Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 7, No. 4, November, Pp. 359-370

Electronic medical record data analysis and prediction of stroke disease using explainable artificial intelligence

Yuri Pamungkas*[1], Adhi Dharma Wibawa[2], Meiliana Dwi Cahya[3]
Medical Technology, Department of Biomedical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[1']
Medical Technology and Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[2]
Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Negeri Malang, Indonesia[3]

Article Info
Keywords:
Electronic Medical Record, Statistical Analysis, Stroke Prediction, XAI Algorithm, Probabilistic Neural Network

Article history:
Received: September 04, 2022
Accepted: October 24, 2022
Published: November 30, 2022

Cite:
Y. Pamungkas, A. D. Wibawa, and M. D. Cahya, "Electronic Medical Record Data Analysis and Prediction of Stroke Disease Using Explainable Artificial Intelligence (XAI)", *KINETIK*, vol. 7, no. 4, Nov. 2022. https://doi.org/10.22219/kinetik.v7i4.1535

*Corresponding author.
Yuri Pamungkas
E-mail address:
yuri@its.ac.id

Abstract
The deficiency of oxygen in the brain will cause the cells to die, and the body parts controlled by the brain cells will become dysfunctional. Damage or rupture of blood vessels in the brain is better known as a stroke. Many factors affect stroke. These factors certainly need to be observed and alerted to prevent the high number of stroke sufferers. Therefore, this study aims to analyze the variables that influence stroke in medical records using statistical analysis (correlation) and stroke prediction using the XAI algorithm. Factors analyzed included gender, age, hypertension, heart disease, marital status, residence type, occupation, glucose level, BMI, and smoking. Based on the study results, we found that women have a higher risk of stroke than men, and even people who do not have hypertension and heart disease (hypertension and heart disease are not detected early) still have a high risk of stroke. Married people also have a higher risk of stroke than unmarried people. In addition, bad habits such as smoking, working with very intense thoughts and activities, and the type of living environment that is not conducive can also trigger a stroke. Increasing age, BMI, and glucose levels certainly affect a person's stroke risk. We have also succeeded in predicting stroke using the EMR data with high accuracy, sensitivity, and precision. Based on the performance matrix, PNN has the highest accuracy, sensitivity, and F-measure levels of 95%, 100%, and 97% compared to other algorithms, such as RF, NB, SVM, and KNN.

1. Introduction

In recent decades, stroke has become the second leading cause of death and the third disability worldwide. Every year, about 17.7 million new cases of stroke and about 6.7 million deaths occur due to stroke [1]. According to the World Health Organization (WHO), stroke is a condition or clinical sign of focal and global neurologic deficits that can be burdensome and last a long time [2]. Blockage or damage to blood vessels in the brain is the cause of stroke. It can result in part of the brain not getting a blood supply and can lead to death without any other apparent cause other than vascular. In addition, stroke is also a contributing factor to dementia and depression.

As part of the cardiocerebrovascular disease (classified as a catastrophic disease), stroke has a broad economic and social impact. This disease can cause permanent disability, so it will automatically affect the productivity of the sufferer and increase the burden of health financing. According to data from BPJS (Social Security Administering Agency in Indonesia) Health, there was an increase in total financing for catastrophic disease services in the JKN (National Health Insurance in Indonesia) in 2016-2018 by 4 trillion rupiahs. Stroke cost health care costs 2.56 trillion rupiahs in 2018, making it one of Indonesia's diseases with the highest medical costs [3]. Therefore, stroke management requires serious attention because it significantly impacts the country's socio-economic development.

The adverse effects of stroke can be minimized if the stroke is recognized more quickly and get immediate help. However, many obstacles are faced in treating stroke symptoms quickly and accurately. One of the factors constraining stroke management in Indonesia is the not yet optimal screening process or early detection of stroke. It can be seen from the increasing prevalence of risk factors and the low achievement of health screening for productive age. The delay in handling cases is because people do not recognize the early signs of a stroke. In addition, not all health services have a complete and integrated diagnostic system or stroke management team.

The use of AI (Artificial Intelligence) to build an accurate disease diagnostic system (especially stroke) has become the most feasible and efficient option. Many disease screening systems have utilized AI in the decision-making process for their predictions. Choi et al. [4] performed stroke prediction and classification using the Convolutional Neural Network - Bidirectional LSTM method. Obtained an accuracy rate of 94% with 6% False Positive Result and 5.7% False Negative Result for the process of prediction and classification of stroke data. Fang et al. [5] also used the Deep Learning method to predict ischemic stroke. Deep learning includes Residual networks, Convolutional Neural networks,

and Long-Short Term Memory. The results obtained stroke prediction accuracy rates of 82% (for Residual Network), 83% (for CNN), and 82% (for LSTM). In addition, in a study, Kaur et al. [6] also used Deep Learning to predict the signs of stroke occurrence. Accuracy levels of 95.6% were obtained (when using the Gate Recurrent Unit algorithm), 91% (when using the Bidirectional LSTM), 87% (when using the Bidirectional LSTM), and 83% (when using the FFNN algorithm) in the early prediction of stroke symptoms.

It cannot be denied that the Deep Learning used in previous studies has a high level of accuracy. However, this high level of accuracy is not accompanied by the transparency of data processing in Deep Learning [7]. Data processing that occurs in deep learning tends to be difficult for users to understand because of the complex architecture of its constituents [8]. In other words, the decision-making process in deep learning tends to be hidden in a black box that is difficult to solve [9]. Moreover, in the process of predicting diseases such as stroke, it will be very risky if the decision-making process in a method/algorithm tends to be hidden in a black box. Therefore, the use of Explainable AI is expected to overcome some of the weaknesses of Deep learning, such as the transparency of the data processing process or decision making [10].

Explainable AI (XAI) utilizes mathematical or statistical calculations simpler than Deep learning, such as regressions, probability values, the distance of the nearest neighbours, voting on tree distribution results, and others [11]. Simple mathematical or statistical calculations can undoubtedly make it easier for users to understand how existing methods work and if there are errors in the process, it will be easier to make improvements or evaluations [12]. Some examples of methods belonging to XAI include Probabilistic Neural Network, Naive Bayes, K-Nearest Neighbor, Random Forest, and Support Vector Machine. The above method is based on mathematical calculations much simpler than deep learning methods such as R-CNN, Bi-LSTM, Resnet, Etc.

Based on some of the considerations above related to AI-based stroke prediction, this study utilizes several methods belonging to XAI for the early prediction process of stroke using EMR data. Researchers also tried to analyze the factors that influence stroke using statistical analysis or correlation values between parameters in the Electronic Medical Record. EMR data for the prediction and analysis process is crucial because it contains much patient personal information and medical history that can be used as a reference to determine what factors influence the onset of a stroke. In addition, EMR data in hospitals is still rare to analyze in more detail because of the large number and requires a very long time when analyzed manually. Therefore, we hypothesize that using XAI to process EMR data for stroke prediction is one of the right steps. The use of XAI in this study is also expected to provide precise and accurate predictive results for people at risk of stroke.

2. Research Method

This section will explain the research that has been carried out step by step. The stages include EMR data collection, data exploration, data pre-processing, statistical analysis, stroke prediction using XAI, and analysis of results. Each of these stages will also be explained in each subsequent sub-section.

2.1 EMR Data Collection

This study used Hospital Electronic Medical Record (EMR) data from 3412 patients, including 201 stroke patients and 3211 non-stroke patients. EMR data for each patient (both stroke and non-stroke) includes gender, age, history of hypertension, heart disease, marital status, type of occupation, area of residence, glucose level, BMI, and smoking history. Table 1 is a breakdown of the EMR data used in this study.

Table 1. Electronic Medical Record of Stroke and Non-Stroke Patients

No.	Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence Type	Glucose Level	BMI	Smoking Status	Condition
1	Male	79	No	No	Yes	Private	Rural	72,73	28,40	No	Stroke
2	Female	63	No	No	Yes	Govt	Rural	205,35	42,20	Yes	Stroke
3	Female	81	No	No	No	Govt	Urban	70,30	25,80	Yes	Stroke
4	Male	49	No	No	No	Private	Rural	104,86	31,90	Yes	Stroke
5	Female	81	No	No	Yes	Private	Rural	184,40	27,50	No	Stroke
…	…	…	…	…	…	…	…	…	…	…	…
3408	Male	67	No	No	Yes	Private	Urban	190,70	36,00	Yes	Non-Stroke
3409	Female	45	No	No	Yes	Govt	Urban	113,63	27,50	Yes	Non-Stroke
3410	Female	66	No	No	Yes	Private	Rural	141,24	28,50	No	Non-Stroke
3411	Male	58	Yes	No	Yes	Govt	Rural	56,96	26,80	Yes	Non-Stroke
3412	Male	69	No	No	Yes	Private	Rural	203,04	33,60	No	Non-Stroke

2.2 Data Exploration

Stroke is a disease that arises because of damage to brain blood vessels, so the blood supply that carries oxygen to the brain is disrupted [13]. Depriving the brain of oxygen will cause its cells to die, and the body parts controlled by these brain cells will become dysfunctional [14]. Factors that cause stroke can be divided into two types: modifiable factors (gender, age, heart disease, genetics, Etc) and non-modifiable factors (hypertension, smoking, glucose, BMI, physical activity, Etc) [15]. Figure 1 presents the distribution of stroke and non-stroke patient data based on gender, hypertension, heart disease, marital status, smoking, work type, and residence type.

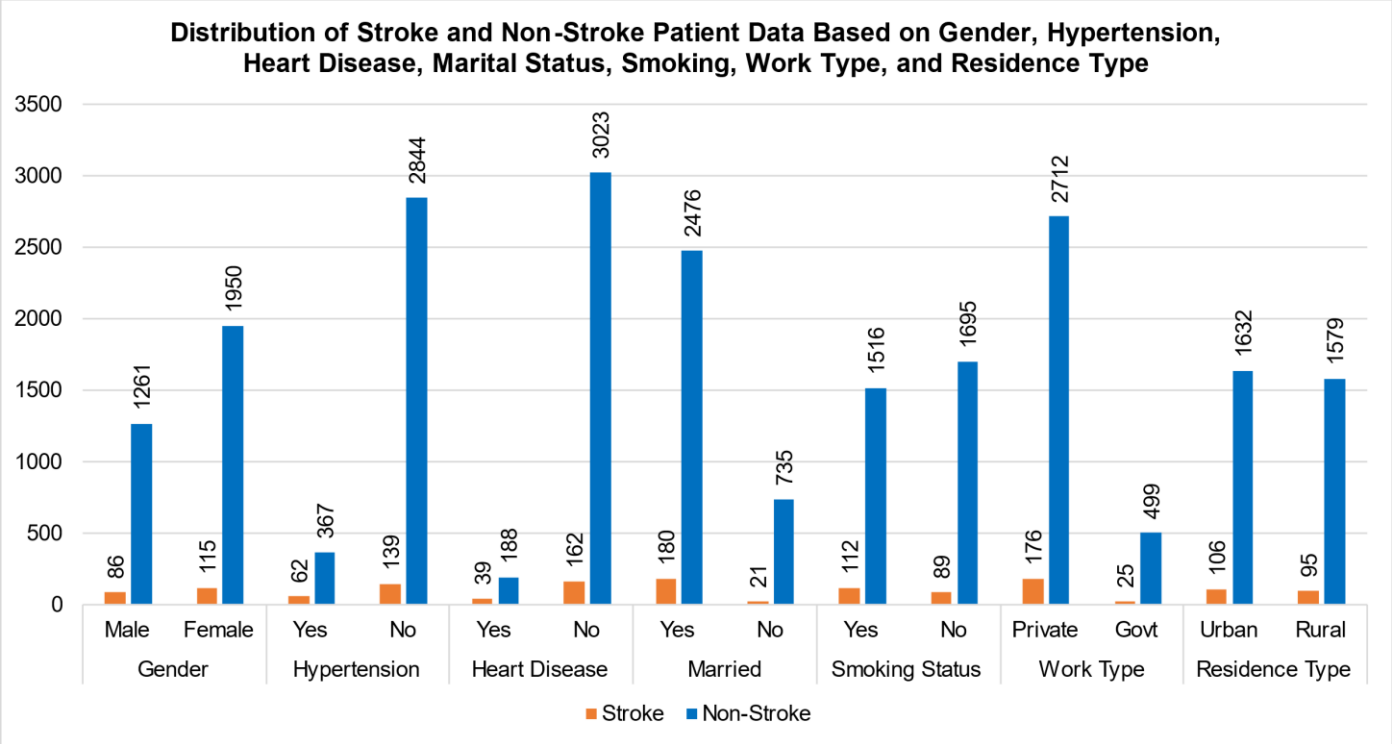

Figure 1. Distribution of Stroke and Non-Stroke Patient Data Based on Gender, Hypertension, Heart Disease, Marital Status, Smoking, Work Type, and Residence Type

If seen at a glance from the distribution of EMR data, the number of female stroke patients (115) was higher than that of male stroke patients (86). So, it can be assumed that women have a higher risk of stroke than men. However, if we look at the number of stroke patients who have a history of hypertension (62) and heart disease (39) with those without hypertension (139) and heart disease (162), it is seen that people who do not appear to have hypertension and heart disease are not necessarily free from the threat of stroke. It may be because hypertension and heart disease has not been detected before, and a stroke occurs suddenly in someone without realising it. So, it can be assumed that people who do not have hypertension and heart disease (hypertension and heart disease have not been detected early) still have a high risk of stroke.

Furthermore, married people also have a higher risk of stroke than unmarried people. It can be seen from the EMR data that there are 180 married stroke patients and 21 unmarried stroke patients. Marriages expected to make a person happier are sometimes also interspersed with conflicts between individuals or couples. Then, a person's feelings of stress will begin to appear and cause a stroke. Smoking also increases a person's risk of stroke (112 stroke patients are smokers, and 89 are nonsmokers). In addition, people who work harder also have a higher risk of stroke than people who work moderately or casually. Private workers sometimes work harder than government employees. Based on EMR data, 176 stroke patients were private workers, and 25 stroke patients were government employees. The type of residence also affects the distribution of stroke patients. A total of 106 stroke patients live in urban areas, and 95 stroke patients live in rural areas. It is probably due to the pattern or lifestyle of urban communities that are not healthy. Figure 2 presents the distribution of stroke and non-stroke patient data based on age, BMI, and glucose level.

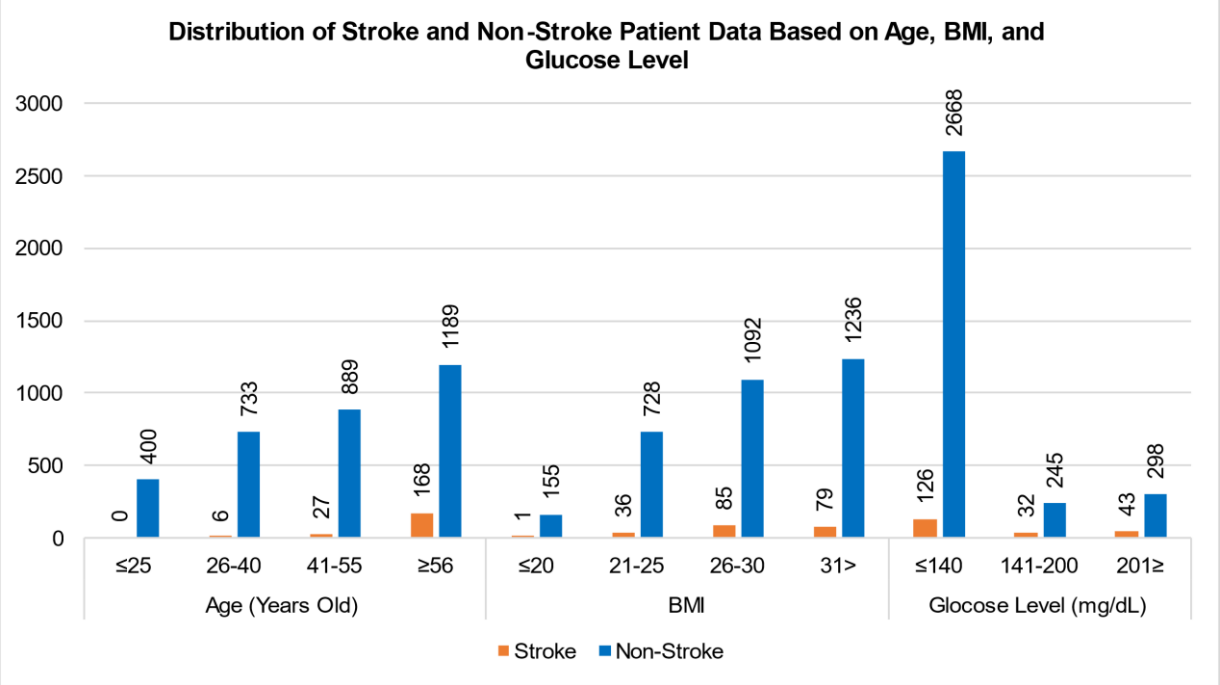

Figure 2. Distribution of Stroke and Non-Stroke Patient Data Based on Age, BMI, and Glucose Level

If seen at a glance from the distribution of EMR data, the older a person ages, the higher the BMI value will increase a person's risk of having a stroke. At the age above 56 years, stroke patients were the highest (168 patients). Likewise, when the BMI value is higher, the number of stroke patients will also increase. Nevertheless, people with low glucose levels are still at high risk of stroke related to blood glucose levels.

2.3 Data Pre-Processing

At the pre-data processing stage, the EMR data of stroke and non-stroke patients was converted into quantitative form (from the previous qualitative form). There is no need for age, BMI, and glucose level data to be converted to numeric values because they were already in numeric form. Meanwhile, the data converted to numeric were data on gender ("1" for male and "0" for female), hypertension ("1" for "yes" statement and "0" for "no" statement), heart disease ("1" for "yes" statements and "0" for "no" statements), marital status ("1" for "yes" statements and "0" for "no" statements), smoking ("1" for "yes" statements " and "0" for the "no" statement), the work type ("1" for the "private" work type and "0" for the "government" work type), and the type of residence ("1" for the "urban" and "0" for "rural" statements). In addition, for "stroke" conditions it is labeled "1" and for non-stroke conditions it is labeled "0". Table 2 below shows the results of converting EMR data to numeric form.

Table 2. Electronic Medical Record of Stroke and Non-Stroke Patients in Numeric Form

No.	Gender	Age	Hypertension	Heart Disease	Marital Status	Work Type	Residence Type	Glucose Level	BMI	Smoking Status	Condition
1	1	79	0	0	1	1	0	72,73	28,40	0	1
2	0	63	0	0	1	0	0	205,35	42,20	1	1
3	0	81	0	0	0	0	1	70,30	25,80	1	1
4	1	49	0	0	0	1	0	104,86	31,90	1	1
5	0	81	0	0	1	1	0	184,40	27,50	0	1
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
3408	1	67	0	0	1	1	1	190,70	36,00	1	0
3409	0	45	0	0	1	0	1	113,63	27,50	1	0
3410	0	66	0	0	1	1	0	141,24	28,50	0	0
3411	1	58	1	0	1	0	0	56,96	26,80	1	0
3412	1	69	0	0	1	1	0	203,04	33,60	0	0

2.4 Statistical Analysis

At this stage, statistical analysis is carried out to determine the factors influencing stroke. By calculating the mean, median, and standard deviation of several parameters (such as age, BMI, and glucose level), it is possible to know the onset values (age, BMI, and glucose level) of stroke and non-stroke patients. In other words, if the onset value is reached, a person is more at risk of having a stroke than when the onset value has not been reached. In addition, stroke parameters were also correlated (age, gender, hypertension, heart disease, marital status, work type, type of residence, glucose level, BMI, and smoking) with the patient's actual condition (stroke or non-stroke) using the Pearson correlation. It is to test how strong each stroke parameter's influence was on the patient's actual condition. The Pearson correlation formula can be described as follows Equarion 1 [16].

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad or \quad r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum {x_i}^2 - (\sum x_i)^2}\sqrt{n\sum {y_i}^2 - (\sum y_i)^2}} \tag{1}$$

Where r = Pearson correlation coefficient; x_i = the value of the *x* variable in the sample data; $\bar{x}$ = mean value of variable *x*; y_i = the value of the *y* variable in the sample data; $\bar{y}$ = mean value of variable *y*; and n = amount of data.

2.5 Stroke Prediction using Explainable Artificial Intelligence (XAI)

As previously explained, XAI is an AI model with more transparent processes than deep learning [7]. The decision-making process in XAI is also based on mathematical calculations that are easier for users to understand [11]. So that if an error occurs in the process, users can evaluate or improve the process quickly and accurately [12]. In this study, there are several types of XAI methods used for the stroke prediction process, including PNN (Probabilistic Neural Network), K-NN (K-Nearest Neighbor), SVM (Support Vector Machine), NB (Naive Bayes), and RF (Random Forest).

2.5.1 Probabilistic Neural Network (PNN)

A Probabilistic Neural Network (PNN) is a kind of NN that does not require extensive forward or backward feedback calculations [17]. PNN is commonly used to perform the classification process or pattern recognition with various types of training data [18]. The advantage of PNN is the use of probability in its decision-making to minimize misclassification when applied to the data classification process [19]. In addition, in the PNN process, there is a PDF function calculation of each data class using non-parametric functions and the Parzen window technique [20]. The stages in PNN begin with new data input X_{new}. Then calculate the Gaussian Kernel of each input vector using Equation 2, Equation 3, and Equation 4.

$$\omega_{i,j} = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\left(X_{new} - X_{i,j}\right)^T (X_{new} - X_{i,j})}{2\sigma^2}\right) \tag{2}$$

Then, the conditional class probability is calculated using the Kernel classmate.

$$P_{NC} = \frac{1}{|C_{NC}|}\sum_{j=1}^{|C_{NC}|} \omega_{NCj} \text{ and } \left\{\omega_{NC,1}, \omega_{NC,2}, \omega_{NC,3}, \dots, \omega_{NC,|C_{NC}|}\right\} \tag{3}$$

Class selection is based on the high probability of conditional class:

$$argmax\{P_i\} \text{ and } 1 \le i \le NC \tag{4}$$

Where $\omega_{i,j}$ = gaussian kernel calculation; P_{NC} = class-conditional probability; and $argmax\{P_i\}$ = selection of the high class-conditional probability.

2.5.2 K-Nearest Neighbour (K-NN)

K-NN is a classification method based on similarity or proximity between data [21]. The proximity is based on the closest K-data value and is used as a reference in determining the new data class [22]. The technique for finding the nearest k generally uses the Euclidean distance formula [23]. Here is the Euclidean distance formula Equation 5 in the KNN algorithm.

$$dis = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + \cdots} \tag{5}$$

Where dis = Euclidean Distance; x_i = the value of the *x* variable in the sample data; y_i = the value of the *y* variable in the sample data; and n = amount of data.

In addition to the Euclidean distance, the distance between data can be calculated based on the Hamming Distance (finding the distance between data based on binary code), Manhattan Distance (finding the distance between data based on p and q vectors in n-dimensional space), and Minkowski Distance (finding the distance between data based on the hybridization of Euclidean Distance and Manhattan Distance).

2.5.3 Support Vector Machine (SVM)

SVM is a classification method that works by finding the best hyperplane or separator function to separate data classes [24]. Hyperplanes in SVM are functions that act as constraints to classify data points [25]. Therefore, data points on both sides of the hyperplane can be interpreted as different classes. In addition to hyperplanes, support vectors are used to maximize classifier margins. In other words, the support vector is the data point closest to the hyperplane and affects the position and orientation of the hyperplane.

Initially Equation 6, the available data is denoted as $\overrightarrow{X_i} \epsilon \Re^d$ while each label is denoted $y_i \in \{-1, +1\}$ for $i = 1,2, \dots, l$, where l is the numbers of data. It is assumed that both classes -1 and +1 can be entirely separated by the defined *d*-dimensional hyperplane.

$$\vec{w}\vec{x} + b = 0 \tag{6}$$

Pattern $\overrightarrow{x_i}$ which belongs to class -1 (negative sample), can be formulated Equaiton 7 as a pattern that satisfies the inequality.

$$\vec{w}\vec{x} + b \leq -1 \tag{7}$$

While the pattern $\overrightarrow{x_i}$ which includes class +1 (positive sample) of Equation 8.

$$\vec{w}\vec{x} + b \geq +1 \tag{8}$$

The most significant margin can be found by maximizing the value of the distance between the hyperplane and its closest point, which is $1/\|\vec{w}\|$. It can be formulated as a Quadratic Programming (QP) problem, which is to find the minimum point of Equation 9, by considering the constraints Equation 10.

$$\min_{\vec{w}} \tau(w) = \frac{1}{2} \|\vec{w}\|^2 \tag{9}$$

$$y_i(\overrightarrow{x_i}.\vec{w} + b) - 1 \geq 0, \forall i \tag{10}$$

This problem can be solved by various computational techniques, including the Langrange Multiplier.

$$L(\vec{w}, b, \alpha) = \frac{1}{2}|\vec{w}|^2 - \sum_{i=l}^{l} \alpha_i \left(y_i\left((\overrightarrow{x_i}.\vec{w} + b) - 1\right)\right) for\ (i = 1,2, \dots, l) \tag{11}$$

α_i is Lagrange Multipliers, which are zero or positive $(\alpha_i \geq 0)$. The optimal value of Equation 11 can be calculated by minimizing L concerning $\vec{w}$ and b, and maximizing L concerning α_i. By considering the characteristics of the optimal point gradient $L = 0$, equation (11) can be modified as a maximization problem that only contains i, as Equation 12 below.

Maximize:

$$\sum_{i=l}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=l}^{l} \alpha_i\, \alpha_j\, y_i\, y_j \overrightarrow{x_i}\, \overrightarrow{x_j} \tag{12}$$

Subject to Equation 13:

$$\alpha_i \geq 0\ (i = 1,2, \dots, l) \sum_{i=l}^{l} \alpha_i y_i = 0 \tag{13}$$

From the results of this calculation, α_i is primarily positive. The data that is positively correlated with α_i is called the support vector.

2.5.4 Naïve Bayes (NB)

Naive Bayes is a classifier that uses conditional probabilities or statistical calculations [26]. In other words, conditional probability is a measure of the occurrence of an event based on other events [27]. Therefore, this method is considered simple and effective to be applied in the classification and prediction process. This method applies Bayes' Theorem in determining the data class [28]. Classification of data classes based on Bayes' theorem makes Naive Bayes have a performance comparable to decision trees and neural networks. Several types of Naive Bayes methods exist, such as Multinomial, Bernoulli, and Gaussian. The classification of several Naive Bayes methods is based on the features used for the classification process. For example, the features used in the Naive Bayes Multinomial process are taken from a simple Multinomial distribution, and the features represent discrete quantities. The features used in the Bernoulli Naive Bayes process are the results of data conversion to binary (0 and 1) form. Meanwhile, the features used in the Gaussian Naive Bayes process are taken from a simple Gaussian distribution. Related to the Bayes theorem formula used in Naive Bayes, it can be formulated as follows Equation 14.

$$Prob(h|x) = \frac{prob(x|h)\ .\ prob(h)}{prob(x)} = \frac{prob(x|h)}{prob(x)} \times prob(h) \tag{14}$$

Where $prob(h|x) =$ data whose class is still unknown; $h =$ data hypothesis x; $prob(h|x) =$ hypothesis probability h based on x; $prob(h)$ = hypothesis probability h; $prob(x|h) =$ hypothesis probability x based on h; and $prob\ (x) =$ probability of x

2.5.5 Random Forest (RF)

Random forest is a bagging method, which is a method that generates several trees from sample data where the creation of one tree during training does not depend on the previous tree, then the decision is taken based on the most voting [29]. Two concepts that form the basis of random forest are building an ensemble of trees via bagging with replacement and a random selection of features for each tree built. First, each sample taken from the dataset for the training tree can be used again for another training tree. Second, the features used during training for each tree are a subset of the features owned by the dataset [30]. Ensemble-based classification will have maximum performance if there is a low correlation between essential learners. An ensemble must build a weak primary learner because a strong learner is likely to have a high correlation and usually causes an overfit. At the same time, a random forest minimizes correlations and maintains classification strength by randomizing the training process by selecting a number of features randomly. Of all the features in each training tree, then use the selected features to get the optimal branching tree.

Classify the sample data by combining the prediction results of several *k* tree classifiers based on the majority vote rule. Equation 15, the combined splitter H is created as an aggregation of each splitter $H_k, k = 1, \dots, K$ and the example d_i is classified to class c_j according to the number of votes obtained from the particular splitter H_k.

$$H(D_i, c_j) = sign \sum_{k=1}^{K} H_k(D_i, c_j) \tag{15}$$

The probability value of $H(d_i, c_j)$ ranges between 0 and 1. The cut off value used is 0.5. If $H(d_i, c_j)$ < 0.5 then it will be classified as category 0. Suppose $\widehat{H_k}(d_i, c_j)$ is prediction class of the k-tree random forest, then $\widehat{H_{rf}^K}(d_i, c_j) = majority\ vote\ \{\widehat{H_k}\ (d_i, c_j)\}_1^K$. The prediction model's performance generated based on analysis using the random forest method can be measured based on errors or errors formed from the prediction results. The error value can be obtained by Equation 16 calculating the Mean Absolute Percentage Error (MAPE).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|X_i - F_i|}{X_i} \times 100 \tag{16}$$

Where X_i = *i* - actual value*;* F_i = *i* - prediction value*; and n* = total of data*.*

Based on the MAPE formula, which describes the error value of the prediction model, the accuracy of the model can be obtained by the following calculation Equation 17.

$$Accuracy = 100\% - MAPE \tag{17}$$

2.5.6 Evaluation Metrics Performance

A performance matrix is used to determine the performance of a model. Several parameters are used to measure a model's performance, namely precision, sensitivity, accuracy, and F-Measure. The following is the Equation 18, Equation 19, Equation 20, and Equation 21 for each parameter of the performance matrix [31].

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (18)$$

$$Sensitivity = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (19)$$

$$F - Measure = \frac{2\ x\ Precision\ x\ Sensitivity}{(Precision + Sensitivity)} \quad (20)$$

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + True\ Negative +\ False\ Positive + False\ Negative)} \quad (21)$$

3. Results and Discussion

Disrupting blood supply to the blood-brain can cause a stroke [13], [41]. This condition causes certain areas of the brain not to receive oxygen and nutrients, resulting in the death of brain cells [14]. A stroke is a medical emergency because, without a supply of oxygen and nutrients, the cells in the affected part of the brain can die in just a matter of minutes [1], [42]. As a result, the body parts controlled by these brain areas cannot function properly [2]. The causes of stroke are generally divided become two, namely the presence of a blood clot in a blood vessel in the brain and a rupture of a blood vessel in the brain [13]. Narrowing or rupture of these blood vessels can occur due to several factors, such as high blood pressure, use of blood-thinning drugs, brain aneurysms, and brain trauma [2]. Some of the factors that cause stroke can be divided into two factors: modifiable and non-modifiable factors [15]. The modifiable factors were gender, age, heart disease, and genetics, while the non-modifiable factors were hypertension, smoking, glucose, BMI, and physical activity. In addition to the factors above, several factors may cause a stroke, such as marital status, type of work, and residence.

Based on the distribution of EMR data used in this study, female stroke patients have a higher percentage than male stroke patients. It shows that women have a higher risk of stroke than men. Another study also revealed that women have a 25.1% risk of stroke compared to men (24.7%) [32]. However, this percentage is still subject to change depending on regional variations. In Eastern Europe, women have a 36.5% risk of stroke, while in East Asia, women have a 36.3% risk of stroke [32]. In the United States, women also have a higher risk of stroke than men, which is 20-21% (for women) compared to 14-17% (for men) [33]. Some factors that make women more prone to stroke are pregnancy, high blood pressure during pregnancy, using certain types of birth control drugs, having higher rates of depression, and other gender-specific factors [15]. However, if we look at the number of stroke patients who have a history of hypertension (62 patients) and heart disease (39 patients) with those without hypertension (139 patients) and heart disease (162 patients), it is seen that people who do not appear to have hypertension and heart disease are not necessarily free from the threat of stroke. It may be because hypertension and heart disease has not been detected before, and a stroke occurs suddenly in someone without realising it. Hypertension and heart disease are leading causes of stroke [34]. Hypertension occurs when the blood vessels are not relaxed enough [35]. It creates a higher resistance to pumping blood through the circulatory system, including the heart [34]. Therefore, hypertension affects the health conditions of the heart and blood vessels that can cause stroke [35].

In addition, married people also have a higher risk than those who have never been married. It can be caused by the burden on the mind of someone married to be more than the unmarried. This burden of thought can cause feelings of stress and trigger a stroke [36]. The link between stress and stroke is significantly correlated and undeniable [37]. First, stress can cause the heart to work harder, raise blood pressure, and increase blood sugar and fat levels. If left continuously, these things are very likely to increase the risk of forming clots in the heart or brain, causing strokes [36]. In other words, chronic stress can trigger problems with the blockage or rupture of blood vessels in the brain (stroke). Smoking also has a very close relationship with the risk of stroke [38]. This habit can increase the risk of stroke by up to two times by raising blood pressure and reducing oxygen in the blood. Because when someone smokes, the chemicals in cigarettes (carbon monoxide, formaldehyde, arsenic, and cyanide) enter the lungs and are transferred into the bloodstream [39]. Blood containing these chemicals then flows throughout the body, changing and damaging its cells and affecting how people's bodies work [38]. These changes can then increase the risk of stroke. In addition, smoking can also make the blood thicker, increasing the risk of forming blood clots that can cause strokes [39].

Based on previous study, people who work harder also have a higher risk of stroke than people who work moderately or casually [40]. Private workers sometimes work harder than government employees. The workload of private employees is more than that of government employees, and the welfare/income of government employees is more secure than private workers. Therefore, the number of stroke patients from the private sector tends to be more than the number of stroke patients from government employees. The type of residence also affects the distribution of

stroke patients. Stroke sufferers in urban areas tend to be more than in rural areas. It is due to the pattern or lifestyle of urban communities that are not healthy, such as rarely doing sports, irregular eating patterns, and even a lot of workloads. In addition, several other factors, such as age, BMI (Body Mass Index), and glucose levels, affect the occurrence of stroke. Figure 3 presents the statistics (mean, median, and standard deviation) values of several factors (age, BMI, and glucose levels) that cause a stroke.

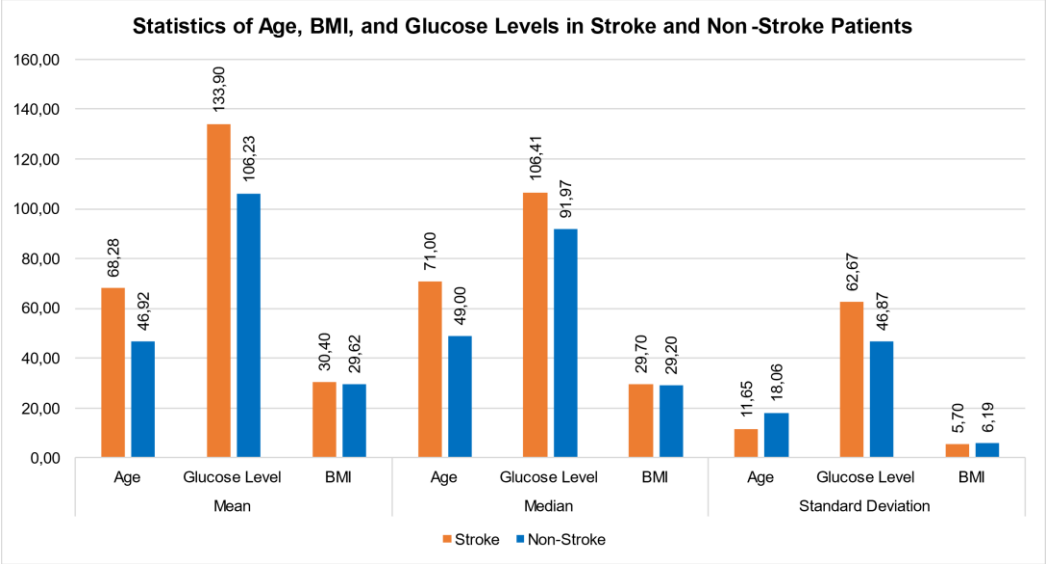

Figure 3. Statistic of Age, BMI, and Glucose Levels in Stroke and Non-Stroke Patients

Based on the statistics above, stroke tends to be suffered by the elderly with an average age of 62.28 years (the median value based on EMR data is 71 years). The glucose level in stroke patients is also high, at an average level of 133.9 mg/dL (the mean value based on EMR data is 106.41 mg/dL). Meanwhile, BMI in stroke patients also tends to be higher than in non-stroke patients, with an average of 30.4 (the mean value based on EMR data is 29.7). The standard deviation values for each of the factors that cause stroke tend to vary (sequentially 11.65 for the age factor, 62.67 for the BMI, and 5.7 for the glucose level) because the standard deviation value aims to determine the closeness of the data from the sample. Based on the results of the Pearson correlation using several factors that may cause stroke, the correlation coefficient values are 0.252 (for age), 0.129 (for glucose levels), 0.018 (for BMI), 0.138 (for hypertension), 0.128 (for heart disease), 0.071 (for marital status), 0.020 (for type of work), 0.009 (for type of residence), and 0.040 (for smoking). The age factor is the most dominant in triggering a stroke when viewed from the coefficient value. Other factors that are quite dominant include glucose levels, hypertension, heart disease, and marital status. As for the BMI factor, type of work, residence, and smoking have a relatively small correlation value compared to other factors. Figure 4 is the Pearson correlation value of several factors that cause a stroke.

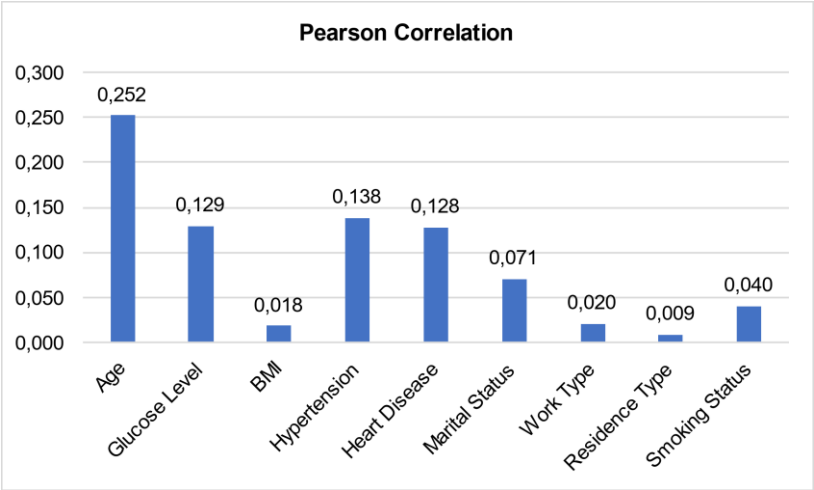

Figure 4. Pearson Correlation of Stroke Factors

In addition, we also predict strokes using XAI algorithms such as Probabilistic Neural Network (PNN), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB). The input parameters used in the stroke prediction are age, glucose level, BMI, hypertension, heart disease, marital status, type of occupation, type of residence, and smoking. The data used for stroke prediction are 3412 (201 stroke patient data and 3211 non-stroke), with a portion of 70% for the training process and the remaining 30% for the testing process. Several parameters were used as benchmarks to determine the performance level of the stroke prediction model, such as precision, sensitivity, F-measure, and accuracy values. The value of the performance matrix of each XAI algorithm for stroke prediction is shown in Figure 5.

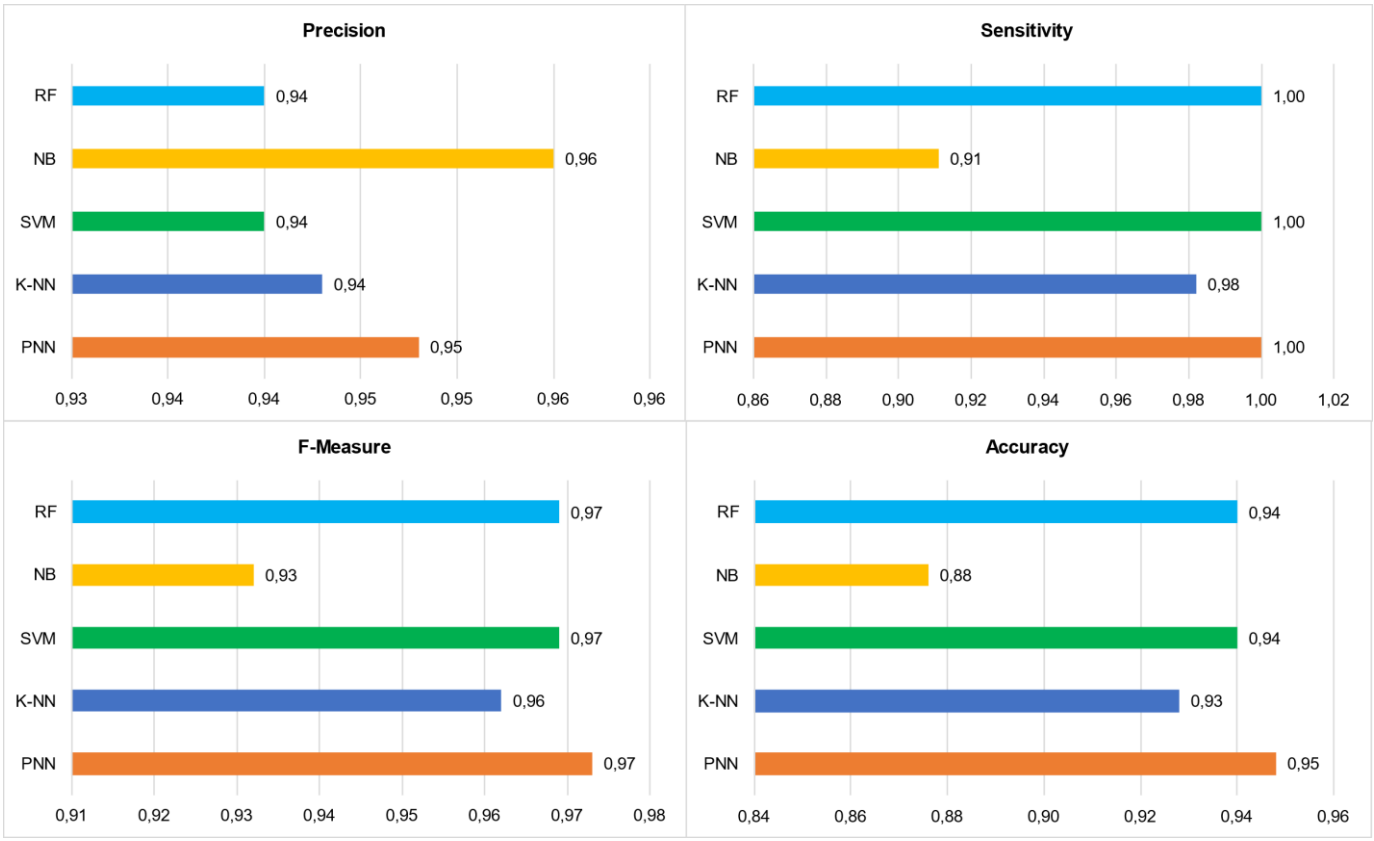

Figure 5. XAI Algorithms Performance Metrics for Stroke Prediction

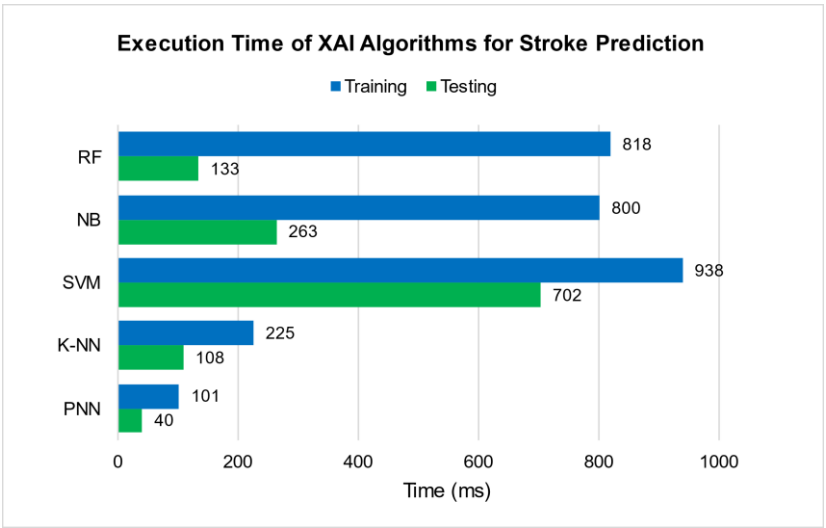

Figure 6. Execution Time of XAI Algorithms for Stroke Prediction

Based on the performance matrix, PNN has the highest accuracy, sensitivity, and F-measure levels of 95%, 100%, and 97% (respectively) compared to other algorithms such as RF, NB, SVM, and KNN. The precision level of PNN is 95% (below NB, which is 96%). In addition, the accuracy values obtained are 94% for RF, 88% for NB, 94% for SVM, and 93% for K-NN. The sensitivity values for RF are 100%, 91% for NB, 100% for SVM, and 98% for K-NN. The precision values for RF are 94%, 96% for NB, 94% for SVM, and 94% for K-NN. At the same time, the F-Measure value for RF is 97%, 93% for NB, 97% for SVM, and 96% for K-NN. The execution time of the PNN algorithm is also faster than other XAI algorithms. As a comparison, studies on stroke prediction using deep learning algorithms such as GRU, LSTM, bi-LSTM, and FFNN have accuracy rates of 95.6%, 87%, 91%, and 83% [6]. It proves that stroke prediction in this research using several XAI algorithms can be made very well, considering that the accuracy, sensitivity, precision, and F-Measure values are high.

4. Conclusion

This study aims to analyze the several factors that influence stroke in medical records using statistical analysis (correlation value) and stroke prediction using the XAI algorithm. Factors analyzed included gender, age, hypertension, heart disease, marital status, residence type, occupation, glucose level, BMI, and smoking. Based on the study results, we found that women have a higher risk of stroke than men, and even people who do not have hypertension and heart disease (hypertension and heart disease are not detected early) still have a high risk of stroke. Married people also have a higher risk of stroke than unmarried people. It may be triggered by conflicts between partners during their life together. Thus, a person's feelings of stress will begin to appear and cause a stroke. In addition, bad habits such as smoking, working with very intense thoughts and activities, and the type of living environment that is not conducive can also trigger a stroke. Increasing age, BMI, and glucose levels certainly affect a person's stroke risk. When viewed based on the average age, BMI, and glucose levels of stroke patients, they were 68.28 years, 133.90 mg/dL, and 30.40. However, stroke patients can have age, BMI, and glucose levels above or below average.

We have also succeeded in predicting stroke using the EMR data with high accuracy, sensitivity, and precision. Based on the performance matrix, PNN has the highest accuracy, sensitivity, and F-measure levels of 95%, 100%, and 97% (respectively) compared to other algorithms such as RF, NB, SVM, and KNN. The PNN precision level is 95% (below the NB, which is 96%). The execution time of the PNN algorithm is also faster than other XAI algorithms. For further research, more data on the EMR of stroke patients is needed in the study. It is to strengthen the pattern of information related to the dominant factors for the emergence of stroke in a person. In addition, it is necessary to modify the XAI-based algorithm to increase the accuracy, precision, and sensitivity of predictions.

Acknowledgement

Thank you to the research team and lecturers in the ITS Medical Technology study program, who have become a forum for researchers to develop this journal research. Hopefully, this research can significantly contribute to advancing AI-based medical system development.

References

[1] Shikany, J. M., Safford, M. M., Soroka, O., Newby, P., Brown, T. M., Durant, R. W., & Judd, S. E. “Associations of Dietary Patterns and Risk of Sudden Cardiac Death in The Reasons for Geographic and Racial Differences in Stroke Study Differ by History of Coronary Heart Disease.” *Circulation*, 141(1), 2020. https://doi.org/10.1161/circ.141.suppl_1.P520

[2] Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., & Carson, A. P. “Heart Disease and Stroke Statistics-2019 Update: A report from the American Heart Association.” *Circulation*, 139(10), e56-e528, 2019. https://doi.org/10.1161/CIR.0000000000000659

[3] Aprianda, R. “Annual Report of the Indonesian Health Social Security Administering Agency.” *Data and Information Center of the Ministry of Health*, 2019.

[4] Choi, Y. A., Park, S. J., Jun, J. A., Pyo, C. S., Cho, K. H., Lee, H. S., & Yu, J. H. “Deep Learning-Based Stroke Disease Prediction System Using Real-Time Bio Signals.” *Sensors, 21*(4269), 1-17, 2021. https://doi.org/10.3390/s21134269

[5] Fang, G., Huang, Z., & Wang, Z. “Predicting Ischemic Stroke Outcome Using Deep Learning Approaches.” *Frontiers in Genetic, 12,* 1-8, 2022. https://doi.org/10.3389/fgene.2021.827522

[6] Kaur, M., Sakhare, S. R., Wanjale, K., & Akter, F. “Early Stroke Prediction Methods for Prevention of Strokes.” *Behavioural Neurology*, *2022*, 1-9, 2022. https://doi.org/10.1155/2022/7725597

[7] Tocchetti, A., & Brambilla, M. “The Role of Human Knowledge in Explainable AI.” *Data*, *7*(93), 1-20, 2022. https://doi.org/10.3390/data7070093

[8] Ennab, M., & Mcheick, H. “Designing an Interpretability-Based Model to Explain the Artificial Intelligence Algorithms in Healthcare.” *Diagnostics*, *12*(1557), 1-17, 2022. https://doi.org/10.3390/diagnostics12071557

[9] Madanu, R., Abbod, M. F., Hsiao, F. J., Chen, W. T., & Shieh, J. S. “Explainable AI (XAI) Applied in Machine Learning for Pain Modeling: A Review.” *Technologies*, *10*(3), 1-15, 2022. https://doi.org/10.3390/technologies10030074

[10] Sarp, S., Kuzlu, M., Wilson, E., Cali, U., & Guler, O. “The Enlightening Role of Explainable Artificial Intelligence in Chronic Wound Classification.” *Electronics*, *10*(1406), 1-15, 2021. https://doi.org/10.3390/electronics10121406

[11] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. “Explainable AI: A Review of Machine Learning Interpretability Methods.” *Entropy*, *23*(18), 1-45, 2021. https://dx.doi.org/10.3390/e23010018

[12] Obayya, M., Nemri, N., Nour, M.K., Al Duhayyim, M., Mohsen, H., Rizwanullah, M., Sarwar Zamani, A., & Motwakel, A. “Explainable Artificial Intelligence Enabled TeleOphthalmology for Diabetic Retinopathy Grading and Classification.” *Applied Sciences*, *7*(93), 1-18, 2022. https://doi.org/10.3390/app12178749

[13] Alanazi, E. M., Abdou, A., & Luo, J. "Predicting Risk of Stroke from Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models." *JMIR Formative Research*, *5*(12), 1-10, 2021. https://doi.org/10.2196/23440
[14] Guo, Q., Kawahata, I., Cheng, A., Jia, W., Wang, H., & Fukunaga, K. "Fatty Acid-Binding Proteins: Their Roles in Ischemic Stroke and Potential as Drug Targets." *International Journal of Molecular Science*, *23*(17),1-16, 2022. https://doi.org/10.3390/ijms23179648
[15] Kuriakose, D., & Xiao, Z. "Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives." *International Journal of Molecular Sciences*, *21*(20), 1-24, 2020. https://doi.org/10.3390/ijms21207609
[16] Profillidis, V. A., & Botzoris, G. N. "Modeling of Transport Demand." *Amsterdams: Elsevier Inc*, 2019.
[17] Sun, C., Hu, Y., & Shi, P. "Probabilistic Neural Network-Based Seabed Sediment Recognition Method for Side-Scan Sonar Imagery." *Sedimentary Geology*, 410, 2020. https://doi.org/10.1016/j.sedgeo.2020.105792
[18] Chaki, S., Routray, A., & Mohanty, W. K. "A Probabilistic Neural Network (PNN) based Framework for Lithology Classification using Seismic Attributes." *Journal of Applied Geophysics,* 199, 2022. https://doi.org/10.1016/j.jappgeo.2022.104578
[19] Specht, D. F. "Probabilistic Neural Networks." *Neural Networks*, 3(1), 109-118, 1990. https://doi.org/10.1016/0893-6080(90)90049-q
[20] Ancona, F., Colla, A. M., Rovetta, S., & Zunino, R. "Implementing Probabilistic Neural Networks." *Neural Computing & Applications*, 5(3), 152–159, 1997. https://doi.org/10.1007/bf01413860
[21] Rozos, E., Koutsoyiannis, D., & Montanari, A. "KNN vs Bluecat-Machine Learning vs Classical Statistics." *Hydrology,* 9(101), 1-15, 2022. https://doi.org/10.3390/hydrology9060101
[22] Rungskunroch, P., Shen, Z. J., & Kaewunruen, S. "Benchmarking Socio-Economic Impacts of High-Speed Rail Networks Using K-Nearest Neighbour and Pearson's Correlation Coefficient Techniques through Computational Model-Based Analysis." *Applied Sciences*, *12*(1520), 1-26, 2022. https://doi.org/10.3390/app12031520
[23] Miao, Y., Hunter, A., & Georgilas, I. "An Occupancy Mapping Method Based on K-Nearest Neighbours." *Sensors*, *22*(1), 1-17, 2022. https://doi.org/10.3390/s22010139
[24] Rath, S. K., Sahu, M., Das, S. P., Bisoy, S. K., & Sain, M. "A Comparative Analysis of SVM and ELM Classification on Software Reliability Prediction Model." *Electronics*, *11*(2707), 1-14, 2022. https://doi.org/10.3390/electronics11
[25] Li, L., Ke, Y., Zhang, T., Zhao, J., & Huang, Z. "A Human Defecation Prediction Method Based on Multi-Domain Features and Improved Support Vector Machine." *Symmetry*, *14*(9), 1-23, 2022. https://doi.org/10.3390/sym14091763
[26] Barros, W. K. P., Barbosa, M. T., Dias, L. A., & Fernandes, M. A. C. "Fully Parallel Proposal of Naive Bayes on FPGA." *Electronics*, *11*(2565), 1-15, 2022. https://doi.org/10.3390/electronics11162565
[27] Alsolai, H., & Roper, M. "The Impact of Ensemble Techniques on Software Maintenance Change Prediction: An Empirical Study." *Applied Sciences, 12*(5234), 1-28, 2022. https://doi.org/10.3390/app12105234
[28] Boonnam, N., Udomchaipitak, T., Puttinaovarat, S., Chaichana, T., Boonjing, V., & Muangprathub, J. "Coral Reef Bleaching under Climate Change: Prediction Modeling and Machine Learning." *Sustainability*, *14*(10), 1-13, 2022. https://doi.org/10.3390/su14106161
[29] Song, J., Gao, J., Zhang, Y., Li, F., Man, W., Liu, M., Wang, J., Li, M., Zheng, H., & Yang, X. "Estimation of Soil Organic Carbon Content in Coastal Wetlands with Measured VIS-NIR Spectroscopy Using Optimized Support Vector Machines and Random Forests." *Remote Sens*, *14*(17),1-21, 2022. https://doi.org/10.3390/rs14174372
[30] Corradino, C., Amato, E., Torrisi, F., & Del Negro, C. "Data-Driven Random Forest Models for Detecting Volcanic Hot Spots in Sentinel-2 MSI Images." *Remote Sens*, *14*(17), 1-18, 2022. https://doi.org/10.3390/rs14174370
[31] Kumar, V., Lalotra, G. S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., Shorfuzzaman, M., Alsufyani, A., & Uddin, M. "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques." *Healthcare*, *10*(1293), 1-28, 2022. https://doi.org/10.3390/healthcare10071293
[32] Feigin, V. L., Nguyen, G., Cercy, K., Johnson, C. O., Alam, T., Parmar, P. G., Abajobir, A. A., Abate, K. H., Abdullah, F., & Abejie, A. N. "Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016." *N Engl J Med, 379,* 2429–2437, 2018. https://doi.org/10.1056/NEJMoa1804492
[33] Seshadri, S., Beiser, A., Hayes, K. M., Kase, C. S., Au, R., Kannel, W. B., & Wolf, P. A. "The Lifetime Risk of Stroke: Estimates from the Framingham Study." *Stroke*, *37*, 345–350, 2006. https://doi.org/10.1161/01.STR.0000199613.38911.b2
[34] Bassa, B., Güntürkün, F., Craemer, E. M., Lamadé, M. U., Jacobi, C., Bassa, A., & Becher, H. "Diabetes, Hypertension, Atrial Fibrillation and Subsequent Stroke-Shift towards Young Ages in Brunei Darussalam." *Int J Environ Res Public Health*, *19*(14), 1-9, 2022. https://doi.org/10.3390/ijerph19148455
[35] Sobierajski, T., Surma, S., Roma, M., Labuzek, K., Filipiak, K. J., & Oparil, S. "What Is or What Is Not a Risk Factor for Arterial Hypertension? Not Hamlet, but Medical Students Answer That Question." *Int J Environ Res Public Health*, *19*(13), 1-12, 2022. https://doi.org/10.3390/ijerph19138206
[36] Iriepa, D., Knez, D., Gobec, S., Iriepa, I., Ríos, C., Bravo, I., Munoz, F., Contelles, J., & Litina, D. "Polyfunctionalized α-Phenyl-tert-butyl(benzyl)nitrones: Multifunctional Antioxidants for Stroke Treatment." *Antioxidants, 11*(1735), 1-20, 2022. https://doi.org/10.3390/antiox11091735
[37] Wang, X., Thiel, L., & Graff, N. D. "Mindfulness and Relaxation Techniques for Stroke Survivors with Aphasia: A Feasibility and Acceptability Study." *Healthcare*, *10*(1409), 1-15, 2022. https://doi.org/10.3390/healthcare10081409
[38] Ohlrogge, A. H., Frost, L., & Schnabel, R. B. "Harmful Impact of Tobacco Smoking and Alcohol Consumption on the Atrial Myocardium." *Cells*, *11*(2576), 1-29, 2022. https://doi.org/10.3390/cells11162576
[39] Sifat, A. E., Nozohouri, S., Archie, S. R., Chowdhury, E. A., & Abbruscato, T. J. "Brain Energy Metabolism in Ischemic Stroke: Effects of Smoking and Diabetes." *International Journal of Molecular Sciences*, *23*(15), 1-25, 2022. https://doi.org/10.3390/ijms23158512
[40] Wicht, C. A., Chavan, C. F., Annoni, J. M., Balmer, P., Aellen, J., Humm, A.M., Roten, F., Spierer, L., & Medlin, F. "Predictors for Returning to Paid Work after Transient Ischemic Attack and Minor Ischemic Stroke." *Journal of Personalized Medicine, 12*(7), 1-11, 2022. https://doi.org/10.3390/jpm12071109
[41] Tadi, P., & Lui, F. Acute Stroke. Treasure Island, Florida: StatPearls Publishing, 2022.
[42] Hui, C., Tadi, P., & Patti, L. Ischemic Stroke. Treasure Island, Florida: StatPearls Publishing, 2022.