Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 4 (2024) 555 - 562 e-ISSN: 2580-0760

Increasing the Accuracy of Brain Stroke Classification using Random Forest Algorithm with Mutual Information Feature Selection

Fachruddin[1*], Errissya Rasywir[2], Yovi Pratama[3]
[1]Information Systems Department, Computer Science Faculty, Dinamika Bangsa University, Jambi, Indonesia
[2,3]Informatics Engineering Department, Computer Science Faculty, Dinamika Bangsa University, Jambi, Indonesia
[1]fachruddin.stikom@gmail.com, [2] errissya.rasywir@gmail.com, [3] yovi.pratama@gmail.com

Abstract

Brain stroke stands out as a leading cause of death, distinguishing it from common illnesses and highlighting the critical need to utilize machine learning techniques to identify symptoms. Among these techniques, the Random Forest (RF) algorithm emerged as the main candidate because of its optimal accuracy values. RF was chosen for its ensemble learning properties that optimize accuracy while simultaneously, bagging all outputs (DT), thus increasing its efficacy. Feature Selection, an important data analysis step, which is mainly achieved through pre-processing, aims to identify influential features and ignore less impactful features. Mutual Information serves as an important feature selection method. Specifically, the highest level of accuracy was achieved by cross-validating the test data - 10, resulting in 0.7760 without feature selection and 0.7790 with mutual information. Most of the attributes in the brain stroke dataset show relevance to the stroke disease class, but the resulting decision tree shows age as a particularly important node. So, the research results show that the selection feature (Mutual Information) can increase the accuracy of brain stroke classification, although it is not significant, namely an increase of 0.0030%. With an increase, where there is no significant difference, it can be said that almost all the attributes contained in the brain stroke dataset used have an influence on their relevance to the stroke disease class.

Keywords: *random forest; strokes; brain; mutual information; features*

How to Cite: F. Fachruddin, E. Rasywir, and Y. Pratama, "Increasing the Accuracy of Brain Stroke Classification using Random Forest Algorithm with Mutual Information Feature Selection", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 4, pp. 555 - 562, Aug. 2024.
DOI: https://doi.org/10.29207/resti.v8i4.5795

1. Introduction

Stroke is a type of disease that is most commonly suffered by Indonesian people and has a high mortality rate [1], [2]. Stroke plays a role in damaging or disrupting brain function, where blood flow to the human brain is disrupted and this makes it possible for patients affected by brain stroke to experience paralysis or death [2], [3]. In 2019, the World Health Organization (WHO) ranked brain stroke as seven of the ten main causes of death. The Ministry of Health classifies stroke as a catastrophic disease because of its broad economic and social impact [3], [4]. A brain stroke cannot be considered an ordinary disease because if you look at the symptoms and consequences of a brain stroke, this disease must be considered a serious disease [3]–[5]. Therefore, it is important for us to find out information about the symptoms or causes of brain stroke using one of the data mining techniques, namely classification and also being able to classify brain stroke itself [6], [7].

In related research, the classification of brain stroke diseases has been conducted using several methods, including an integrated machine learning approach used to select features as prognosis factors of stroke on The International Stroke Trial (IST) dataset [3], yielding good results. Other studies related to this have also shown that Machine Learning (ML)[3] delivers accurate and quick prediction outcomes, becoming a powerful tool in healthcare settings, providing personalized clinical care for stroke patients [8]. Generally, it can be said that the methods used in previous research include Recursive Feature Elimination with cross-validation (RFECV) [3], [9], [10], Random Forest Classifier [3], [6], [7], Extra Trees Classifier [3], [4], [11], [12], AdaBoost Classifier [3], [13], [14], and Multinomial Naïve Bayes Classifier [3], [14]–[16], along with the Random Forest Classifier and Shapiro Wilk algorithm [3].

In this research, the author will use the Random Forest Algorithm because the Random Forest Algorithm is very good for solving cases of problems that involve

Received: 03-05-2024 | Accepted: 13-08-2024 | Published Online: 28-08-2024

data mining techniques, namely classification. Data mining is a process of searching for patterns or even interesting information in selected data using certain methods according to what is needed or desired because the methods or algorithms in data mining are very diverse and have their own advantages based on each goal [3], [6], [7].

Classification is the activity of assessing an object based on data and then grouping it by assigning categories to the available classes [5], [17]–[20]. Classification is a type or type of data analysis activity that is useful for making it easier to determine the label class of the object to be classified [17], [21]–[24]. There are also previous researchers who conducted research using similar algorithms, such as research on the Classification of Stroke Diseases [25], Classifying Student Achievement Index [26], Classification of Attack Detection in Network Protocols [27], [28], Classification of Marketing Personnel Placement [29], Classification of Online Shop Buyer Satisfaction Levels[30]. On the dataset that will be used, the author will carry out classification with an algorithm and find out the accuracy results with 2 different types of tools, as well as get a visual of the shape of the tree that will be produced.

Feature selection is a technique for selecting important and relevant features of data and reducing irrelevant features. Feature selection aims to select the best features from a feature data set. Feature Selection is a modelling or data analysis activity which can generally be carried out using preprocessing and aims to select influential features (optimal features) and exclude features that have no influence. There are many types of feature selection methods used in previous research, including Chi-Square, Information Gain, Mutual Information, weighting using the hashing method and others. In this study, we propose that it is necessary to carry out feature selection to see the attributes that have the most influence on brain stroke classification using the mutual information method.

2. Research Methods

The stages of this research are the steps taken to solve the problem of Increasing the Accuracy of Brain Stroke Prediction using The Random Forest Algorithm with Mutual Information Feature Selection. The research stages used can be seen in Figure 1.

2.1. The Research Architecture

The formulation of the problem in this research is how to apply data mining by Increasing the Accuracy of Brain Stroke Prediction using The Random Forest Algorithm with Mutual Information Feature Selection. The aim to be achieved in this research is to find out how much accuracy Increasing the Accuracy of Brain Stroke Prediction using The Random Forest Algorithm with Mutual Information Feature Selection. Study literature that can achieve research objectives, literature sourced from journals, electronic books, and of course information from the internet such as websites. The literature used will be attached to the bibliography. The schema of this research can be seen in Figure 1.

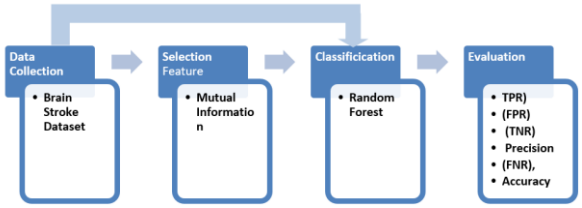

Figure 1. Flow of Brain Stroke Classification using Random Forest Algorithm with Mutual Information Feature Selection Research.

In Figure 1, the research flow of the entire scheme to be conducted is displayed. The following section will explain steps or research methodologies including data collection steps and information related to Brain Stroke Classification with Random Forest experimented with feature selection using Mutual Information and without feature selection. Furthermore, the classification results will be explained with evaluation parameters. The parameters calculated for this stroke classification test include Recall Sensitivity True Positive Rate (TPR), False Positive Rate (FPR), False Alarm rate, Specificity True Negative Rate (TNR), Precision, False Negative Rate (FNR), and Accuracy.

Data and Information Collection: In data collection, the author obtained a dataset online which is found on a website on the internet, namely the Kaggle.com website. *The dataset* that has been used in the Increasing the Accuracy of Brain Stroke Prediction using The Random Forest Algorithm with Mutual Information Feature Selection research is " *Brain* Stroke Dataset ", this data has not been cleaned and has a total of 4,981 data with 11 attributes, namely Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Avg Glucose Level, BMI, Smoking Status, and the last one as label, namely Stroke. From all the data obtained, it will enter the *data cleaning process* to remove *noise* or defects in the data. After going through the *data cleaning process*, the author will use *the holdout* data as *training data* and *testing data*. In research, using 2/3 *training data* and 1/3 *testing data*. In general, the proportion *of training data* and *testing data.*

Table 1. *Brain Stroke Dataset* (Kaggle)

No	Gender	Age	Hyper tension	Heart Disease	Ever Married
1	Male	67	0	1	Yes
2	Male	80	0	1	Yes
3	Female	49	0	0	Yes
4	Female	79	1	0	Yes
5	Male	81	0	0	Yes
….	….	….	….	….	….
4981	Female	80	1	0	Yes

Tables 1 and 2 are excerpts of the Brain Stroke Dataset from Kaggle. Some of the attributes include Gender, Age, Hypertension, Heart Disease, Ever Married, Residence Type, Avg_Glucose Level, BMI, Smoking Status, and Stroke. The data includes both nominal and

numeric types, which can be effectively handled by the Random Forest Algorithm.

Table 2. Kaggle Dataset Attributes

Residence Type	Avg_ Glucose Level	Bmi	Smoking Status	Stroke
Urban	228.69	36.6	formerly smoked	1
Rural	105.92	32.5	never smoked	1
Urban	171.23	34.4	smokes	1
Rural	174.12	24	never smoked	1
Urban	186.21	29	formerly smoked	1
….	….	….	….	….

Selection Feature using Mutual Information: Mutual information (MI) is used as feature selection because it can measure random dependencies between variables, so it is suitable for assessing features of information content in classification tasks. Previous research has proven that classification using MI feature selection is able to produce an accuracy rate of 87%. One feature selection that is often used to calculate the weight of features is mutual information. In MI, it can be seen how much information the presence or absence of a feature contributes to making a correct or incorrect classification decision. Formula 1 is the MI formula [31].

$$I(U;C) = \sum_{ec\in\{1,0\}} P(U=et, C=ec) log2 \frac{P(U=et,C=ec)}{P(U=et)P(C=ec)} \quad (1)$$

U is the random variable with et values; et =1 is the Instance that contains feature t; et = 0 is the Instance that does not contain feature t; C is the Random variable with ec values; ec = 1 is the Instance in class c; ec = 0 is the Instance not in class c.

Table 3. Example of Attribute Matrix with et and ec values

K*	1	1	0	0	1	et		ec		et = 1	et = 1	et = 0	et = 0
Features	A1	A2	A3	A4	A5	1	0	1	0	ec = 1	ec = 0	ec = 1	ec = 0
X	0	1	1	1	0	3	2	3	2	1	2	2	0
Y	1	0	0	0	1	2	3	3	2	2	0	1	2
Z	1	1	0	1	0	3	2	3	2	2	1	1	1

Table 3 represents a simulated example of Mutual Information calculation. The Mutual Information Selection Feature formula involves parameters such as the random variable with values "et," which denotes instances containing feature "t," the instances that do not contain feature "t," the random variable with values "ec," representing instances in class "c," and instances not in class "c." This formula helps in determining the mutual information between features and the target class in the dataset.

Classification: In this section, there is a process of extracting knowledge stored in the large volume Brain Stroke dataset. To obtain knowledge of the dataset, you can use the Random Forest algorithm. *The Random Forest algorithm* is developed by J. Ross Quinlan, with a system in the form of a tree with a branching form that starts with the most significant attribute and continues until there are several branches until the rules are complete. The flow of the *Random Forest Algorithm method* is as seen in Figure 2.

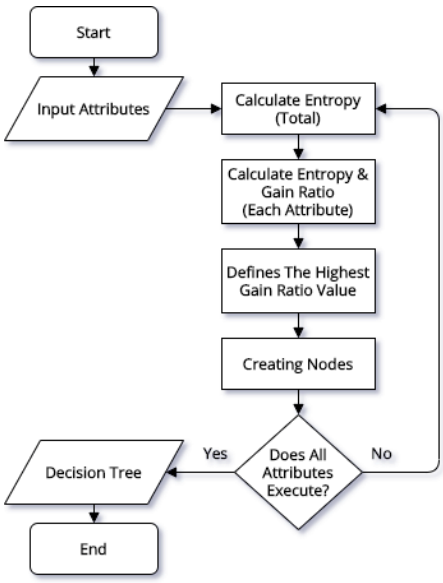

Figure 2. Flow of the Decision Tree Algorithm (C4.5) in Random Forest.

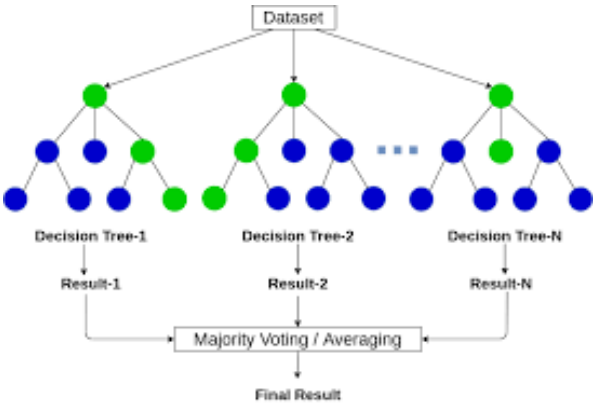

Figure 3. Merging Decision Trees in Random Forest [32]

Picture 3 is a visualization of the combination of decision trees used in the random forest algorithm in the research. The decision tree used in "Increasing the Accuracy of Brain Stroke Classification using Random Forest Algorithm with Mutual Information Feature Selection" is the C4.5 algorithm. The Formula of the Decision Tree we use is shown in the next formula. The entropy and gain values for each attribute are calculated and from the highest gain to the starting node, to calculate the entropy and gain Formula 2 is used.

$$Entropy\ (s) = \sum_{i=1}^{n} -Pi * log_2 Pi \quad (2)$$

S is the set of k cases, n is the number of S partitions, pi is the proportion of Si to S.

$$Gain(S,A) = Entropy(s) - \sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy\ (Si) \quad (3)$$

S is the set of cases, A is the Attributes, n is the number of attributes A partitions, |Si| is the number of cases in partition I, and |S| is the number of cases in S.

Testing and Evaluation: The testing process carried out in this research applies a confusion matrix, which can calculate accuracy, precision and recall values to obtain classification results. The definitions and formulas for calculating accuracy, precision and recall values are shown in Formulas 4 – 9 [33].

Accuracy: Accuracy is the value of the closeness of the results in the classification or classification to the actual value.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Precision: Precision is a level of accuracy that shows the closeness of the difference in value each time it is repeated.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall: Recall is the value of the percentage of the data classification model to its actual class.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

False Positive Rate (FPR) False Alarm rate

$$FPR = \frac{FP}{TN+FP} \quad (7)$$

Specificity True Negative Rate (TNR)

$$TNR = \frac{TN}{TN+FP} \quad (8)$$

False Negative Rate (FNR)

$$FNR = \frac{FN}{FN+TP} \quad (9)$$

Description:
TP is the True Positive, TN is the True Negative, FP is the False Positive, FN is the False Negative

3. Results and Analysis

3.1 Data Analysis

At this stage, data analysis is carried out with the Data Selection and Preprocessing stages.

At the Data Selection stage, the author determines the decision for the dataset to be used. The data taken and used by the author is the "Brain Stroke" dataset obtained from the Kaggle website.

Next, we enter the Preprocessing stage, this stage the author carries out processes such as cleaning data and does not apply feature selection. Data Cleaning is the process of deleting or cleaning a collection of datasets that have noise, for example, duplicate data, data with unclear symbols, errors in writing data, or data that is not complete. This was done so that the dataset could be used properly and accurately and get maximum results, with a total of 4,981 raw data and after cleaning the total data became 3,481 data.

Data Transformation: At this stage, the process of changing the contents of the data to be used is carried out, such as the attributes:

Age: In this attribute, the data is categorized by age based on the age numbers from patient data such as 10-12 years (Children); 13-25 (Teenagers); 26-45 (Mature); and 46-82 (Old).

Hypertension: In this attribute, data containing the number 1 becomes " Yes " and 0 becomes " No ".

Heart Disease: In this attribute, data containing the number 1 becomes " Yes " and 0 becomes " No ".

Average Glucose Level: The data is categorized as a value less than 140 (Normal), a value 141-199 (Prediabetes), and a value above 200 (Diabetes).

BMI (Body Mass Index): In this attribute, the author also categorizes values less than 19 (Underweight), values 19-25 (Normal), values between 26-30 (Overweight), and values above 30 (Obesity).

Strokes (Class): In this attribute, the author changes the value of the number 1 to "Yes " and 0 to "No".

Data Splitting: Data splitting divides the dataset into two parts: training data, which is used to create the model, and testing data, which is employed for model evaluation. In this study, the data is split into three proportions, with a 60% split being one of them. The table below displays the proportions resulting from the data-splitting process. The Random Forest Algorithm computations are conducted using the Python programming language. The formula utilized for Random Forest algorithm calculation in this study is provided below. The outcomes of the Random Forest algorithm, obtained through the tools employed, include the utilization of 60% training data, and cross-validation with ratios of 10 and 5. Subsequently, the Random Forest algorithm results are presented.

Random Forest Testing with MI Feature Selection with Split 60%: The section displays excerpts of the visualization of trees generated with the highest accuracy from each evaluation scheme conducted, including using a 60% test data split test scheme, followed by cross-validation with ratios of 10 and 5.

Figure 4 is an excerpt of the tree generated from the highest classification result using a 60% data split scheme. With 60% of the training data, the Random Forest algorithm yielded a total of 38 nodes. The highest node is found in the "Age" attribute.

Random Forest testing with MI Feature selection with Cross Validation-10: Figure 5 is an excerpt of the tree generated from the highest classification result using a Cross Validation-10 scheme. With Cross Validation-10, the Random Forest algorithm yielded a total of 33 nodes. The highest node is found in the "Age" attribute.

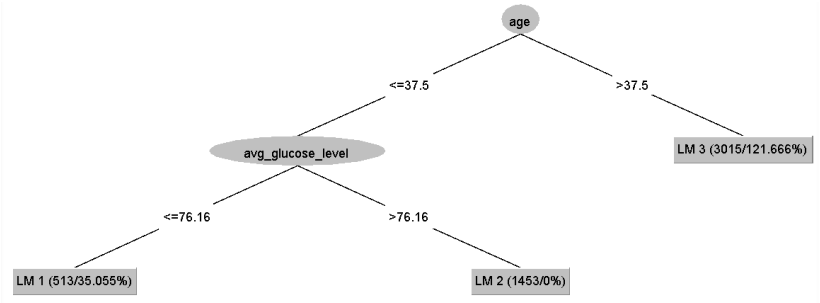

Figure 4. Snapshot of Classification Results with Random Forest with 60% Training Data.

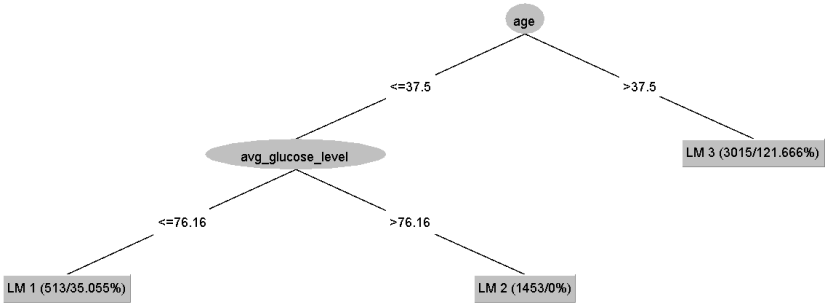

Figure 5. Random Forest with Mutual Information with Training Cross Validation-10

Random Forest testing with MI Feature selection with Cross Validation-5: Figure 6 is an excerpt of the tree generated from the highest classification result using a Cross Validation-5 scheme. With Cross Validation-5, the Random Forest algorithm yielded a total of 32 nodes. The highest node is found in the "Age" attribute.

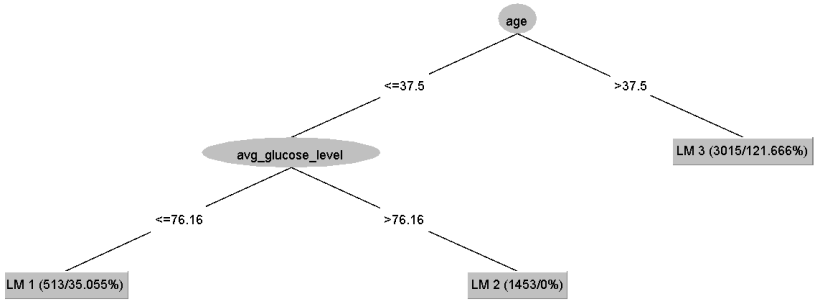

Figure 6. Random Forest with Mutual Information with Training Cross Validation-5

Evaluation of Random Forest Algorithm Classification: In this section, the classification results for stroke are displayed, and tests carried out include using classification with the Random Forest algorithm. The Decision tree that we use is the C.45 algorithm. The classification process is divided into tests subject to feature selection and without feature selection. By using the Mutual Information (MI) feature selection algorithm. The parameters calculated for this stroke classification test include Recall Sensitivity True Positive Rate (TPR), False Positive Rate (FPR) False Alarm rate, Specificity True Negative Rate (TNR), Precision and False Negative Rate (FNR), and Accuracy.

Table 4. Evaluation of TPR, FPR, Precision and Accuracy (60% Split Data Testing)

Evaluation	Split 60%		
Parameter	RF	RF+MI	Difference
(TPR)	0.8162	0.8162	0
(FPR)	0.8285	0.7571	-0.0714
Precision	0.8346	0.8393	0.0047
Accuracy	0.76175	0.76635	0.0046

In Table 4 and Figure 7, the evaluation results of Recall Sensitivity True Positive Rate (TPR), False Positive Rate (FPR), False Alarm rate, and Accuracy from the classification of brain stroke data using mutual information feature selection and without feature selection are displayed.

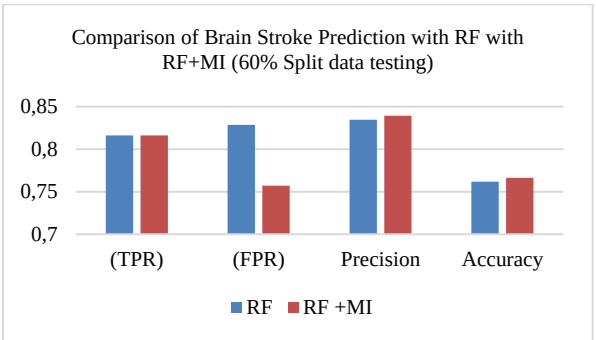

Figure 7. Comparison of Brain Stroke Prediction with RF with RF+MI (60% Split data testing).

By testing using split data of 60%, the highest result is accuracy using feature selection which is higher than without feature selection of 0.0046. The results in the table are visualised in Figure 6.

Table 5. Evaluation of TPR, FPR, Precision and Accuracy (CV-10)

Evaluation	CV 10		
Parameter	RF	RF+MI	Difference
(TPR)	0.8441	0.8346	-0.0094
(FPR)	0.7333	0.7125	-0.0208
Precision	0.8253	0.8346	0.0093
Accuracy	0.7760	0.7790	0.0030

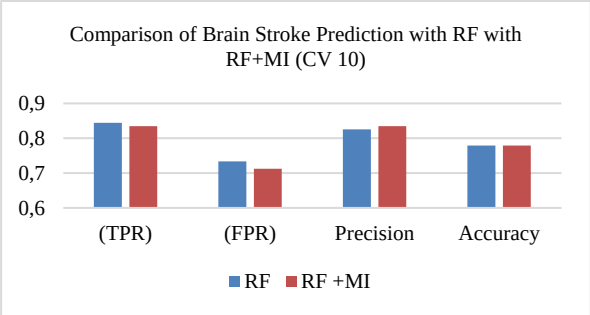

Figure 8. Comparison of Brain Stroke Prediction with RF with RF+MI (CV-10)

In Table 5 and Figure 8, the evaluation results of Recall Sensitivity True Positive Rate (TPR), False Positive Rate (FPR) False Alarm rate, and Accuracy from brain stroke data classification using mutual information feature selection and without feature selection are displayed. By testing using cross-validation of 10, the highest result is accuracy using feature selection which is higher than without feature selection of 0.0030. The results in the table are visualised in Figure 8.

Table 6. Evaluation of TPR, FPR, Precision and Accuracy (CV-5)

Evaluation	CV 5		
Parameter	RF	RF+MI	Difference
(TPR)	0.8207	0.8207	0
(FPR)	0.8230	0.7461	-0.0769
Precision	0.8393	0.8441	0.0047
Accuracy	0.7697	0.7744	0.0046

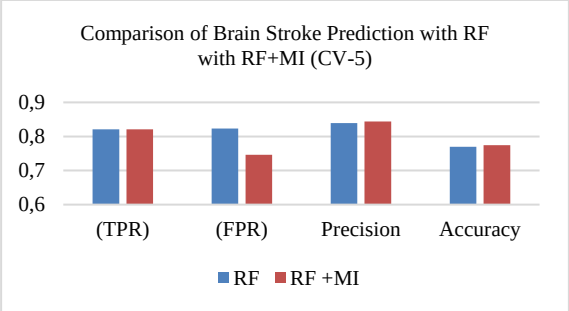

Figure 9. Comparison of Brain Stroke Prediction with RF with RF+MI (CV-5)

In Table 6 and Figure 9, the evaluation results of Recall Sensitivity True Positive Rate (TPR), False Positive Rate (FPR), False Alarm rate, and Accuracy from brain stroke data classification using mutual information feature selection and without feature selection are displayed. By testing using cross-validation of 5, the highest result is accuracy using feature selection which is higher than without feature selection of 0.0046. The results in the table are visualised in Figure 8.

Table 7. Evaluation of TNR, and FPR (60%)

Evaluation	Split 60%		
Parameter	RF	RF+MI	Difference
(TNR)	-0.0285	0.0428	0.0714
(FNR)	-0.0162	-0.0162	0

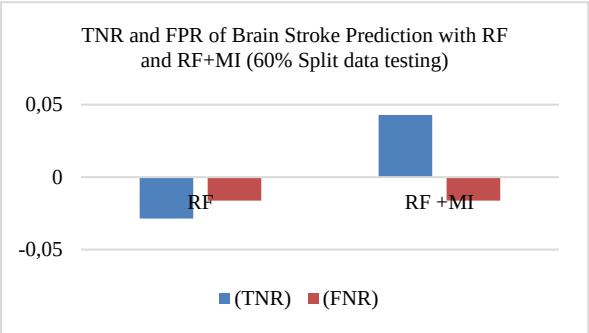

Figure 10. TNR & FPR of Comparison of Brain Stroke Prediction with RF with RF+MI (60% Split Data Testing)

Table 7 shows the evaluation results of Specificity True Negative Rate (TNR), Precision and False Negative Rate (FNR), from the classification of brain stroke data using mutual information feature selection and without feature selection. By using a split data test of 60%, the highest result was TNR using a higher feature selection than without a feature selection of 0.0714. The results in the table are visualised in Figure 10.

Table 8. Evaluation of TNR and FPR (CV-10)

Evaluation	CV 10		
Parameter	RF	RF+MI	Difference
(TNR)	0.0666	0.0875	0.0208
(FNR)	-0.0441	-0.0346	0.0094

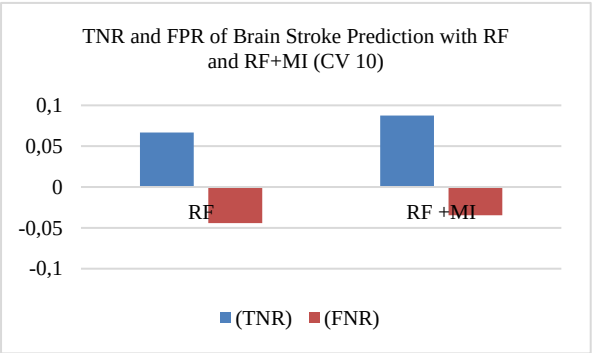

Figure 11. TNR & FPR of Comparison of Brain Stroke Prediction with RF with RF+MI (CV-10)

In Table 8 and Figure 11, the evaluation results of Specificity True Negative Rate (TNR), Precision and False Negative Rate (FNR) are displayed from the classification of brain stroke data using mutual information feature selection and without feature selection. By testing using cross validation-10, the highest result is TNR using feature selection which is higher than without feature selection of 0.0208. The results in the table are visualised in Figure 11.

Table 9. Evaluation of TNR and FPR (CV-5)

Evaluation	CV 5		
Parameter	RF	RF+MI	Difference
(TNR)	-0.0230	0.0538	0.0769
(FNR)	-0.0207	-0.0207	0

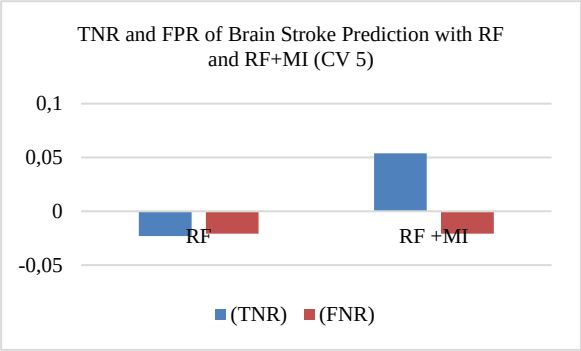

Figure 12. TNR & FPR of Comparison of Brain Stroke Prediction with RF with RF+MI (CV-5)

In Table 9 and Figure 12, the evaluation results of Specificity True Negative Rate (TNR), Precision and False Negative Rate (FNR) are displayed from the classification of brain stroke data using mutual information feature selection and without feature selection. By testing using cross validation-10, the highest result was TNR using feature selection which was higher than without feature selection of 0.0769.

4. Conclusions

Finally, the classification results for stroke, tests carried out include using classification with the Random Forest algorithm with the decision that we use is the C.45 algorithm. The classification process is divided into tests subject to feature selection and without feature selection. By using the Mutual Information (MI) feature selection algorithm. The parameters calculated for this stroke classification test include Recall Sensitivity True Positive Rate (TPR), False Positive Rate (FPR) False Alarm rate, Specificity True Negative Rate (TNR), Precision and False Negative Rate (FNR), and Accuracy. From the classification of brain stroke data using mutual information feature selection and without feature selection. By testing using split data test 60%, cross validation-10, both with cross validation-5, the highest results are using feature selection which is higher than without feature selection. For the highest accuracy with feature selection, the cross-validation scheme -10 is 0.7760 without feature selection, and 0.7790 with mutual information feature selection. With an increase of 0.0030, where there is no significant difference, it can be said that almost all the attributes contained in the brain stroke dataset used all have an influence on their relevance to the stroke disease class. This research has the potential to automatically make health care decisions and patient outcomes in predicting brain stroke by giving the greatest influence to the parameter, namely age. However, these results have limited information because they were only tested on machine learning which is not very popular. This research needs to be continued using other, more sophisticated methods.

Acknowledgments

This work was supported by a research grant from the Dinamika Bangsa Foundation. Thank you for the material and non-material support that has been provided.

References

[1] L. Bai, F. Ciravegna, R. Bond, and M. Mulvenna, "A Low Cost Indoor Positioning System Using Bluetooth Low Energy," *IEEE Access*, vol. 8, pp. 136858–136871, 2020, doi: 10.1109/ACCESS.2020.3012342.

[2] H. Angga Yuwono, S. Kusuma Wijaya, and P. Prajitno, "Feature selection with Lasso for classification of ischemic strokes based on EEG signals," *J. Phys. Conf. Ser.*, vol. 1528, no. 1, 2020, doi: 10.1088/1742-6596/1528/1/012029.

[3] G. Fang, W. Liu, and L. Wang, "A machine learning approach to select features important to stroke prognosis," *Comput. Biol. Chem.*, vol. 88, no. June, p. 107316, 2020, doi: 10.1016/j.compbiolchem.2020.107316.

[4] S. Ray, K. Alshouiliy, A. Roy, A. Alghamdi, and D. P. Agrawal, "Chi-Squared Based Feature Selection for Stroke Prediction using AzureML," *2020 Intermt. Eng. Technol. Comput. IETC 2020*, 2020, doi: 10.1109/IETC47856.2020.9249117.

[5] T. G. Debelee, S. R. Kebede, F. Schwenker, and Z. M. Shewarega, "Deep Learning in Selected Cancers' Image Analysis—A Survey," *J. Imaging*, vol. 6, no. 11, pp. 1–40, 2020, doi: 10.3390/jimaging6110121.

[6] P. Narasimhaiah and C. Nagaraju, "Breast Cancer Screening Tool Using Gabor Filter-Based Ensemble Machine Learning Algorithms," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2, pp. 936–947, 2023.

[7] R. Tugay and Ş. G. Ögüdücü, "Demand prediction using machine learning methods and stacked generalization," *Proc. 6th Int. Conf. Data Sci. Technol. Appl.*, pp. 216–222, 2020, doi: 10.5220/0006431602160222.

[8] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine Learning for Brain Stroke: A Review," *J. Stroke Cerebrovasc. Dis.*, vol. 29, no. 10, 2020, doi: 10.1016/j.jstrokecerebrovasdis.2020.105162.

[9] B. El Boudani *et al.*, "Implementing deep learning techniques in 5g iot networks for 3d indoor positioning: Delta (deep learning-based co-operative architecture)," *Sensors (Switzerland)*, vol. 20, no. 19, pp. 1–20, 2020, doi: 10.3390/s20195495.

[10] L. Buch and A. Andrzejak, "Learning-Based Recursive Aggregation of Abstract Syntax Trees for Code Clone Detection," *SANER 2019 - Proc. 2019 IEEE 26th Int. Conf. Softw. Anal. Evol. Reengineering*, pp. 95–104, 2019, doi: 10.1109/SANER.2019.8668039.

[11] J. Tolan *et al.*, "Sub-meter resolution canopy height maps using self-supervised learning and a vision transformer trained on Aerial and GEDI Lidar," 2023, [Online]. Available: http://arxiv.org/abs/2304.07213.

[12] Y. H. Haw *et al.*, "Classification of basal stem rot using deep learning: a review of digital data collection and palm disease classification methods," *PeerJ Comput. Sci.*, vol. 9, pp. 1–30, 2023, doi: 10.7717/PEERJ-CS.1325.

[13] J. You, W. Liu, and J. Lee, "A DNN-based semantic segmentation for detecting weed and crop," *Comput. Electron. Agric.*, vol. 178, no. September, p. 105750, 2020, doi: 10.1016/j.compag.2020.105750.

[14] Fachruddin, Saparudin, E. Rasywir, Y. Pratama, and B. Irawan, "Extraction of object image features with gradation contour," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 6, pp. 1913–1923, 2021, doi: 10.12928/TELKOMNIKA.v19i6.19491.

[15] H. Saleh, S. Mostafa, A. Alharbi, S. El-Sappagh, and T. Alkhalifah, "Heterogeneous Ensemble Deep Learning

Model for Enhanced Arabic Sentiment Analysis," *Sensors*, vol. 22, no. 10, pp. 1–28, 2022, doi: 10.3390/s22103707.

[16] K. Park, J. S. Hong, and W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," *Appl. Artif. Intell.*, vol. 34, no. 5, pp. 396–411, 2020, doi: 10.1080/08839514.2020.1723868.

[17] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, "Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT," no. Dl, 2022, [Online]. Available: http://arxiv.org/abs/2207.07919.

[18] M. Ju, H. Luo, Z. Wang, B. Hui, and Z. Chang, "The application of improved YOLO V3 in multi-scale target detection," *Appl. Sci.*, vol. 9, no. 18, 2019, doi: 10.3390/app9183775.

[19] G. Liu, J. C. Nouaze, P. L. T. Mbouembe, and J. H. Kim, "YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3," *Sensors (Switzerland)*, vol. 20, no. 7, pp. 1–20, 2020, doi: 10.3390/s20072145.

[20] G. Neelakantam, D. D. Onthoni, and P. K. Sahoo, "Fog computing enabled locality based product demand prediction and decision making using reinforcement learning," *Electron.*, vol. 10, no. 3, pp. 1–16, 2021, doi: 10.3390/electronics10030227.

[21] L. Alzubaidi *et al.*, *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021.

[22] F. Utami, S. Suhendri, and M. Abdul Mujib, "Implementasi Algoritma Haar Cascade pada Aplikasi Pengenalan Wajah," *J. Inf. Technol.*, vol. 3, no. 1, pp. 33–38, 2021, doi: 10.47292/joint.v3i1.45.

[23] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00329-2.

[24] F.-Z. Nakach, A. Idri, and H. Zerouaoui, "Deep Hybrid Bagging Ensembles for Classifying Histopathological Breast Cancer Images," in *Proceedings ofthe 15th International Conference on Agents and Artificial Intelligence (ICAART2023) - Volume 2, pa*, 2023, vol. 2, no. Icaart, pp. 289–300, doi: 10.5220/0011704200003393.

[25] H. Lee *et al.*, "Machine Learning Approach to Identify Stroke Within 4.5 Hours," *Stroke*, vol. 51, no. 3, pp. 860–866, 2020, doi: 10.1161/STROKEAHA.119.027611.

[26] T. Kakkonen and M. Mozgovoy, "HERMETIC AND WEB PLAGIARISM DETECTION SYSTEMS FOR STUDENT ESSAYS — AN EVALUATION OF THE STATE-OF-THE-ART," *J. Educ. Comput. Res.*, vol. 42, no. 2, pp. 135–159, 2010, doi: 10.2190/EC.42.2.a.

[27] F. Ullah, M. R. Naeem, L. Mostarda, and S. A. Shah, "Clone detection in 5G-enabled social IoT system using graph semantics and deep learning model," *Int. J. Mach. Learn. Cybern.*, 2021, doi: 10.1007/s13042-020-01246-9.

[28] S. A. El-Regaily, M. A. M. Salem, M. H. Abdel Aziz, and M. I. Roushdy, "Multi-view Convolutional Neural Network for lung nodule false positive reduction," *Expert Syst. Appl.*, vol. 162, p. 113017, 2020, doi: https://doi.org/10.1016/j.eswa.2019.113017.

[29] H. Matta, R. Gupta, and S. Agarwal, "Search Engine optimization in Digital Marketing: Present Scenario and Future Scope," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 2020, pp. 530–534, doi: 10.1109/ICIEM48762.2020.9160016.

[30] R. A. Wilis and A. Nurwulandari, "The effect of E-Service Quality, E-Trust, Price and Brand Image Towards E-Satisfaction and Its Impact on E-Loyalty of Traveloka's Customer," *JIMEA J. Ilm. MEA (Manajemen, Ekon. Akuntansi)*, vol. 4, no. 3, pp. 1061–1099, 2020.

[31] S. Assegaff, E. Rasywir, and Y. Pratama, "Experimental of vectorizer and classifier for scrapped social media data," vol. 21, no. 4, pp. 815–824, 2023, doi: 10.12928/TELKOMNIKA.v21i4.24180.

[32] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning," *Rev. d'Intelligence Artif.*, vol. 34, no. 6, pp. 753–761, 2020, doi: 10.18280/RIA.340609.

[33] Fachruddin, Saparudin, E. Rasywir, and Y. Pratama, "Network and layer experiment using convolutional neural network for content based image retrieval work," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 20, no. 1, pp. 118–128, 2022, doi: 10.12928/TELKOMNIKA.v20i1.19759.

