Classification of Brain Stroke: Predictive Analysis using Random Forest

Steve Aditya Gandha
Mathematics and Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
steve.gandha@binus.ac.id

Sherly Patricia
Mathematics and Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
sherly.patricia001@binus.ac.id

Dionisius Avelino
Mathematics and Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
dionisius.avelino@binus.ac.id

Manisha Arie Paramartha
Mathematics and Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
manisha.paramartha@binus.ac.id

Yonathan Immanuel Nurhan
Mathematics and Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
yonathan.nurhan@binus.ac.id

***Abstract*—The brain is the center of human body control that plays lots of important roles in every process of the human body. When the most important part of a human body is not functioning well, the body will not be able to operate properly. This research aims to create a random forest model that can predict a person's risk of having a brain stroke. By giving information related to the required variables, people can be more aware of its warning signs and symptoms, which are very helpful in preventing brain stroke. The model that is used to provide a predictive analysis in this research is Random Forest. By using the random forest model, the result shows 81% of accuracy for the 78% of not stroke cases that are predicted. As for the 79% positive stroke cases that are predicted, it is predicted with an accuracy of 82%. Therefore, this research shows that random forest can be used to predict upcoming brain stroke cases for early prevention in the future. However, further studies should be conducted with larger dataset and maybe using other machine learning methods in order to make more accurate predictors.**

Keywords—Brain Stroke, Prediction, Random Forest

I. INTRODUCTION

An early brain stroke prediction is a form of prevention against strokes that is required at this time as it is the second leading cause of death in the world, with a death rate of around 5.5 million people per year [1]. This makes brain stroke a beneficial topic to elevate since the brain also plays an integral part in the human body. The brain is the center of human body control that plays lots of important roles in every process of the human body. Subsequently, it is the organ that sends signals to other parts of the body, allowing the body to be used for day-to-day activities. Besides that, the brain also allows humans to form thoughts, control emotions, and understand languages that are important for daily life. Like other parts of the human body, the brain is also prone to diseases or problems, such as brain stroke. Brain stroke occurs when there is insufficient blood flowing to the brain in turn causing a lack of oxygen as blood carries oxygen which ultimately causes brain cells to die [2]. Furthermore, brain stroke is the most common cause of disability in people. It can be categorized into two different types. One of them occurs because of bleeding on the brain or because of a blood blockage. Strokes caused by blood blockage are also known as ischemic strokes, which account for 80% of brain strokes. These can occur due to several things, such as a clot in the blood vessel of the brain or neck, a clot originating from another part of the body, or a narrowing of arteries. Another stroke caused by bleeding is called hemorrhagic stroke [3].

This research aims to create a random forest model that can predict a person's risk of having a brain stroke. By giving information related to the required variables, people can be more aware of its warning signs and symptoms, which are very helpful in preventing brain stroke. According to the previous facts and backgrounds, preventing brain stroke is very important. These not only improve the quality of life of people who are susceptible to brain stroke but also those around them because caring for someone who has had a stroke requires a lot of comprehensive treatment.

Advances in technology allow researchers to utilize modern methods, such as a machine learning model, to perform analysis on topics like these. Before proceeding with the machine learning model, data cleaning will be done first by handling duplicate values, outliers, and null values. Then, a machine learning model known as the Random Forest will be utilized. The Random Forest is a supervised machine-learning model that is popularly used for making predictions. Furthermore, the dataset used for this research was taken from Kaggle and it works perfectly with Random Forest. The effectiveness of the models will be deduced based on the accuracy it produces [4]. In this way, researchers will know whether this model can be used to predict brain stroke or vice versa. There are several research done on this particular topic using Machine Learning models. However, most of them utilize a group of methods at once and perform a comparison, while this research will focus more on one method, Random Forest [5].

II. RELATED WORKS

Various research have been done related to this particular topic by using the machine learning model approach. First, Sirsat, M.S. et al. used four categories of machine learning models to do their prediction which was done in 2020 using a database of brain strokes from 2007-2019. They found out that out of the 4 models they used, the Support Vector Machine (SVM) gave the best accuracy [6]. Second, Bandi, V. et al. used seven different machine learning models in 2021 to perform a brain stroke prediction. They also came to a pretty similar conclusion to the one mentioned prior [5]. Third, Krishna, Vempati, et al. also did a similar research in 2021 where they took on the topic of early detection by using

machine learning models, such as K-nearest neighbor and logistic regression. However, what differentiates theirs from the one prior is that they did not compare the results of the different models [7].

Various machine learning classification models, such as AdaBoost, Gaussian, and many others to perform the prediction, were done by Emon, M.U., et al. They compared things like false positive and negative rates of the different classifiers and they concluded that the weighted voting classifier performed better with an accuracy of 97% [8]. Mahesh, K. A. et al. also did a machine learning models approach on this topic to the conclusion that models, such as decision trees and neural networks are proper for this topic since it has an acceptable accuracy [9]. Furthermore, Dristas, E. et al. in 2022 concluded that their experiments showed that stacking classification outperformed other methods with an accuracy of 98% [10].

Subsequently, Rahman, S. et al. in 2023 where they did a comparison between machine learning and neural network methods. They concluded that the best of the machine learning models produced a 99% accuracy which is higher than the best from the neural network side which is only 92% [11]. In 2022, Mariano, V. et al. combined the use of machine learning classification methods with linearized scattering operators and produced promising results. Overall, all of the research done on this topic using machine learning methods has been a bunch and it can be observed that it is a very effective approach for this type of problem [12].

III. Methodology

This research begins with dataset selection that will be used for the model to develop predictive analysis. The methods that are used for this research consist of data selection, data analytics, data preprocessing, data splitting, data training, data testing, evaluation, and result that is shown in Fig. 1.

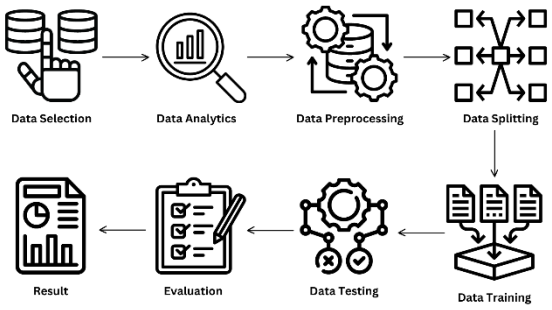

Fig. 1. The workflow of research methods.

A. Datasets

Dataset plays an important role in training the model. For this model, the dataset used is taken from Kaggle. This dataset contains information from around the globe. It contains 4981 different observations with 11 different variables for each observation. However, there are only seven independent variables that are used, such as gender, age, hypertension, heart disease, average glucose level, BMI, and smoking status. The target variable or dependent variable from this dataset is the stroke variable which tells whether the person has a stroke or not. The dataset will then be split into 80% for training and 20% for testing, with 0 being chosen as the random state for this research. Fig. 2, Fig. 3, Fig. 4 and Fig. 5 show the information and examples of dataset that are provided.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
2	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
3	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
4	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1

Fig. 2. Dataset variables and the first five samples.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
4976	Male	41.0	0	0	No	Private	Rural	70.15	29.8	formerly smoked	0
4977	Male	40.0	0	0	Yes	Private	Urban	191.15	31.1	smokes	0
4978	Female	45.0	1	0	Yes	Govt_job	Rural	95.02	31.8	smokes	0
4979	Male	40.0	0	0	Yes	Private	Rural	83.94	30.0	smokes	0
4980	Female	80.0	1	0	Yes	Private	Urban	83.75	29.1	never smoked	0

Fig. 3. Dataset variables and the last five samples.

```
RangeIndex: 4981 entries, 0 to 4980
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             4981 non-null   object
 1   age                4981 non-null   float64
 2   hypertension       4981 non-null   int64
 3   heart_disease      4981 non-null   int64
 4   ever_married       4981 non-null   object
 5   work_type          4981 non-null   object
 6   Residence_type     4981 non-null   object
 7   avg_glucose_level  4981 non-null   float64
 8   bmi                4981 non-null   float64
 9   smoking_status     4981 non-null   object
 10  stroke             4981 non-null   int64
dtypes: float64(3), int64(3), object(5)
```

Fig. 4. Data information from the dataset.

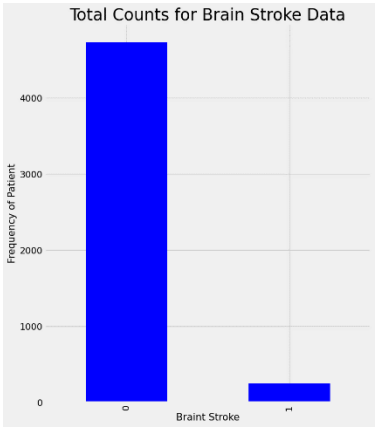

Fig. 5. The total samples of Brain Stroke.

B. Feature Engineering

Feature engineering is the process of creating new features or transforming existing features from the dataset to obtain features that are relevant and suitable for the machine learning models. The better the quality of the features used to train the models, the better the machine learning models will be [13].

a. Binning

Data binning, also known as data bucketing, is a data pre-processing technique that is often used in machine learning to convert or group a continuous value into an interval or bins. It can help a lot with the model since it smoothen data fluctuations and also reduces the amount of errors that are in the data. For this dataset specifically binning will be done on the age variable where they will be grouped into a new variable called ‘age band’ [14] .

TABLE I
THE RESULT OF DATA BINNING

age	age_band
0-16	0
17-32	1
33-48	2
49-64	3
>64	4

b. Normalization

Normalization is a data pre-processing technique that is used to transform variables in a dataset into a specific scale or range, commonly it is converted into [0, 1] [15]. It is very useful in helping a machine learning model's accuracy. For this dataset, the two variables that are going to be normalized are "average_glucose_level" and "BMI".

$$Normalized[column] = \frac{data[column]}{max(abs(data[column]))} \quad (1)$$

c. Encoding

A feature that is used to transform categorical variables in the dataset into numerical variables. This is because most of the machine learning models can only work with numerical variables. In other words, it cannot handle categorical variables. It is why transforming the variables into a numerical one is necessary and that will lead to a better machine learning model [16]. For this dataset, the two categorical variables that are going to be transformed into numerical variables are 'gender' with one-hot encoding and 'smoking_status' with label encoding.

TABLE II
THE RESULT OF DATA ENCODING

variable	**old**	**new**
gender	Male	0
	Female	1
smoking_status	formerly smoked	0
	never smoked	1
	smokes	2
	Unknown	3

d. Random Sampling

The simplest form of sampling where data points are selected randomly from a dataset without any predetermined pattern. This method also assumes that every point in a dataset has an equal chance of being selected [17]. For this model, random sampling will be done to reduce the model's bias when predicting, to be specific it will be done based on the 'stroke' variable. Since the dataset has an unbalanced amount of stroke and nonstroke observations, it will be balanced by randomly sampling the nonstroke observations.

C. Model

A model is important in this research in which machine learning is being applied. By creating a good model, a good prediction related to brain stroke can be made. The model that id used to provide a predictive analysis in this research is Random Forest. Random forest is a very widely used machine learning algorithm that combines the output of multiple decision trees and comes up with a single result. It can handle both classification and regression problems [18]. For this dataset, the random forest model will be used to handle classification (handling categorical variables).

D. Metrics

Evaluation Metrics are measures that are used to assess the performance of a model. It can help to know how well and effectively a model is used to make predictions. By using evaluation metrics, we can compare and choose which models are fit for a given task [19]. For this dataset which involves classification, the evaluation metrics used are accuracy, precision, recall, and F1 score.

a. Class

Class is the separation of the different variables in the dataset into specific groups or categories. For this dataset the variable that will be split into different classes is the target variable 'stroke' where a person with stroke will be class 1 and 0 for nonstroke.

b. Accuracy

Accuracy is the ratio of correctly predicted true positive and true negative to the total instances. It is used to measure how well the model performed.

$$\text{Accuracy} = \frac{(True\ Positive + True\ Negative)}{Total\ Number\ Predictions} \quad (2)$$

c. Precision

Precision measures the accuracy of the positive predictions made by the model. It is the ratio of correctly predicted positives to the total of the predicted positives. For this dataset, high precision means that when the model predicts a stroke, it is likely to be correct.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

d. Recall

Recall is the proportion of the True predictions that were identified (True Positive / (True positive + False Negative)), it helps us understand how complete the results are. It's useful in a medical case where a false alarm doesn't matter as much.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

e. F1 Score

F1 Score is the harmonic mean of precision and recall. It is often used in cases where the False Positive

and False Negative are equally costly or the True Negative is high.

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2x\frac{precision\ x\ recall}{precision + recall} = \frac{True\ Positive}{True\ Positive + \frac{False\ Negative\ +\ False\ Positive}{2}} \quad (5)$$

IV. Result and Discussions

In this research, we performed this model by coding on Google Colab. This research chose predictive analysis as it focused by using one of the machine learning methods in purpose to make a brain stroke prediction model that can be developed and used in the future as brain stroke prevention.

TABLE III
CLASSIFICATION REPORT

Random Forest		Class 0				Class 1			
Accuracy	Total support	Precision	Recall	F1	Support	Precision	Recall	F1	Support
0.8	100	0.81	0.78	0.8	50	0.79	0.82	0.8	50

From Table III, it can be observed that in the case of:

a. Not Stroke: 81% percent of the time it is predicted correctly and 78% of the "not stroke" cases are predicted, though recall is not that important it is still an interesting variable to look at. The F1 score or commonly known as the harmonic mean of precision and recall is also very high.
b. Stroke: 79% of the time the stroke prediction is correct and the model manages to catch 82% of all the positive cases which for this class is very important since nobody wants the model to tell that someone is fine when they are not. It also has the same F1 score as the other class.

Both classes are supported by 50 data each. Overall, the model achieved an accuracy of 80% which is very good. Therefore, based on the results provided by the report, the random forest model can really help with predicting possible brain stroke when provided by certain variables. Few things to highlight are most of the time it will not make false stroke calls and will catch the possible stroke attack.

This research focused on using Random Forest rather than using a group of machine learning models to make brain stroke predictions. It is because we would like to determine how effective and accurate a random forest model is in predicting the brain stroke by itself, without any combinations of other machine learning methods. An accuracy of 80% is high enough for a single model to achieve when it is compared to the accuracy of existing research that used a group of machine learning models.

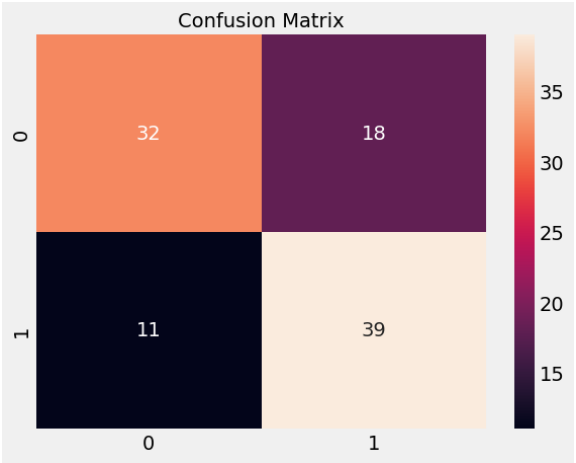

Fig.6. Confusion Matrix of Random Forest.

Figure 6 shows the confusion matrix of the model, which is the Random Forest. The random forest model achieved an accuracy of 80%.

In this research, researchers can evaluate the performance of a random forest is to predict brain stroke by seeing the evaluation performance metrics. Researcher can also explore deeper related to this random forest model. This research can also provide further insights and conclusions that can be developed and used in the future.

V. Conclusion

Brain plays an integral part in the human body. It is the center of human body control that plays lots of important roles in every process of the human body. the brain is not functioning well, the body will not be able to operate properly. One of the machine learning models was used in this research in purpose to determine the performance of Random Forest in predicting brain stroke. A Random Forest model was used where the model went through data pre-processing, training, splitting, and testing which resulted in an accuracy of 80%, which is considered good enough for a single model to perform as brain stroke predictor. This means that in the future the model can be used to predict upcoming brain stroke or at least predict people with high risk, so that they can be more aware and do some early prevention.

Through this research, we can observe that lots of existing research that examined a combination of machine learning models provide higher accuracy compared to a single model of Random Forest. Therefore, further studies should be conducted with larger datasets and maybe using other machine learning methods. Consequently, it can hopefully be a more accurate predictor and make it even more useful for society.

References

[1] E. S. Donkor, "Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life," Stroke Research and Treatment, vol. 2018, pp. 1–10, Nov. 2018.
[2] Centers for Disease Control and Prevention, "About Stroke," https://www.cdc.gov/stroke/about/?CDC_AAref_Val=https://www.cdc.gov/stroke/about.htm. [Accessed: May 19, 2024].
[3] National Institute of Neurological Disorders and Stroke, "Brain Basics: Preventing Stroke," https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-preventing-stroke#:~:text=It%20is%20the%20most%20common,that%20burden%20through%20biomedical%20research. [Accessed: May 19, 2024].
[4] Javatpoint, "Machine Learning - Random Forest Algorithm," https://www.javatpoint.com/machine-learning-random-forest-algorithm. [Accessed: May 19, 2024].
[5] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of Brain Stroke Severity Using Machine Learning," Revue d'Intelligence Artificielle, vol. 34, no. 6, pp. 753-761, Dec. 2020
[6] M. S. Sirsat, E. Ferm´e, and J. Cˆamara, "Machine Learning for Brain Stroke: A Review," Journal of Stroke

and Cerebrovascular Diseases, vol. 29, no. 10, p. 105162, Oct. 2020.
[7] V. Krishna, J. S. Kiran, P. P. Rao, G. C. Babu, and G. J. Babu, “Early Detection of Brain Stroke Using Machine Learning Techniques.” Trichy, India: IEEE, 2021, pp. 1489–1495.
[8] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, Nov. 2020.
[9] K. Akash, H. Shashank, S. .S, and T. A.M, “Prediction of Stroke Using Machine Learning,” 06 2020.
[10] E. Dritsas and M. Trigka, “Stroke Risk Prediction with Machine Learning Techniques,” Sensors, vol. 22, no. 13, p. 4670, Jun. 2022.
[11] V. Mariano, J. A. T. Vasquez, M. R. Casu, and F. Vipiana, “Efficient Data Generation for Stroke Classification Via Multilayer Perceptron.” Denver, CO, USA: IEEE, 2022, pp. 890–891.
[12] S. Rahman, M. Hasan, and A. K. Sarkar, “Prediction of Brain Stroke Using Machine Learning Algorithms and Deep Neural Network Techniques,” European Journal of Electrical Engineering and Computer Science, vol. 7, no. 1, pp. 23–30, Jan. 2023.
[13] GeeksForGeeks, "What is Feature Engineering" https://www.geeksforgeeks.org/what-is-feature-engineering. [Accessed: June 10, 2024].
[14] M, Kabungo, "Data Bining Explained" https://medium.com/@mose.kabungo/binning-explained-557aa3cce591. [Accessed: June 10, 2024].
[15] Deepchecks, "Normalization in Machine Learning" https://deepchecks.com/glossary/normalization-in-machine-learning/#:~:text=Normalization%20in%20machine%20learning%20is%20the%20process%20of%20translating%20data,when%20Euclidean%20distance%20is%20used [Accessed: June 10, 2024].
[16] GeeksForGeeks, "Feature Encoding Techniques Machine Learning" https://www.geeksforgeeks.org/feature-encoding-techniques-machine-learning/ [Accessed: June 10, 2024].
[17] R. Gangwal, "Types of Sampling and Sampling Technique" https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/. [Accessed: June 10, 2024].
[18] IBM, "What is Random Forest" https://www.ibm.com/topics/random-forest [Accessed: June 10, 2024].
[19] S. Kumar Agwal, "Metrics to Evaluate Your Classification Model to Take the Right Decisions" https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/[Accessed: June 10, 2024].