Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 2 Iss. 1 July 2024, pp: 34-43
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Performance Comparison of Random Forest, Support Vector Machine and Neural Network in Health Classification of Stroke Patients

Windy Junita Sari[1*], Nasya Amirah Melyani[2] , Fadlan Arrazak[3], Muhammad Asyraf Bin Anahar[4], Ezza Addini[5], Zaid Husham Al-Sawaff[6], Selvakumar Manickam[7]

[1,2,3]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[4]Department of Data Science and Computational Intelligence, Faculty of Computer Science,
International Islamic University of Malaysia, Malaysia
[5]Department of Medicine, Faculty of Medicine, Ankara Yıldırım Beyazıt Üniversitesi, Turkey
[6]Department of Medical Instrumentation Technology,
Center of Technical Research Northern Technical University Mosul, Iraq
[7]National Advanced IPv6 Centre (NAv6), Universiti Sains Malaysia, Malaysia

E-Mail: [1]12150324759@students.uin-suska.ac.id, [2]12150323859@students.uin-suska.ac.id
[3]12150312250@students.uin-suska.ac.id, [4]asyrafanaharwork@gmail.com,
[5]addiniezza@gmail.com, [6]zaidalsawaff@ntu.edu.iq, [7]selva@usm.my

Received Dec 27th 2023; Revised Mar 14th 2024; Accepted Apr 20th 2024
Corresponding Author: Windy Junita Sari

Abstract

Stroke is the second most common cause of death globally, making up about 11% of all deaths from health-related deaths each year, the condition varies from mild to severe, with the potential for permanent or temporary damage, caused by non-traumatic cerebral circulatory disorders. This research began with data understanding through the acquisition of a stroke patient health dataset from Kaggle, consisting of 5110 records. The pre-processing stage involved transforming the data to optimize processing, converting numeric attributes to nominal, and preparing training and test data. The focus then shifted to stroke disease classification using Random Forest, Support Vector Machine, and Neural Network algorithms. Data processing results from the Kaggle dataset showed high performance, with Random Forest achieving 98.58% accuracy, SVM 94.11%, and Neural Network 95.72%. Although SVM has the highest recall (99.41%), while Random Forest and ANN have high but slightly lower recall rates, 98.58% and 95.72% respectively. Model selection depends on the needs of the application, either focusing on precision, recall, or a balance of both. This research contributes to further understanding of stroke diagnosis and introduces new potential for classifying the disease.

Keywords: Classification, Neural Network, Random Forest, Stroke, Support Vector Machine

1. INTRODUCTION

Worldwide, stroke ranks as the second most prevalent cause of death, accounting for around 11% of all deaths from health-related deaths each year. Stroke falls into two main categories: hemorrhagic and ischemic stroke [1][2][3]. The condition can vary from mild to very severe, with damage that may be permanent or temporary. Hemorrhagic, which is rare, involves the rupture of a cerebral blood artery that is bleeding. Stroke is a disease of the brain that causes localized and or global impairment of nerve function, which is sudden, progressive, and rapid. It is caused by non-traumatic circulatory disorders of the brain [4]. Tissue Plasminogen Activator (TPA) is the only drug permitted by the Food and Drug Administration (FDA) to treat ischemic stroke. Only 1% to 8% of patients who might be qualified have received treatment with tPA [5]. As a result, tPA therapy is time-dependent and raises concerns about bleeding complications. As a result, the clinical need for treatment of both ischemic and hemorrhagic stroke has not been met, despite recommendations to manage hemorrhagic stroke as a supportive treatment [6].

The World Stroke Organization says that as many as approximately 5.5 million individuals are predicted to die as a result of the 13 million stroke victims that occur each year [7]. With disability as the leading cause, stroke ranks second in the globe in terms of causes of mortality [8][9]. The World Health Organization (WHO) also reports that low- and middle-income people account for 87% of stroke-related mortality in nations, accounting for 70% of all stroke-related deaths globally. According to projections, stroke will account for

14.4% of all fatalities by 2030, 8.2% of all deaths by that year, and 14.4% of all deaths by that year [10]. Strokes continue to rise in low- and lower-middle-income countries, where they cause 89.0% of Disabilitas Adjust Life Years (DALYs) and 86.0% of deaths. The atherogenic effects of some Antiretroviral (ARV) drugs, HIV-related inflammation, immunological activation, microbial translocation, hypertension, diabetes, dyslipidemia, renal impairment, and smoking are a few prevalent risk factors for Cardiovascular Disease (CVD). Intracranial vasculopathy can also result from inflammation and hypercoagulability [11].

This has prompted a large number of studies on stroke diseases, to obtain faster and more accurate results by using certain algorithms to perform classification and manage very large data sets [12]. In general, classification algorithms are divided both non-parametric and parametric methods. While non-parametric procedures can be employed with optional data distributions without understanding the shape of the underlying data structure, parametric approaches require predicting the statistical distribution of the data [13]. Random Forest, Support Vector Machine, and Neural Network are some of the classification algorithms that will be applied in this research. I used these methods in my research because the previous research was not accurate enough. Therefore, I tried these methods in my new research to be tested again.

Reranking query expansion results using the Random Forest algorithm can enhance the efficacy of recommendations derived from medical literature, according to research done in 2019[14]. To overcome class imbalance and dimensionality reduction, research was conducted again in 2023 using the vector support machine algorithm to process PCa datasets using SVM-PCa-EDD. The logistic regression model (LR)'s area under the Receiver Operating Characteristic (ROC) curve (AUC) for the sample data set was 58.4% after accounting for class imbalance, while the AUC of the LR for the unbalanced dataset ROC was 58 [15]. In 2019, research has been conducted using a neural network classification algorithm. The suggested framework was tested on Pima Indian diabetic patient data, consisting of 768 records with eight characteristics. With 86.26% classification accuracy, we succeeded. For the classification of Type 2 Diabetes data, we propose a Deep Learning framework based on stacked autoencoders. When the technique was tested using machine learning data from UCI, it outperformed other current categorization techniques [16].

Based on this, researchers compared Random Forest, Support Vector Machine, and Neural Network algorithms using the Kaggle stroke patient health dataset.The data was then processed with RapidMiner software, and the analysis was performed automatically.The results of stroke data processing are expected to increase public understanding and knowledge and find new opportunities or potential for classifying stroke diseases.

2. MATERIAL AND METHOD

2.1 Dataset

The stroke patient health dataset we use is a dataset we obtained from Kaggle, which has 5110 records. By sharing the processed data, this data will be used for application and level analysis. Divide the test data and training data, then model the Random Forest algorithm, Support Vector Machine, and Neural Network. Then using the algorithm will be used to run the data training process and the test data is used to learn the model. Figure 1 shows the stages of research conducted in this research.

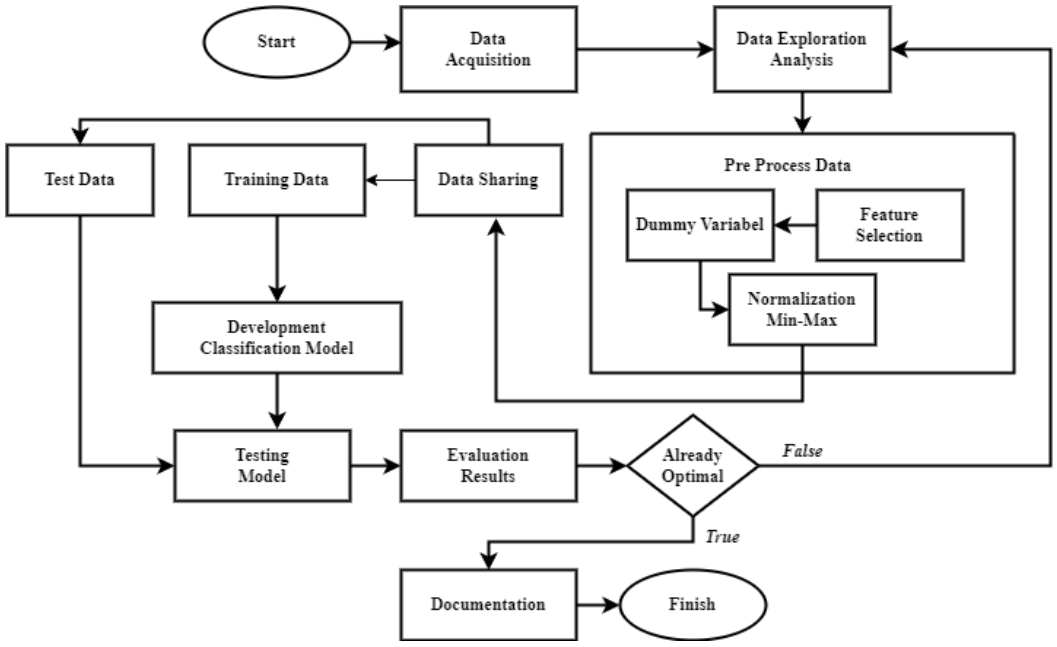

Figure 1. Research Methodology

2.2 Data Acquisition

Data acquisition is the stage of collecting data for processing. Data can be obtained in a variety of ways, including direct interviews, questionnaires, observation, and data collection from existing sources. The method used for data collection should be in accordance with the objectives of the research and the given hypothesis. Data can also be obtained by taking data from sources such as data on the internet, government data, and hospital data[17].

2.3 Pre-process

Data preprocessing is an important part of the machine learning cycle and is an essential step in data analytics [18]. In the healthcare field, access to clean and complete data sets is essential. A model's ability to learn and generalize is affected by data quality [19].

2.4 Normalization

One of the preprocessing techniques for scaling or mapping is normalization. At this stage, the scale of the data is changed to a smaller scale. The new data scale can help classification by removing noisy and less relevant features. Formula 1 can be used to obtain the normalization value for each data set [1].

$$A = \frac{A - nilai\ minA}{nilai\ max.A - nilai\ min.A} \quad (1)$$

2.5 Random Forest Algorithm

Many decision trees' outputs are combined by the popular Random Forest (RF) machine learning technique to produce a single result: a prediction model for various interests[20]. Random Forest uses two methods to compensate for the classification deficiencies of individual regression trees. First, for each tree in the Random Forest, a "bootstrap" dataset is created by sampling from the input training dataset. Second, by forcing all decision trees by reducing the statistical association across every decision tree, "random feature selection" allows decision trees to use a randomly chosen portion of their nodes' predictor subset[21]. In addition, the random forest algorithm can greatly enhance the robustness and accuracy of the model [14].

Random forest, also referred to as random ensemble, is a combination of decision trees. There are three main components, namely [2]:

1. Creating prediction trees by bootstrap sampling.
2. Utilizing randomized predictors for each decision tree
3. The random forest will create a majority vote for classification or average for regression, or use the result of the tree immediately.

Then, using a "boot-strap sample" from the data set, a large number of decision trees are randomly generated. The branching of each tree is determined by randomly selected predictors at the nodes. The impact of each tree on the RF estimate differs depending on its weight, but the average of all the results is the final RF estimate. This method uses a "black box" feature, so each tree is not viewed in isolation. The Random Forest algorithm is more robust than other machine learning algorithms because it can create random trees and randomly select training data from subsets. In addition, the random forest algorithm keeps the overfitting rate constant, even though boot-strap sampling is used to randomly select subsets of data for training [22][23].

2.6 Support Vector Machine (SVM)

The main principle of SVM is to find a dividing line function, or hyperplane, that can maximally separate the data of the two classes. Maximum means that the dividing line can separate the data of the two classes with the best margin, which is the distance of the hyperplane line to the nearest class member. The margin that has the ability to maximally separate the classes is called the Optimal Hyperplane [3].

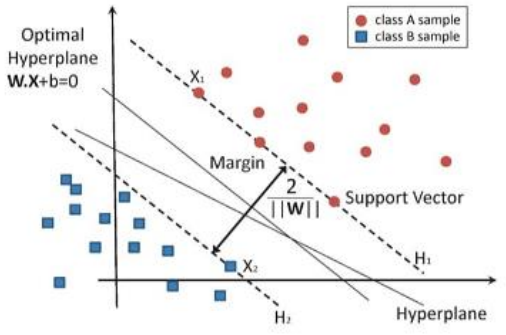

Figure 2. Data Classification using Support Vector Machine (SVM)

The SVM calculation formula is as follows: SVM can determine the right hyperplane to classify into two groups[2].

1. Data point:
   $X_i = \{X_{1,}\, X_{2,} \dots\dots, X_{n,}\} \in R^n$

2. Data class : $y_i \in \{-1, y1\}$

3. Data and class pairs: $\{(x1, y1)\}\, N_{i=1}$

4. Maximize function as equation 2.

$$\text{Ld} = \sum_{i=1}^{N} \alpha_{i-} \quad \sum_{i=1}^{N} \sum_{i=1}^{N} \quad \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2}$$

   Terms : $0 \leq \alpha_i \leq C \; dan \; \sum_{i=1}^{N} \alpha_i \; y_i$

5. Calculation of w and b, as equation 3.

$$\text{W} = \sum_{i}^{N} \alpha_i y_i x_i \quad \text{b} = -\frac{1}{2}\, \text{w}.x^{+} + \text{w}.x^{-} \tag{3}$$

6. Sign of the classification decision function, as equation 4.

$$(\text{f(x)}) = \sum_{i=1}^{m} \alpha_i y_i \; K\; (x_{,} x_i) + \text{b} \tag{4}$$

   Explanation :

N	: Amount of data
N	: Data dimension or number of features
Ld	: Lagrange Multiplier Duality
αi	: Weight value of each data point
C	: Constant value
m	: Number of support vectors/data points that have αi > 0 K(x,xi) : kernel function

2.7 Neural Network Algorithm

Three layers: an input layer, an output layer, and one or more hidden layers make up a neural network. One or more cells, or units, make up each layer. The weighted total of all of a cell's inputs determines the output (yi, j) for each cell. An activation function for a basic neural network is represented by the following equation 5 and 6.

$$y_{ij} = \text{f}\,(\sum[W_{ij}\,(1, j-1) \times \; y_{ij-1}] + b_{ij}) \tag{5}$$

Where, here is how the sigmoid function, f(), is defined:

$$\text{f(x)} = \frac{1}{1+\exp(-x)} \tag{6}$$

A simple Neural Network (NN) is made up of three layers: : a hidden layer with ten sigmoid cells, an output layer with one cell activating the identity function, and an input layer with P input cells (P equals the number of prior breaks of the observed GW level). Cell i at layer j and cell l at layer j-1 are represented by the weighted variables yi,j, bi,j, and Wi,j (l, j-1).

All of the cells in the following layer are linked to every cell within the preceding layer when a simple NN is fully connected. The quantity of cells within the obscure layer of the hidden layer × [number of input cells + number of output cells + 1] + quantity of the output layer's cells is the number of parameters of a basic neural network (P × 10)+21. For instance, a basic Neural Network (NN) has 51 parameters, such as biases and connection weights, when the network input is three lags behind the GW rate that was previously measured. It is necessary to tune these parameters in order to reduce the network's prediction error [24].

3. RESULTS AND ANALYSIS

3.1 Data Understanding

Data understanding, or data comprehension, is the next step after research comprehension. The purpose of this stage is in order to better comprehend the characteristics of the data to be processed. A sample of 5103 data from the Kaggle website [25] will be used for the data mining process. The initial dataset has twelve

attributes, one of which is shown as a label or target, and the other eleven as attributes, the initial dataset can be seen in Table 1.

Table 1. Initial Data

Id	Gender	Age	Hypertension	…	BMI	Smooking Status	Stroke
9046	Male	67	0	…	36,6	Formely Smoked	1
51676	Female	61	0	…	N/A	Never Smoked	1
31112	Male	80	0	…	32,5	Never Smoked	1
60182	Female	49	0	…	34,4	Smokes	1
1665	Female	79	1	…	24	Never Smokes	1
56669	Male	81	0	…	29	Formely Smoked	1
…	…	…	…	…	…	…	…
19723	Female	35	0	…	30,6	Never Smoked	0
37544	Male	51	0	…	25,6	Fomerly Smoked	0
44679	Female	44	0	…	26,2	Unknown	0

In the next stage, the initial dataset was transformed. There are some attributes that have numerical values that are converted to nominal through transformation. The converted numerical values include age, hypertension, heart disease, baseline glucose level, BMI, and stroke. While the nominal values were not converted, such as gender, continuously married, type of employment, type of residence, and smoking. Table 2 displays the data transformation results for each attribute.

Table 2. Data Transformation Results

Id	Gender	Age	Hypertension	…	BMI	Smooking Status	Stroke
9046	1	67	0	…	36,6	3	1
51676	2	61	0	…	29,1	2	1
31112	1	80	0	…	32,5	2	1
60182	2	49	0	…	34,4	1	1
1665	2	79	1	…	24	2	1
56669	1	81	0	…	29	3	1
…	….	….	….	…	…	…	…
44873	2	35	0	…	40	2	0
19723	1	51	0	…	30,6	2	0
37544	2	44	0	…	25,6	3	0

Table 2 displays the results of transforming the data set for each attribute. The transformation process converts numerical values into corresponding category or nominal values. Each column in the table provides an explanation of the relevant attribute. For instance, the "Id" column gives a unique identification for each entry in the dataset, which remains unchanged. The "Gender" column denotes the gender of the individual, where 1 and 2 may represent specific genders, such as male and female. The "Age" column indicates the age of the individual, while the "hypertension" column is already in numeric form, where a value of 0 indicates the absence of hypertension and a value of 1 indicates the presence of hypertension. The "BMI" column shows an individual's Body Mass Index (BMI) in numeric form, which can be categorized into categories such as lean, normal, obese, etc. "Smoking Status" column indicates one's smoking status, with values of 1 to 3 representing categories such as non-smoker, light smoker, heavy smoker, and so on. Finally, the "Stroke" column indicates a person's stroke status, with a value of 0 indicating no stroke and a value of 1 indicating a stroke without any changes.

3.2 Random Forest

The study used n_estimators of 200 to train the random slope model because the random slope consists of several slope options. This number of n_estimators indicate the number of slope options used in the random forest model, the evaluation of view table 3.

Table 3. Random Forest Evaluation Results

Model	Performance		
	Accuracy	Precission	Recall
Random Forest	98.58%	98.65%	98.58%

The Random Forest model applied to the dataset showed satisfactory performance. The accuracy of the model reached 98.58%, indicating the success rate in correctly classifying the data. The precision of 98.65%

indicates the proportion of correct positives in the model's positive predictions, while the recall value of about 98.58% indicates the model's ability to identify most of the positive classes in the dataset.

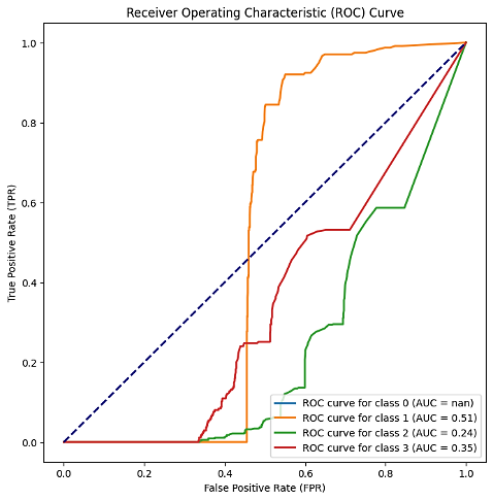

Figure 3. ROC Random Forest

The Receiver Operating Characteristic (ROC) curves for several classes in the Random Forest model are shown in Figure 3. The ROC curve of class 0 has an incomputable Area Under the Curve (AUC) value (nan). The class 1 ROC curve has two curves, with AUCs of 0.51 and 0.24, respectively, and the class 3 ROC curve has an AUC of 0.35. This information provides an idea of the model's performance in predicting true positives and false positives across different classes.

3.3 Support Vector Machine (SVM)

In this study, the SVM model was trained with the following hyperparameters, can be seen as table 4.

Table 4. Hyperparameters of the SVM model

Kernel	Rbf
C	1000
Gamma	0.1

The trained SVM model performs very well when hyperparameters are added to it, the experimental results can be seen in table 5.

Table 5. Support Vector Machine Evaluation Results

Model	Performance		
	Accuracy	Precision	Recall
SVM	94.11%	88.57%	99.41%

The SVM model applied to the dataset showed satisfactory performance. The accuracy of the model reached 94.11%, indicating the success rate in correctly classifying the data. The precision of 88.57% indicates the proportion of true positives in the model's positive predictions, while the recall value of about 94.11% indicates the model's ability to identify most of the positive classes in the dataset, The trained SVM model performs very well when hyperparameters are added to it, the experimental results can be seen in figure 4.

An area AUC value of 0.63 on the ROC curve of a SVM model signifies its capability in distinguishing between positive and negative classes, suggesting performance better than random guessing yet with room for improvement. However, the interpretation of AUC values depends on the context and requirements of the problem [26]. It's crucial to assess trade-offs between sensitivity and specificity, compare with baseline models, analyze feature importance, and conduct model tuning for potential enhancements. Additionally, validation techniques like cross-validation are essential to evaluate the model's generalization ability. Further refinement may be necessary to achieve better performance based on specific task needs and performance thresholds [27].

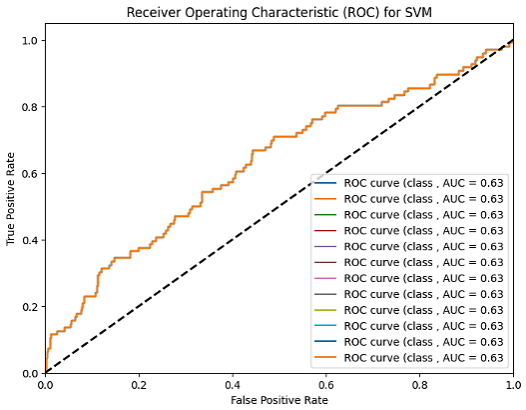

Figure 4. ROC SVM

3.4 Neural Network (NN)

This experiment involves the process of training the model using the previously described method. Here are the details of the parameters used in the 'fit' function to train the model, as table 6.

Table 6. ANN model hyperparameters

Parameter	Value
Optimizer	Adam
Activation Function	ReLU
Number of Epochs	100
Batch Size	128
Verbosity	1

The trained SVM model performs very well when hyperparameters are added to it, the experimental results can be seen in table 7.

Table 7. Artificial Neural Network Evaluation Results

Model	Performance		
	Accuracy	Precission	Recall
ANN	95.72%	95.15%	95.72%

The dataset analyzed by the Neural Network model had an impressive accuracy rate of 95.72%. Accurately classifying data is crucial in any machine learning application and this is evident from the high accuracy achieved by the model. Moreover, the model's accuracy rate of 95.15% in making positive predictions highlights its proficiency. Precision is even more critical when false positives are expensive, as it indicates the model's ability to identify positive cases among all instances classified as positive. It's essential to note that the model has a recall rate of approximately 95.72%, which means that it can correctly identify a significant percentage of positive instances in the dataset. In cases where identifying every positive instance is crucial, recall (also known as sensitivity) becomes essential, even if it results in misclassifying some negative instances as positive, A visualization of the performance data from the experiment can be seen in Figure 5.

Striking an ideal balance between recall and precision is crucial to accurately evaluate the performance of a model. High recall means that the model is good at identifying positive cases, while high precision indicates that the model is careful in predicting positive cases. In machine learning, it can be challenging to achieve a balance between these two metrics because improving one may come at the expense of the other. Therefore, the Neural Network model's ability to achieve an excellent balance between precision and recall is a testament to its effectiveness in classification tasks.

In Figure 5, we can be seen the ROC curve of the Artificial Neural Network with two parts, namely the ROC curve for class 0 (AUC = 1.00) and the ROC curve for class 1 (AUC = 1.00).

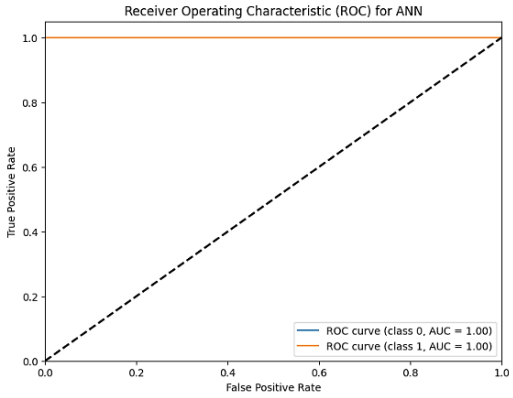

Figure 5. ROC Artificial Neural Network

The performance comparison of the three classification models used, namely Random Forest, SVM, and ANN can be seen in Table 8.

Table 8. Classification Algorithm Comparison Results

Model	Performance		
	Accuracy	Precision	Recall
Random Forest	98.58%	98.65%	98.58%
Support Vector Machine	94.11%	88.57%	99.41%
Artificial Neural Network	95.72%	95.15%	95.72%

Random Forest showed the highest accuracy (98.58%), followed by ANN (95.72%), and SVM (94.11%). All three models also had high precision rates, with Random Forest reaching 98.65%, ANN 95.15%, and SVM 88.57%. It should be noted that SVM has a very high recall rate (99.41%), While Random Forest and ANN exhibit high recall rates, with Random Forest achieving 98.58% and ANN achieving 95.72%, it's worth noting that there's a slight difference in their performance compared to Support Vector Machine (SVM). Recall, also known as sensitivity, measures the proportion of actual positive cases that are correctly identified by the model. The high recall rates of Random Forest and ANN indicate their effectiveness in capturing the majority of positive instances within the dataset. This suggests that both Random Forest and ANN are proficient at minimizing false negatives, which is crucial in scenarios where identifying all positive cases is a priority, such as medical diagnostics. However, despite their high recall rates, it's essential to conduct a comprehensive analysis to determine the overall performance of each algorithm, considering other evaluation metrics such as precision, accuracy, and F1-score. Additionally, understanding the trade-offs associated with each algorithm, such as computational complexity and interpretability, is vital for informed decision-making in model selection. Therefore, while Random Forest and ANN demonstrate commendable recall rates, a holistic assessment considering various factors is necessary to determine the most suitable algorithm for a particular task or application. These results point to whether prioritizing precision, recall, or a balance of both for specific application needs or goals determines the choice of model [28].

4. CONCLUSION

The purpose of this study is to assess the effectiveness of three categorization algorithms: Neural Network, Support Vector Machine, and Random Forests in the context of stroke patient health classification using a dataset from Kaggle. However, the assessment did not find the effectiveness and advantages of the three algorithms. Results show that Random Forest performs best with an accuracy of 98.58%, outperforming Neural Network (95.72%) and Support Vector Machine (94.11%). Random Forest can be considered as the optimal choice for classifying stroke patients' health data. This research provides practical insights and contributes to understanding the benefits of classification models in the healthcare field. Suggestions for future research involve exploring factors that affect model performance, improving data pre-processing, and further investigating the interpretation of model results. Overall, this study has the potential to serve as a basis for further research and provide practical guidance in the selection of classification models.

REFERENCES

[1] G. Sailasya and G. L. A. Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.

[2] H. Li, S. Ghorbani, C. C. Ling, V. W. Yong, and M. Xue, “The extracellular matrix as modifier of

neuroinflammation and recovery in ischemic stroke and intracerebral hemorrhage," *Neurobiol. Dis.*, vol. 186, no. September, p. 106282, 2023, doi: 10.1016/j.nbd.2023.106282.

[3] G. Fekadu *et al.*, "Management protocols and encountered complications among stroke patients admitted to stroke unit of Jimma university medical center, Southwest Ethiopia: Prospective observational study," *Ann. Med. Surg.*, vol. 48, no. September, pp. 135–143, 2019, doi: 10.1016/j.amsu.2019.11.003.

[4] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke," *J. Tek.*, vol. 12 No.02, no. 02, pp. 72–80, 2023, doi: http://dx.doi.org/10.31000/jt.v12i2.9099.

[5] T. Imai, S. Iwata, D. Miyo, S. Nakamura, M. Shimazawa, and H. Hara, "A novel free radical scavenger, NSP-116, ameliorated the brain injury in both ischemic and hemorrhagic stroke models," *J. Pharmacol. Sci.*, vol. 141, no. 3, pp. 119–126, 2019, doi: 10.1016/j.jphs.2019.09.012.

[6] H. Y. Cheng, Y. S. Wang, P. Y. Hsu, C. Y. Chen, Y. C. Liao, and S. H. H. Juo, "miR-195 Has a Potential to Treat Ischemic and Hemorrhagic Stroke through Neurovascular Protection and Neurogenesis," *Mol. Ther. - Methods Clin. Dev.*, vol. 13, no. June, pp. 121–132, 2019, doi: 10.1016/j.omtm.2018.11.011.

[7] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors*, vol. 22, no. 13, 2022, doi: 10.3390/s22134670.

[8] Q. Huang *et al.*, "Association between genetic predisposition and disease burden of stroke in China: a genetic epidemiological study," *Lancet Reg. Heal. - West. Pacific*, vol. 36, no. 27, p. 100779, 2023, doi: 10.1016/j.lanwpc.2023.100779.

[9] E. Natarajan, F. Augustin, M. K. A. Kaabar, C. R. Kenneth, and K. Yenoke, "Various defuzzification and ranking techniques for the heptagonal fuzzy number to prioritize the vulnerable countries of stroke disease," *Results Control Optim.*, vol. 12, no. June, p. 100248, 2023, doi: 10.1016/j.rico.2023.100248.

[10] A. Byna and M. Basit, "Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 9, no. 3, pp. 407–411, 2020, doi: 10.32736/sisfokom.v9i3.1023.

[11] C. I. Hatleberg *et al.*, "Predictors of Ischemic and Hemorrhagic Strokes Among People Living With HIV: The D:A:D International Prospective Multicohort Study," *EClinicalMedicine*, vol. 13, pp. 91–100, 2019, doi: 10.1016/j.eclinm.2019.07.008.

[12] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking," *Informatics Med. Unlocked*, vol. 15, no. January, p. 100180, 2019, doi: 10.1016/j.imu.2019.100180.

[13] N. B. Toosi, A. R. Soffianian, S. Fakheran, S. Pourmanafi, C. Ginzler, and L. T. Waser, "Comparing different classification algorithms for monitoring mangrove cover changes in southern Iran," *Glob. Ecol. Conserv.*, vol. 19, 2019, doi: 10.1016/j.gecco.2019.e00662.

[14] B. Cui, H. Ding, S. Li, and G. Zhuang, "Recommendation of Clinical Diagnostic Literature based on Random Forest Model and Query Expansion," *Procedia Comput. Sci.*, vol. 162, no. Itqm 2019, pp. 59–67, 2019, doi: 10.1016/j.procs.2019.11.258.

[15] B. A. Akinnuwesi *et al.*, "Application of support vector machine algorithm for early differential diagnosis of prostate cancer," *Data Sci. Manag.*, vol. 6, no. 1, pp. 1–12, 2023, doi: 10.1016/j.dsm.2022.10.001.

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

[17] A. Putri *et al.*, "Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 20–26, 2023, doi: 10.57152/malcom.v3i1.610.

[18] E. Oluwasakin *et al.*, "Machine Learning with Applications Minimization of high computational cost in data preprocessing and modeling using MPI4Py," *Mach. Learn. with Appl.*, vol. 13, no. May, p. 100483, 2023, doi: 10.1016/j.mlwa.2023.100483.

[19] S. Albahra *et al.*, "Seminars in Diagnostic Pathology Artificial intelligence and machine learning overview in pathology & laboratory medicine : A general review of data preprocessing and basic supervised concepts," vol. 40, no. February, pp. 71–87, 2023, doi: 10.1053/j.semdp.2023.02.002.

[20] L. Urso, E. Petermann, F. Gnädinger, and P. Hartmann, "Use of random forest algorithm for predictive modelling of transfer factor soil-plant for radiocaesium: A feasibility study," *J. Environ. Radioact.*, vol. 270, no. October, 2023, doi: 10.1016/j.jenvrad.2023.107309.

[21] P. Josso, A. Hall, C. Williams, T. Le, P. Lusty, and B. Murton, "Application of random-forest machine learning algorithm for mineral predictive mapping of Fe-Mn crusts in the World Ocean," *Ore Geol. Rev.*, vol. 162, no. June, p. 105671, 2023, doi: 10.1016/j.oregeorev.2023.105671.

[22] C. M. YEŞİLKANAT, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm," *Chaos, Solitons and Fractals*, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110210.

[23] A. Hasnain, Y. Sheng, M. Z. Hashmi, U. A. Bhatti, Z. Ahmed, and Y. Zha, "Assessing the ambient air quality patterns associated to the COVID-19 outbreak in the Yangtze River Delta: A random forest approach," *Chemosphere*, vol. 314, no. October 2022, p. 137638, 2023, doi: 10.1016/j.chemosphere.2022.137638.

[24] R. Solgi, H. A. Loáiciga, and M. Kram, "Long short-term memory neural network (LSTM-NN) for aquifer level time series forecasting using in-situ piezometric observations," *J. Hydrol.*, vol. 601, 2021, doi: 10.1016/j.jhydrol.2021.126800.

[25] SAMSON TONTOYE, "healthcare dataset stroke data," *Kaggle*, 2021. https://www.kaggle.com/datasets/godfatherfigure/healthcare-dataset-stroke-data (accessed Mar. 10, 2024).

[26] M. F. Fayyad and D. T. Savra, "Sentiment Analysis of Towards Electric Cars using Naive Bayes Classifier and Support Vector Machine Algorithm," vol. 1, no. July, pp. 1–9, 2023, doi: https://doi.org/10.57152/predatecs.v1i1.814.

[27] C. P. Trisya, N. W. Azani, and L. M. Sari, "Performance Comparison of ARIMA , LSTM and SVM Models for Electric Energy Consumption Analysis," vol. 1, no. January, pp. 85–94, 2024, doi: https://doi.org/10.57152/predatecs.v1i2.869.

[28] G. Fu, "Tuning model parameters in class-imbalanced learning with precision-recall curve," no. August, 2021, doi: 10.1002/bimj.201800148.