ILKOM Jurnal Ilmiah Vol. 14, No. 3, December 2022, pp. 218-228
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Classification of stroke patients using data mining with AdaBoost, Decision Tree and Random Forest models

Bahtiar Imran [a,1]; **Erfan Wahyudi** [b,2,*]; **Ahmad Subki** [a,3] ; **Salman** [a,4] ; **Ahmad Yani** [a,5]
[a] *Universitas Teknologi Mataram, Jl.Kampus Universitas Teknologi Mataram, Kekalik Mataram, Mataram and 83115, Indonesia*
[b] *Institut Pemerintahan Dalam Negeri, Jl. Gajah Mada No.1, Lenen, Nusa Tenggara Barat, Praya and 83522, Indonesia*
** Corresponding author*

Article history: Received July 17, 2022; Revised October 12, 2022; Accepted November 29, 2022; Available online December 20, 2022

Abstract

A stroke is a fatal disease that usually occurs to the people over the age of 65. The treatment progress of the medical field is growing rapidly, especially with the technological advance, with the emergence of various medical record data sets that can be used in medical records to identify trends in these data sets using data mining. The purpose of this study was to propose a model to classify stroke survivors using data mining, by utilizing data from the kaggle sharing dataset. The models proposed in this study were AdaBoost, Decision Tree and Random Forest, evaluation results using Confusion Matrix and ROC Analysis. The results obtained were that the decision tree model was able to provide the best accuracy results compared to the other models, which was 0.953 for Number of Folds 5 and 10. From the results of this study, the decision tree model was able to provide good classification results for stroke sufferers.

Keywords: Data Mining; AdaBoost; Decision Tree; Random Forest; Stroke.

Introduction

Atrial fibrillation can result in a stroke which is potentially fatal. Stroke is a fatal disease where people over the age of 65 are prone to it. Stroke has been acknowledged as the third leading cause of death in the United States [1] and other industrialized countries because it harms the brain similar to a "coronary episode" that puts the heart at risk [2]. Stroke has become one of the main factors contributing to disability [3] worldwide [4]–[6].

Stroke or cerebrovascular accident (CVA) can be defined as a loss of brain function caused by a sudden interruption of the blood supply to a part of the brain. This condition occurs because of circulatory issues in the brain that results in paralysis or death. Stroke detection is difficult because people do not regularly monitor their brain and heart health. Computed Tomography (CT) scan, Magnetic Resonance Imaging (MRI), and Electrocardiogram have been used in general diagnostic procedures (ECG or EKG) [7]. Stroke survivors are more likely to experience sudden onset, and stroke is the leading cause of adult epilepsy. This can be an ischemic or hemorrhagic stroke (rupture of a blood vessel). Ischemic stroke accounts for 80-85 percent of all strokes[8]. However, over the years, self-knowledge, experience, and qualifications have played an important role in the pre-diagnosis phase of acute stroke type [8]. The progress of the medical field is increasing rapidly, especially with the advent of technology, with the emergence of various medical record data sets that can be used in medical records to identify trends in these data sets using data mining[9].

Research on stroke has been widely carried out by utilizing technology assistance and by using different methods to obtain many references related to stroke. The study[2] of stroke prediction was carried out using a machine learning algorithm, from the five models used to obtain good accuracy results. In [4] using data mining for the stroke prediction and control, the results obtained that KNN and Decision Tree 4.5 can be used to predict stroke. In [7] the stroke diagonal used hyperparameters from deep learning, and indicated that optimization using Bayesian was superior in time optimization [10] this study detected stroke using deep learning by utilizing data from MRI images. This study obtained SSD results with a precision of 89.77%, [11] the results obtained in this study were, the random forest method obtained a high accuracy result of 96. In [12] stroke prediction using artificial intelligence by utilizing Real Time Electromyography Signals. This study reported that random forest accuracy was 90.38% and LSTM was 98.95%. In the study of [13] data mining techniques were used to predict stroke with 95% accuracy, patients with heart disease conditions, immune diseases, diabetes, kidney failure, hyperlipidemia, epilepsy had a tendency to suffer a stroke. Furthermore data mining was employed on ischemic stroke prediction [14], where the result

http://dx.doi.org/10.33096/ilkom.v14i3.1328.218-228

indicated that SVM obtained the highest accuracy with 0.9789. This study [15] used factorization and machine learning models in detecting stroke, the SVM classification method and XGBoost obtained 98% accuracy result. In[16] machine learning was used to predict stroke risk from the results of lab tests, this study focused on the development and evaluation of the proposed prediction model, the random forest model got the best results than the other models. [17] This study used machine learning to predict stores, the models used were XGBoost, random forest, KNN, logistic regression and SVM. XGBoost got the highest accuracy with 93.68% compared to other models. [18] using artificial neural networks and machine learning for stroke type prediction, artificial neural networks got 91.7995% accuracy and nave Bayes got 99.1983% accuracy. [19] using natural language processing based on machine learning by utilizing reports from MRI radiology, the multi-CNN algorithm showed the best classification performance (0.805), and the CNN algorithm (0.799). In a study using a machine learning classification algorithm for the analysis and performance of stroke prediction reported that of the five proposed methods the nave Bayes model provided an accuracy of 82% [20].

Many classification methods have been proposed to classify stroke or other diseases, including data mining. Most of the previous studies used different methods with different results. No previous studies were related to the classification of stroke patients using data mining with the AdaBoost model, decision tree and random forest model. Therefore, this study aimed to classify stroke patients using data mining with the AdaBoost model, decision tree and random forest model. The reasons for choosing these three models are, AdaBoost is a model that provides good performance in classifying [21], decision tree gives high accuracy results for the classification process [22], while the random forest model gives good accuracy results in classifying [23]. The overall data used in this study was 5110 with 11 features, the data used as training was 80% of the total data and the other 20% was used as testing data.

Method

The research method employed in this study can be seen in **Figure 1**.

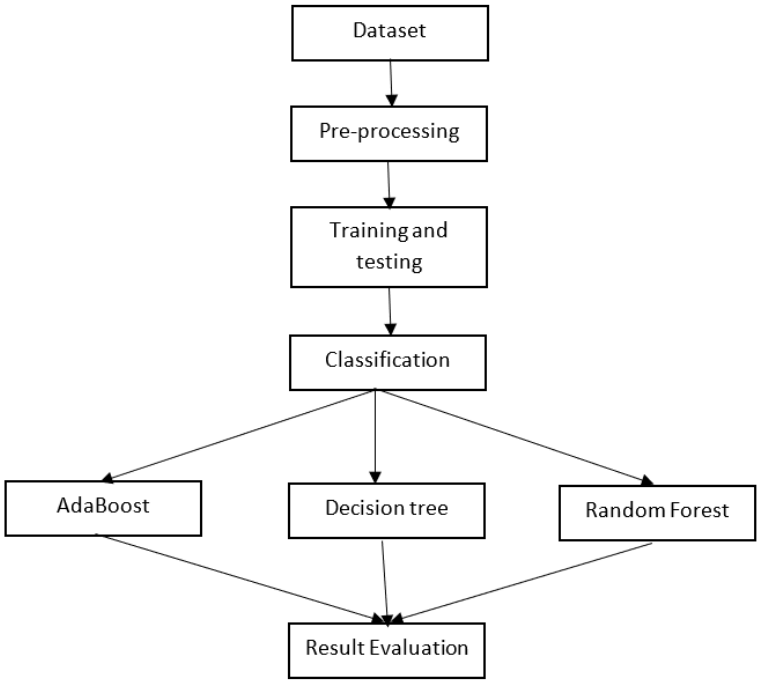

Figure 1. Research flow

A. Dataset

The dataset used in this study was obtained from the dataset sharing website, namely Kaggle.com with the link https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. The data used in the form of Ms. Excel, and the number of datasets used was 5100 with 11 attributes. **Figure 2** is an example of the dataset used. The attributes used can be seen in **Table 1**.

Tabel 1. Attributes Used

Atribute	
No	***Atribute***
1	id
2	gender
3	age
4	hypertension
5	heart_disease
6	ever_married
7	work_type
8	Residence_type
9	avg_glucose_level
10	bmi
11	smoking_status

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	gender	age	hypertens	heart_disease	ever_mar	work_type	Residence_type	avg_glucose_lev	bmi	smoking_status	stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
3	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
16	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
17	58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
18	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
19	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
20	27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
21	25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
22	70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
23	13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1
24	68794	Female	79	0	0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1

Figure 2. Example dataset

B. Pre-processing

There were still a lot of redundant, null, and missing data after the data collection stage was complete. So that pre-processing was needed to eliminate redundant and null data [23]. This pre-processing method used the Normalize to Interval [0,1] feature to normalize the feature. **Figure 3** is an example of a dataset after pre-processing.

Data Table - Orange

	gender	ever_married	work_type	Residence_type	smoking_status	bmi	id	age	hypertension	heart_disease	avg_glucose_level	stroke
1	Male	Yes	Private	Urban	formerly smoked	36.6	0.1232144	0.816895	0.000	1.000	0.8012649	1
2	Female	Yes	Self-employed	Rural	never smoked	N/A	0.7082047	0.743652	0.000	0.000	0.6790232	1
3	Male	Yes	Private	Rural	never smoked	32.5	0.4260151	0.975586	0.000	1.000	0.2345120	1
4	Female	Yes	Private	Urban	smokes	34.4	0.8249283	0.597168	0.000	0.000	0.5360078	1
5	Female	Yes	Self-employed	Rural	never smoked	24	0.0219286	0.963379	1.000	0.000	0.5493491	1
6	Male	Yes	Private	Urban	formerly smoked	29	0.7767211	0.987793	0.000	0.000	0.6051611	1
7	Male	Yes	Private	Rural	never smoked	27.4	0.7384765	0.902344	1.000	1.000	0.0691072	1
8	Female	No	Private	Urban	never smoked	22.8	0.1422612	0.841309	0.000	0.000	0.1812852	1
9	Female	Yes	Private	Rural	Unknown	N/A	0.3753379	0.719238	0.000	0.000	0.0970824	1
10	Female	Yes	Private	Urban	Unknown	24.2	0.8291686	0.951172	0.000	0.000	0.0159265	1
11	Female	Yes	Private	Rural	never smoked	29.7	0.1652464	0.987793	1.000	0.000	0.1168406	1
12	Female	Yes	Govt_job	Rural	smokes	36.8	0.1650543	0.743652	0.000	1.000	0.3016342	1
13	Female	Yes	Private	Urban	smokes	27.3	0.1661521	0.658203	0.000	0.000	0.2280030	1
14	Male	Yes	Private	Urban	Unknown	N/A	0.1117835	0.951172	0.000	1.000	0.7604099	1
15	Female	Yes	Private	Urban	never smoked	28.2	0.0720431	0.963379	0.000	1.000	0.7338658	1
16	Female	Yes	Self-employed	Rural	never smoked	30.9	0.7977577	0.609375	1.000	0.000	0.5183732	1
17	Male	Yes	Private	Urban	smokes	37.5	0.7690777	0.780273	0.000	1.000	0.6300896	1
18	Male	Yes	Private	Urban	smokes	25.8	0.4672924	0.914551	1.000	0.000	0.7671037	1
19	Female	No	Private	Urban	never smoked	37.8	0.3758731	0.731445	0.000	0.000	0.1574185	1
20	Male	No	Govt_job	Urban	Unknown	N/A	0.3452445	0.694824	0.000	1.000	0.7476687	1
21	Female	Yes	Govt_job	Rural	smokes	22.4	0.9683010	0.865723	0.000	0.000	0.6408457	1
22	Female	Yes	Self-employed	Urban	never smoked	48.9	0.1892882	0.633789	1.000	0.000	0.8225002	1
23	Female	Yes	Self-employed	Urban	never smoked	26.6	0.9431065	0.963379	0.000	0.000	0.8013111	1
24	Male	Yes	Private	Rural	Unknown	32.5	0.8879969	1.0000	0.000	1.000	0.7071369	1
25	Male	Yes	Private	Urban	formerly smoked	27.2	0.0569758	0.865723	0.000	0.000	0.2204321	1
26	Male	Yes	Self-employed	Rural	never smoked	23.5	0.9709357	0.975586	0.000	0.000	0.2262026	1

Figure 3. Example of a dataset after pre-processing

C. Training and Testing

The training phase was carried out to provide training to the model used, while for testing was conducted to provide a trial of the model used. The training data used in this study was 80% of the total data, namely 4088, and the training data used was 20% of the total data, 1022. Then the class used was 2 classes, namely 0 = Normal and 1 = Stroke.

D. Classification

Classification was carried out to obtain classification results from the proposed model. The models proposed in this study were AdaBoost, Decision Tree and Random Forest. The tools used to process the data store were using Orange Data Mining Tools. Orange Data Mining was a data mining tool that gave better results than other tools [24]–[27]. **Figure 4** shows the flow of the classification process.

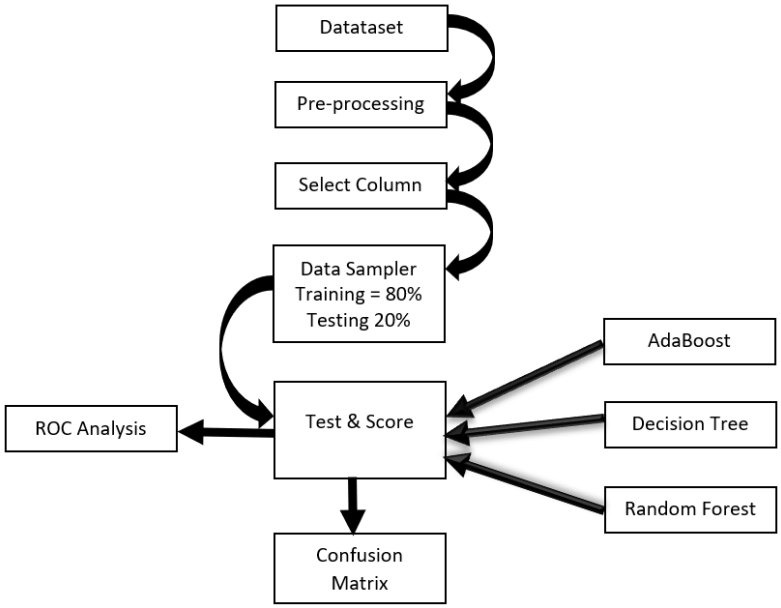

Figure 4. Classification process flow

E. Result Evaluation

The results of the classification in this study were evaluated using a confusion matrix and ROC analysis, the evaluation of the results was carried out to determine the performance of the proposed model [23]. Confusion matrix and ROC analysis obtained the classification results from tests and scores, where tests and scores were first given learning by the AdaBoost, decision tree and random forest models. **Table 2** is an example of the confusion matrix used.

Tabel 2. Confusion Matrix

Actual	**Predicted**	
	Normal	***Stroke***
Normal	True Positives (TP)	False Negatives (FN)
Stroke	False Positives (FP)	True Negatives (TN)

Accuracy was measured using [28]:

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

Precision was measured using [29]:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP+FP}} \quad (2)$$

Recall was measured using [29]:

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}} \quad (3)$$

Results and Discussion

A. Number of Folds 5

The results of the classification test using Number of Folds 5 using the AdaBoost model get an accuracy of .917, then for the decision tree model of .953 and the random forest model of .950, as for the complete test results can be seen in **Table 3**.

Table 3. Number of Folds 5

No	Number of Folds 5				
	Model	***AUC***	***Precision***	***Recall***	***Accuracy***
1	AdaBoost	0.540	0.917	0.917	0.917
2	Decision Tree	0.498	0.907	0.953	0.953
3	Random Forest	0.729	0.923	0.950	0.950

The evaluation used a confusion matrix from the AdaBoost model using Number of Folds 5. The classification model results using AdaBoost were class 0 (Normal) = 3723, and class 1 (Stroke) = 170. As for those that failed to be classified for class 0 (Normal) = 171, and class 1 (Stroke) = 24. **Figure 5** is the result of the confusion matrix of the AdaBoost model.

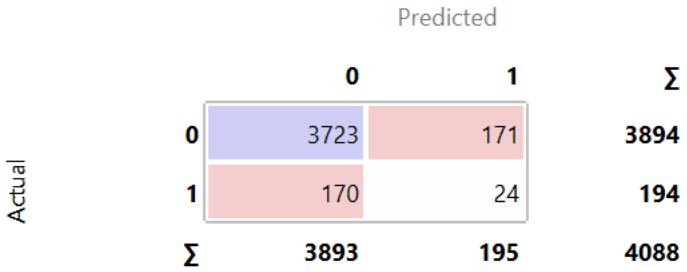

Figure 5. Confusion Matrix with AdaBoost Model

The evaluation used a confusion matrix from the Decision Tree model using Number of Folds 5. The classification model using Decision Tree showed that the classification results were class 0 (Normal) = 3894, and class 1 (Stroke) = 194. As for those that failed to be classified for class 0 (Normal) = 0, and class 1 (Stroke) = 0. **Figure 6** is the result of the confusion matrix of the Decision Tree model.

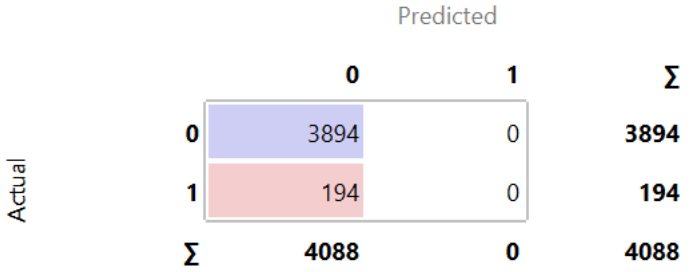

Figure 6. Confusion matrix with decision tree model

The evaluation used a confusion matrix from the Random Forest model using Number of Folds 5. The classification model results using Random Forest obtained class 0 (Normal) = 3878, and class 1 (Stroke) = 187. As for those who fail to be classified for class 0 (Normal) = 16, and for class 1 (Stroke) = 7. **Figure 7** is the result of the confusion matrix of the Random Forest model.

Actual \ Predicted	0	1	Σ
0	3878	16	3894
1	187	7	194
Σ	4065	23	4088

Figure 7. Confusion matrix with random forest model

Figure 8 is the evaluation result using ROC Analysis by calculating the Area Under Curve (AUC) with a target of 0 (Normal) using Number of Folds 5. Based on the evaluation results of the AUC with Number of Folds 5 in the AdaBoost model, the threshold result was 1,000. As for the result of the evaluation of the AUC with the Decision Tree model was .953, and the result of the evaluation of the AUC of the Random Forest model was .990.

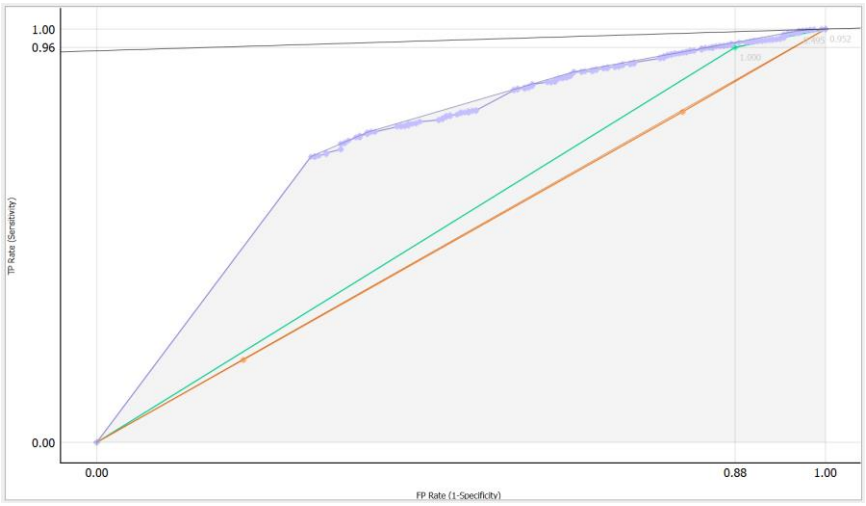

Figure 8. ROC Analysis Target 0 (Normal)

Figure 9 is the result of the evaluation using ROC Analysis by calculating the Area Under Curve (AUC) with a target of 1 (Stroke) using Number of Folds 5. Based on the evaluation results of the AUC with Number of Folds 5 in the AdaBoost model, the threshold result was 1,000. As for the evaluation result from the AUC with the Decision Tree model was .048, and the AUC evaluation result from the Random Forest model was .010.

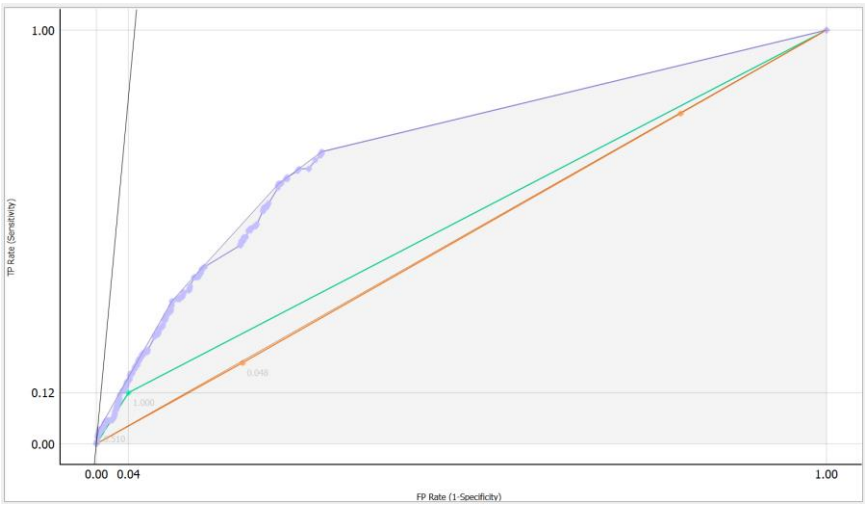

Figure 9. ROC Analysis Of Target 1 (Stroke)

B. Number of Folds 10

The accuracy results of the classification test using Number of Folds 10 using the AdaBoost model was .915, the decision tree model was .953 and the random forest model was .948, as for the complete test results of the three proposed models can be seen in **Table 4**.

Table 4. Number of Folds 10

No	Number of Folds 5				
	Model	***AUC***	***Precision***	***Recall***	***Accuracy***
1	AdaBoost	0.534	0.918	0.915	0.915
2	Decision Tree	0.494	0.907	0.953	0.953
3	Random Forest	0.720	0.912	0.948	0.948

The evaluation using a confusion matrix with the AdaBoost model using Number of Folds 10 has successfully classified class 0 (Normal) = 3715, and class 1 (Stroke) = 167. As for those that failed to be classified, class 0 (Normal) = 179 , and for class 1 (Stroke) = 27. **Figure 10** is the result of the confusion matrix of the AdaBoost model.

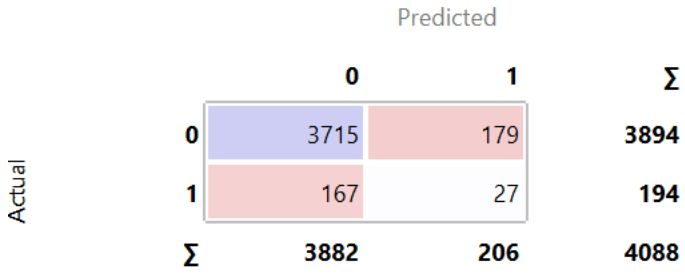

Figure 10. Confusion Matrix with AdaBoost Model

The evaluation using a confusion matrix with a Decision Tree model using Number of Folds 10 has successfully classified for class 0 (Normal) = 3894, and class 1 (Stroke) = 194. As for those that failed to be classified for class 0 (Normal) = 0, and for class 1 (Stroke) = 0. **Figure 11** is the result of the confusion matrix of the Decision Tree model.

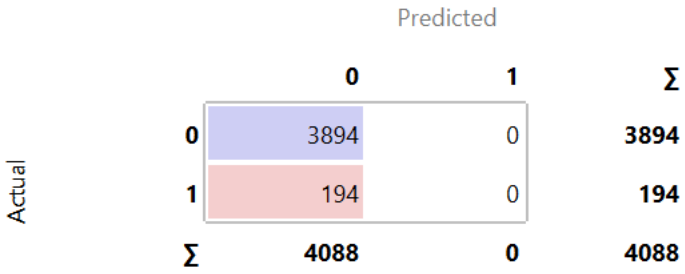

Figure 11. Confusion matrix with decision tree model

The evaluation using a confusion matrix with the Random Forest model using Number of Folds 10 has successfully classified for class 0 (Normal) = 3875, and class 1 (Stroke) = 192. As for those that failed to be classified, class 0 (Normal) = 19, and for class 1 (Stroke) = 2. **Figure 12** is the result of the confusion matrix of the Random Forest model.

Actual \ Predicted	0	1	Σ
0	3875	19	3894
1	192	2	194
Σ	4067	21	4088

Figure 12. Confusion matrix with random forest model

The evaluation used ROC Analysis by calculating the Area Under Curve (AUC) with a target of 0 (Normal) using Number of Folds 10. Based on the evaluation result of the AUC on the AdaBoost model, the threshold result was 1,000. As for the evaluation result with the Decision Tree model was .952, and the AUC evaluation result from the Random Forest model was 1,000. **Figure 13** is a graph using ROC analysis.

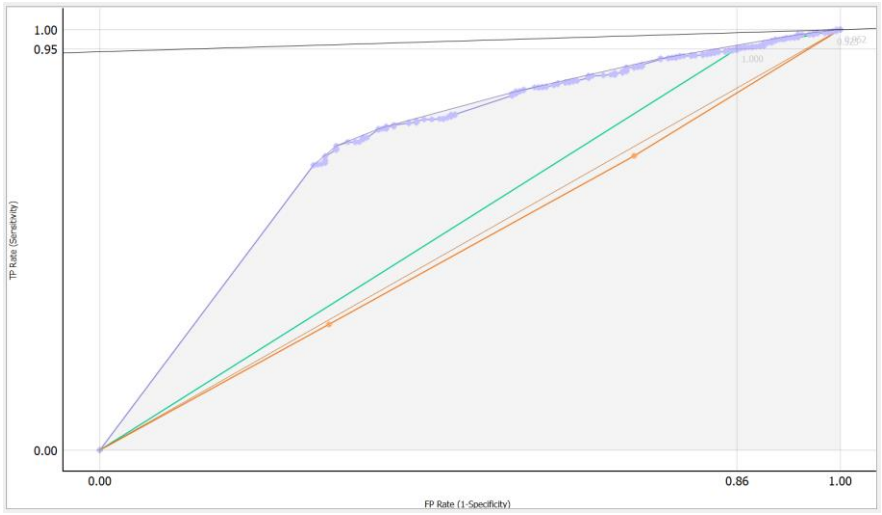

Figure 13. ROC Analysis Target 0 (Normal)

The evaluation used ROC Analysis by calculating the Area Under Curve (AUC) with a target of 1 (Stroke) using Number of Folds 10. Based on the evaluation result of the AUC on the AdaBoost model, the threshold result was 1,000. As for the evaluation result using the Decision Tree model was .0048, and the AUC evaluation result from the Random Forest model was .011. **Figure 14** is a graph using ROC analysis.

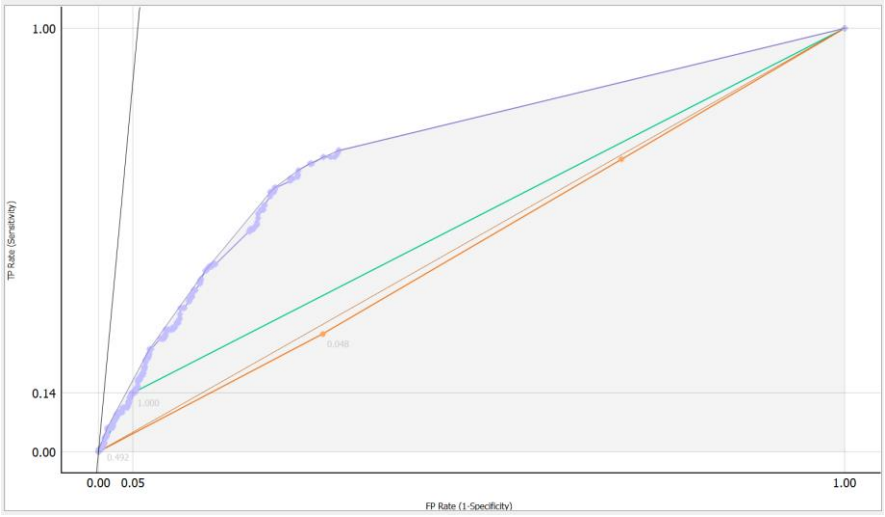

Figure 14. ROC Analysis Target 1 (Stroke)

C. Discussion

Referring the results of tests using Number of Folds 5 and Number of Folds 10, the highest accuracy was the Decision Tree model with .953, while for Number of Folds 10 the Decision Tree model showed that the highest accuracy was .953. This research has also carried out using Number of Folds 2, 3, and 20, but the results obtained were inconsistent and low, so we decided to use Number of Folds 5 and Number of Folds 10. From the evaluation results using the Confusion Matrix model, Decision Tree showed the highest accuracy result. As for the evaluation result, using ROC Analysis on Number of Folds 5 at target 0 (Normal) showed that AdaBoost was the model with the highest threshold in contrary the Decision tree was the lowest. While for target 1 (stroke), the model that gave the highest threshold result was AdaBoost and the lowest was Random Forest. The result of the evaluation using ROC Analysis on Number of Folds 10 at target 0 (Normal) indicated that AdaBoost model showed the highest threshold whereas Random Forest was the lowest, while for target 1 (stroke) the model that obtained the highest threshold result was AdaBoost and the lowest was Random Forest.

Furthermore, the comparison of the results of the methods used in this study with other relevant studies using data mining and different classification models by utilizing stroke data. The results of the comparison can be seen in **Table 5**.

Table 5. Comparison of the proposed model with other relevant research

No	Comparison Table With Other Studies		
	Method	***Classifier***	***Accuracy (%)***
1	Machine Learning [2]	Decision Tree	0.9113
		Logistic Regression	0.955
		Random Forest	0.955
		Support Vector Machine	0.9243
		KNN	0.9524
2	Data Mining [4]	C4.5	95.42
		KNN with K=1	94.18
		KNN with K=3	92.81
		KNN with K=7	93.06
		KNN with K=11	91.82
3	Data Mining [13]	MLP	89.40
		Jrip	92.60
		C4.5	85.50
4	Machine Learning [16]	**Without data resampling**	
		Naïve Bayes	0.82
		BayesNet	0.82
		Decision tree	0.83
		Random forest	0.86
		Data imputation	
		Naïve Bayes	0.81
		BayesNet	0.86
		Decision tree	0.88
		Random forest	0.90
		Data resampling	
		Naïve Bayes	0.82
		BayesNet	0.87
		Decision tree	0.93
		Random forest	0.96
5	Machine Learning [17]	XGBoost	93.68
		Random Forest	91.14
		KNN	88.69
		Logistic Regression	81.18
		SVM	81.10
6	Machine Learning [20]	Logistic Regression	78
		Decision Tree	66
		Random Forest	73
		KNN	80
		SVM	80
		Naïve Bayes	82
7	Our Method	**Number of Folds 5**	
		AdaBoost	0.917
		Decision Tree	**0.953**
		Random Forest	0.950

No	Comparison Table With Other Studies		
	Method	*Classifier*	*Accuracy (%)*
		Number of Folds 10	
		AdaBoost	0.915
		Decision Tree	**0.953**
		Random Forest	0.948

What is new in this study is that we propose a classification system to classify Stroke and normal patients using the AdaBoost, Decision Tree and Random Forest models. Currently, research with the proposed classification model has never been done by previous researchers, from the results of trials conducted using Number of Folds 5 and 10, the model that provides the best results is the Decision Tree.

Conclusion

The data used in this study were 5100 with a distribution of 80% training and 20% testing. There were 11 attributes used and the evaluation results used confusion matrix and ROC analysis. Based on the results of the tests that have been carried out the proposed model was able to provide good classification result. From the proposed model, the Decision Tree model obtained the highest accuracy result, namely .953 for Number of Folds 5, and .953 for Number of Folds 10. The Decision Tree model showed consistent results even though the Number of Folds used was different. From this study, the decision tree can provide a good classification for stroke classification, the results of this study depend on the dataset used. For further research, the proposed model should be different or combine the decision tree model with other models.

Acknowledgement

We would like to extend our sincere thanks to the Institute for Research and Community Service, University of Technology Mataram, for the funds provided for this research.

References

[1] *et al.*, "Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke," *Commun. Med.*, vol. 1, no. 1, pp. 1–18, 2021, doi: 10.1038/s43856-021-00062-8.

[2] H. K. V, H. P, G. Gupta, P. Vaishak, and P. K. B, "Stroke prediction using Machine Learning algorithms," *Int. J. Innov. Res. Eng. Manag.*, vol. 8, no. 4, pp. 553–558, 2021, doi: 10.1109/Confluence52989.2022.9734197.

[3] N. H. Ali, A. R. Abdullah, N. M. Saad, A. S. Muda, T. Sutikno, and M. H. Jopri, "Brain stroke computed tomography images analysis using image processing: A review," *IAES Int. J. Artif. Intell.*, vol. 10, no. 4, pp. 1048–1059, 2021, doi: 10.11591/IJAI.V10.I4.PP1048-1059.

[4] L. Amini *et al.*, "Prediction and control of stroke by data mining," *Int. J. Prev. Med.*, vol. 4, pp. S245–S249, 2013.

[5] A. S, J. V, H. N, and D. M, "An artificial intelligence approach for predicting different types of stroke," *Int. J. Eng. Res. Technol.*, vol. 8, no. 12, pp. 1858–1861, 2019, doi: 10.1109/ICICCT.2018.8473057.

[6] Y. Yu *et al.*, "Use of Deep Learning to predict final ischemic stroke lesions from Initial Magnetic Resonance imaging," *JAMA Netw. Open*, vol. 3, no. 3, pp. 1–13, 2020, doi: 10.1001/jamanetworkopen.2020.0772.

[7] T. Badriyah, D. B. Santoso, I. Syarif, and D. R. Syarif, "Improving stroke diagnosis accuracy using hyperparameter optimized deep learning," *Int. J. Adv. Intell. Informatics*, vol. 5, no. 3, pp. 256–272, 2019, doi: 10.26555/ijain.v5i3.427.

[8] S. Surya, B. Yamini, T. Rajendran, and K. E. Narayanan, "A Comprehensive method for identification of stroke using Deep Learning," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 7, pp. 647–652, 2021.

[9] A. B. URAL, "Computer aided Deep Learning based assessment of stroke from brain Radiological CT Images," *Eur. J. Sci. Technol.*, no. 34, pp. 42–52, 2022, doi: 10.31590/ejosat.1063356.

[10] S. Zhang, S. Xu, L. Tan, H. Wang, and J. Meng, "Stroke lesion detection and analysis in MRI Images based on Deep Learning," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/5524769.

[11] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, "Stroke disease

detection and prediction using Robust Learning Approaches," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/7633381.

[12] J. Yu, S. Park, S. H. Kwon, C. M. B. Ho, C. S. Pyo, and H. Lee, "AI-based stroke disease prediction system using real-time electromyography signals," *Appl. Sci.*, vol. 10, no. 19, 2020, doi: 10.3390/app10196791.

[13] O. Almadani and R. Alshammari, "Prediction of stroke using data mining classification techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 1, pp. 457–460, 2018, doi: 10.14569/IJACSA.2018.090163.

[14] A. K. Arslan, C. Colak, and M. E. Sarihan, "Different medical data mining approaches based prediction of ischemic stroke," *Comput. Methods Programs Biomed.*, vol. 130, pp. 87–92, 2016, doi: 10.1016/j.cmpb.2016.03.022.

[15] M. W. Hassan, A. Keshk, A. A. El-atey, and E. Alfeky, "Brain stroke detection using Tensor Factorization and Machine Learning models," *Int. J. Eng. Technol. Manag. Res.*, vol. 8, no. 8, pp. 1–12, 2021, doi: 10.29121/ijetmr.v8.i8.2021.1006.

[16] E. M. Alanazi, A. Abdou, and J. Luo, "Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models," *JMIR Form. Res.*, vol. 5, no. 12, pp. 1–10, 2021, doi: 10.2196/23440.

[17] A. Kshirsagar, H. Goyal, S. Loya, and A. Khade, "Brain stroke prediction portal using Machine Learning," *Int. J. Res. Eng. Appl. Manag.*, vol. 07, no. 03, 2021.

[18] M. G. M and D. P. M. C, "IRJET- Stroke Type Prediction using Machine Learning and Artificial Neural Networks," *Int. Res. J. Eng. Technol.*, vol. 8, no. 6, 2021.

[19] T. S. Heo *et al.*, "Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI," *J. Pers. Med.*, vol. 10, no. 4, pp. 1–11, 2020, doi: 10.3390/jpm10040286.

[20] 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.

[21] E. R. Kaur and V. Chopra, "Implementing adaboost and enhanced adaboost algorithm in Web Mining," *Int. J. Adanced Res. Comput. Commun. Eng.*, vol. 4, no. 7, pp. 306–311, 2015, doi: 10.17148/IJARCCE.2015.4771.

[22] H. H. Patel and P. Prajapati, "Study and analysis of decision tree based classification Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 56–61, 2018.

[23] B. Imran, Zaeniah, Sriasih, S. Erniwati, and Salman, "Data Mining using a Support Vector Machine , Decision Tree , Logistic Regression and Random Forest for," *J. INFOKUM*, vol. 10, no. 2, pp. 792–802, 2022.

[24] B. Imran, H. Hambali, A. Subki, Z. Zaeniah, A. Yani, and M. R. Alfian, "Data Mining using Random Forest, Naïve Bayes, and Adaboost Models for prediction and classification of benign and malignant breast cancer," *J. Pilar Nusa Mandiri*, vol. 18, no. 1, pp. 37–46, 2022, doi: 10.33480/pilar.v18i1.2912.

[25] S. Kodati and R. Vivekanandam, "Analysis of heart disease using in Data Mining tools orange and weka," *Glob. J. Comput. Sci. Technol. C Softw. Data Eng.*, vol. 18, no. 1, pp. 16–22, 2018.

[26] M. S. Kukasvadiya, D. Nidhi, and H. Divecha, "Analysis of data using Data Mining tool orange," *Int. J. Eng. Dev. Res.*, vol. 5, no. 2, pp. 1836–1840, 2017, [Online]. Available: www.ijedr.org.

[27] G. Manimannan, R. L. Priya, and C. A. Kumar, "Application of orange data mining approach of argiculture productivity index performance in Tamilnadu," *Int. J. Sci. Innov. Math. Res.*, vol. 7, no. 8, pp. 8–16, 2019, doi: 10.20431/2347-3142.0708003.

[28] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *J. Algorithms Comput. Technol.*, vol. 12, no. 2, pp. 119–126, 2018, doi: 10.1177/1748301818756225.

[29] S. M. Ayyoubzadeh, A. Almasizand, and ..., "Early Breast cancer prediction using Dermatoglyphics: Data Mining Pilot Study in a General Hospital in Iran," *Heal. Educ. ...*, vol. 9, no. 3, pp. 279–285, 2021, [Online]. Available: https://biot.modares.ac.ir/article-5-53673-en.html.