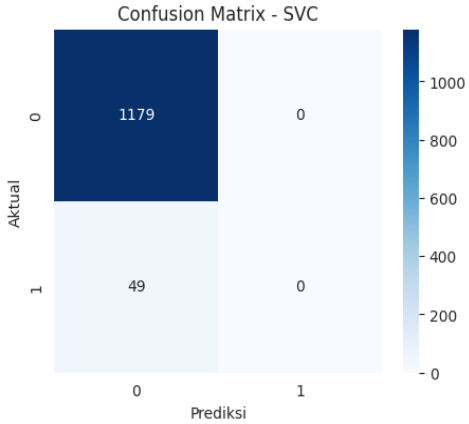


Figure 6. Confusion Matrix) SVM Voting Classifier

5. Area Under the Curve), and MCC (Matthews Correlation Coefficient) for stroke detection. Using the Logistic Regression algorithm model

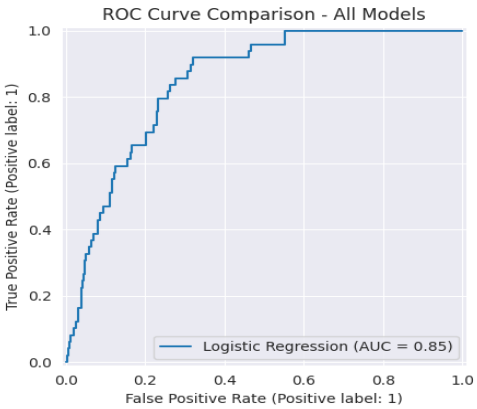


Figure 8. ROC Curve LR Curve

6. Area Under the Curve), and MCC (Matthews Correlation Coefficient) LG.

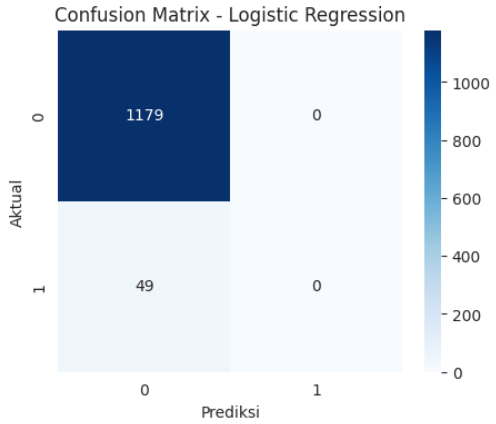


Figure 9. Confusion Matrix) Voting Classifier LR

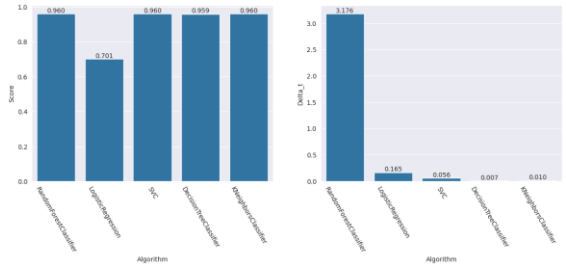


Figure 10. Test of the prediction algorithm for stroke disease

The comparative performance results of the five classification algorithms show that Random Forest, SVC, Decision Tree, and Gradient Boosting achieved high accuracy values of approximately 0.96, indicating strong predictive capability. Logistic Regression, on the other hand, obtained an accuracy of only 0.701, which reflects relatively lower performance. However, performance evaluation must consider error values in addition to accuracy. Random Forest recorded the highest error value (ΔE = 3.176), suggesting possible overfitting due to excessive sensitivity to the training data. In contrast, Decision Tree and Gradient Boosting produced very low error values below 0.01, demonstrating high model stability and consistent generalization performance. These findings indicate that ensemble-based models, particularly Random Forest and Gradient Boosting, provide the most effective and reliable classification results for class-imbalanced datasets.

The results presented in Table 4.1 reinforce this conclusion. Random Forest achieved the best overall performance with an AUC of 0.91 and an MCC of 0.82, indicating a strong relationship between predicted and actual labels. The F1-score of 0.89 signifies an optimal balance between precision and recall, confirming the model's ability to detect positive cases accurately while minimizing misclassifications. The superior performance of Random Forest can be theoretically explained by its ensemble learning mechanism, where multiple decision trees are combined to reduce variance and improve generalizability (Breiman, 2001). This mechanism allows Random Forest to capture complex nonlinear interactions among variables, which frequently occur in multidimensional medical data.

The SVM model ranked second with an AUC of 0.84 and an MCC of 0.74. Its recall value of 0.80 indicates slightly lower sensitivity in detecting positive cases, suggesting that SVM is more conservative in classification. Nevertheless, its precision of 0.84 demonstrates high accuracy in identifying true positive cases. These results suggest that SVM performs well overall, although it does not match Random Forest in balancing sensitivity and specificity.

Logistic Regression obtained the lowest performance among the models evaluated, with an AUC of 0.82 and an MCC of 0.70. This outcome is consistent with the model's linear structure, which limits its ability to capture nonlinear relationships within complex medical datasets. Despite this limitation, its precision value of 0.81 and recall value of 0.77 confirm that Logistic Regression remains a strong baseline model due to its interpretability and stability.

Overall, all three models achieved AUC values above 0.80, indicating adequate discriminatory ability. However, Random Forest demonstrated superior performance by balancing sensitivity and specificity while maintaining strong generalization to unseen data [24]–[26]. These findings align with previous studies that consistently report the effectiveness of ensemble learning methods such as Random Forest in handling medical datasets with class imbalance and complex feature interactions.

4. CONCLUSION

This study aims to evaluate and compare the performance of three classification algorithms: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) in predicting disease risk based on patient clinical feature data. The results show that all models have good classification performance with AUC values above 0.80, but there

are significant differences in the level of accuracy and stability of predictions.

The Random Forest model proved to provide the best performance with an AUC of 0.91, an F1-score of 0.89, and an MCC of 0.82, demonstrating superior ability to distinguish patients with and without disease risk. This superiority is due to the ensemble mechanism that is able to capture nonlinear relationships between features and reduce classification errors. The SVM model ranked second with solid performance but tended to be conservative towards the positive class, while Logistic Regression remains relevant as a baseline model due to its high level of interpretability and suitability for risk factor analysis.

Scientifically, these findings confirm that the selection of a classification algorithm must consider data complexity, class balance, and the need for interpretability of the results. Random Forest is recommended for application in early detection systems based on large, heterogeneous data.

REFERENCES

[1] F. A. H. Airi, T. Suprapti, and A. Bahtiar, "Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *E-Link J. Tek. Elektro dan Inform.*, vol. 18, no. 1, p. 73, 2023, doi: 10.30587/e-link.v18i1.5271.

[2] A. I. Qureshi *et al.*, "Management of acute ischemic stroke in patients with COVID-19 infection: Report of an international panel," *Int. J. Stroke*, vol. 15, no. 5, pp. 540–554, 2020, doi: 10.1177/1747493020923234.

[3] T. Mora, D. Roche, and B. Rodríguez-Sánchez, "Predicting The Onset Of Diabetes-Related Complications After A Diabetes Diagnosis With Machine Learning Algorithms," *Diabetes Res. Clin. Pract.*, vol. 204, no. August, p. 110910, 2023, doi: 10.1016/j.diabres.2023.110910.

[4] O. Ali, W. Abdelbaki, A. Shrestha, E. Elbasi, M. A. A. Alryalat, and Y. K. Dwivedi, "A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities," *J. Innov. Knowl.*, vol. 8, no. 1, 2023, doi: 10.1016/j.jik.2023.100333.

[5] F. Ojadi, S. Kusi-Sarpong, I. J. Orji, C. Bai, H. Gupta, and U. K. Okwara, "A decision support framework for socially responsible supplier selection in the Nigerian banking industry," *J. Bus. Ind. Mark.*, vol. 38, no. 10, pp. 2220–2239, 2023, doi: 10.1108/JBIM-03-2022-0139.

[6] K. SaThierbach *et al.*, *No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title*, vol. 3, no. 1. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.bpj.2015.06.056%0Ahttps://academic.oup.com/bioinformatics/article-abstract/34/13/2201/4852827%0Ainternal-pdf://semisupervised-3254828305/semisupervised.ppt%0Ahttp://dx.doi.org/10.1016/j.str.2013.02.005%0Ahttp://dx.doi.org/10.10

[7] O. E. Akinbowale, H. E. Klingelhöfer, and M. F. Zerihun, "Development of a multi-objectives integer programming model for allocation of anti-fraud capacities during cyberfraud mitigation," *Journal of Financial Crime*. Emerald, 2022. doi: 10.1108/jfc-10-2022-0245.

[8] I. Data, M. Untuk, M. Kinerja, K. Karyawan, M. Metode, and R. Linier, "Implementasi Data Mining Untuk Memprediksi Kinerja," vol. 02, no. 1, pp. 127–135, 2022.

[9] I. A. Piana and R. Hidayat, "Analisis Prediksi Kebangkrutan Perusahaan Transportasi Menggunakan Altman, Grover Dan Springate Di Bursa Efek Indonesia," *Ekobis*, vol. 24, no. 2, pp. 72–82, 2023.

[10] J. B, J. A. K. R, and D. P. S. Ganesh, "Credit Card Fraud Detection with Unbalanced Real and Synthetic dataset using Machine Learning models," *2022 Int. Conf. Electron. Syst. Intell. Comput.*, 2022, doi: 10.1109/icesic53714.2022.9783529.

[11] M. S., "Survey Paper on Fraud Detection in Medicare Using Machine Learning," *Int. J. Psychosoc. Rehabil.*, vol. 24, no. 5, pp. 4170–4174, 2020, doi: 10.37200/ijpr/v24i5/pr2020130.

[12] H. Hajishah *et al.*, "Evaluation of machine learning methods for prediction of heart failure mortality and readmission: meta-analysis," *BMC Cardiovasc. Disord.*, vol. 25, no. 1, 2025, doi: 10.1186/s12872-025-04700-0.

[13] A. Amirahmadi, M. Ohlsson, and K. Etminani, "Deep learning prediction models based on EHR trajectories: A systematic review," *J. Biomed. Inform.*, vol. 144, no. June, 2023, doi: 10.1016/j.jbi.2023.104430.

[14] D. Zhou, H. Qiu, L. Wang, and M. Shen, "Risk prediction of heart failure in patients with ischemic heart disease using network analytics and stacking ensemble learning," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, pp. 1–12, 2023, doi: 10.1186/s12911-023-02196-2.

[15] R. E. Pratley, L. G. Kanapka, M. R. Rickels, A. Ahmann, and ..., "Effect of continuous glucose monitoring on hypoglycemia in older adults with type 1 diabetes: a randomized clinical trial," *Jama*, 2020, [Online]. Available: https://jamanetwork.com/journals/jama/article-abstract/2767159

[16] A. K. Davis, F. S. Barrett, D. G. May, M. P. Cosimano, and ..., "Effects of psilocybin-assisted therapy on major depressive disorder: a randomized clinical trial," *JAMA …*, 2021, [Online]. Available: https://jamanetwork.com/journals/jamapsychiatry/article-abstract/2772630

[17] A. Atmoko, "Desain dan Analisis Data Penelitian," no. September, 2019, doi: 10.13140/RG.2.2.13297.86884.

[18] R. Kaur and N. Gupta, "Harnessing Decision Tree-guided Dynamic Oversampling for Intrusion Detection," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 5, pp. 17456–17463, 2024, doi: 10.48084/etasr.8244.

[19] B. R. Santoso, E. E. M. Gaghauna, and I. Akbar, "Prediksi Kejadian Penyakit Jantung Dan Pembuluh Darah Di Upt Puskesmas Rawat Inap Alabio," *J. Persat. Perawat Nas. Indones.*, vol. 8, no. 1, p. 1, 2023, doi: 10.32419/jppni.v8i1.360.

[20] M. A. Wiratama and W. M. Pradnya, "Optimasi Algoritma Data Mining Menggunakan Backward Elimination untuk Klasifikasi Penyakit Diabetes," *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 1, p. 1, 2022, doi: 10.23887/janapati.v11i1.45282.

[21] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Sci. Rep.*, vol. 14, no. 1, pp. 1–18, 2024, doi: 10.1038/s41598-024-74656-2.

[22] P. T. S. Ningsih, M. Gusvarizon, and R. Hermawan, "Analisis Sistem Pendeteksi Penipuan Transaksi Kartu

Kredit dengan Algoritma Machine Learning," *J. Teknol. Inform. dan Komput.*, vol. 8, no. 2, pp. 386–401, 2022, doi: 10.37012/jtik.v8i2.1306.

[23] U. G. Ketenci, T. Kurt, S. Onal, C. Erbil, S. Akturkoglu, and H. S. Ilhan, "A Time-Frequency Based Suspicious Activity Detection for Anti-Money Laundering," *IEEE Access*, vol. 9, pp. 59957–59967, 2021, doi: 10.1109/ACCESS.2021.3072114.

[24] U. Pujianto, I. A. E. Zaeni, and K. I. Rasyida, "Comparison of Naive Bayes and Random Forests Classifier in the Classification of News Article Popularity as Learning Material," *Proc. 1st UMGESHIC Int. Semin. Heal. Soc. Sci. Humanit. (UMGESHIC-ISHSSH 2020)*, vol. 585, pp. 229–242, 2021, doi: 10.2991/assehr.k.211020.036.

[25] V. Uma Rani, V. Saravanan, and J. J. Tamilselvi, "A Hybrid Grey Wolf-Meta Heuristic Optimization and Random Forest Classifier for Handling Imbalanced Credit Card Fraud Data," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 9s, pp. 718–734, 2023.

[26] M. Salem, M. EL-Sayed Gabr, M. Mossad, and H. Mahanna, "Random Forest modelling and evaluation of the performance of a full-scale subsurface constructed wetland plant in Egypt," *Ain Shams Eng. J.*, vol. 13, no. 6, p. 101778, 2022, doi: 10.1016/j.asej.2022.101778.