Ischemic Stroke Classification using Random Forests Based on Feature Extraction of Convolutional Neural Networks

Glori Stephani Saragih[a], Zuherman Rustam[a,1], Dipo Aldila[a], Rahmat Hidayat[b], Reyhan E. Yunus[c], Jacub Pandelaki[c]

[a] Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia
E-mail: [1]rustam@ui.ac.id

[b] Department of Information Technology, Politeknik Negeri Padang, Padang, 25164, Indonesia
E-mail: rahmat@pnp.ac.id

[c] Department of Radiology, Cipto Mangunkusumo Hospital, Jakarta, 10430, Indonesia

Abstract— Stroke has become a global health problem, due to high mortality and disability, with two-thirds of all strokes occurring in developing countries. In Indonesia, stroke is a disease with the highest mortality rate, namely in the first rank for more than two decades, 1990-2017. Stroke is divided into two types, ischemic and hemorrhagic; however, 87% of stroke sufferers are ischemic stroke. Suppose an ischemic stroke is found, and the patient is a new sufferer. In that case, the patient should get direct treatment because there is a golden period in stroke management that is if 4.5 hours to help and reduce the risk of death or permanent disability. High mortality and disability raise awareness of the importance of early detection of ischemic stroke; therefore, research has been carried out, especially in technology. To carry out automatic diagnosis, machine learning and deep learning can be used, especially because of their ability to provide high accuracy prediction results. In this study, the authors will provide an update in the detection of ischemic stroke based on patient CT scan by replacing NN's role on CNN with random forests. Thus, after feature extraction on CNN, the fully connected layer on CNN is completely replaced by random forests in classifying data. Based on the proposed method, the accuracy of testing is 100% when the percentage of the testing dataset is 10% and the number of trees more than 100 with criterion Gini or entropy.

Keywords— convolutional neural networks; image classification; neuroimaging; random forests; stroke ischemic.

I. INTRODUCTION

According to the Global Burden of Disease Study, stroke is one of the diseases with the most deaths in the world. In 2017, the percentage of total deaths caused by stroke was 11.02%. Stroke can be defined as an injury to the brain caused by blockage of blood vessels and making inadequate blood supply that causes bleeding in the brain parenchyma [1]. The stroke itself is a cardio-cerebrovascular disease classified into a catastrophic disease because it requires therapy with special expertise, uses sophisticated medical devices, and or requires lifelong health services [2]. This disease requires a long time for the healing process so that stroke absorbs the cost of large health claims. According to data from the *Badan Penyelenggara Jaminan Sosial* (BPJS)/ Social Security Administrator that each year, the cost of services for stroke always increases. In 2016, it spent IDR 1.43 trillion, then in 2017, it rose to IDR 2.18 trillion, and in 2018 it reached IDR 2.56 trillion [3]. Therefore, stroke is a big problem for the community and government levels, so early detection of this disease is needed.

Early detection of stroke needs to be given special attention, so we can reduce the number of cases from year to year. When a person experiences paralysis symptom in one part of his body, it is necessary to do a further examination to determine the cause of the paralysis from a stroke, infection, tumor, or something else. In general, investigation through a Computerized Tomography Scan (CT scan) is the first step taken before the patient receives treatment. An alternative to CT scanning is Magnetic Resonance Imaging (MRI), but MRI costs are more expensive. The examination time is longer than a CT scan; therefore, doctors will recommend patients for CT scan [4]. CT scan is useful to distinguish the type of stroke suffered by the patient between ischemic or hemorrhagic stroke. If you find a hemorrhagic stroke, it will generally be given hemorrhage treatment or brought to a neurosurgeon. But suppose an ischemic stroke is found and the patient is a new sufferer. In that case, the

patient will get direct treatment because there is a golden period in stroke management that is if 4.5 hours to reduce the risk of death or permanent disability [5]. So, it is important to detect stroke quickly and precisely, so the type of treatment given is also fast and on target.

Patients with ischemic stroke reach 87%, and the rest are sufferers who experience intracerebral and subarachnoid hemorrhage, which is pathologically a hemorrhagic stroke [6]. The number of ischemic stroke sufferers more than hemorrhagic is why the author chose ischemic stroke for further discussion. Ischemic stroke generally causes changes in density in the brain; these changes can be seen on CT scans with darker results, namely the hypodense area. Hypodense is intended for softer tissue, such as air or liquid. However, the CT scans ischemic stroke location is not very clear, so the diagnosis depends on the doctor in assessing the results. High mortality and disability raise awareness of the importance of early detection of ischemic stroke; therefore, research has been carried out, especially in technology. In recent years, *deep learning* and *machine learning* have provided new directions for detecting CT scan for ischemic stroke sufferers because of their ability to provide predictive results with high accuracy. *Machine learning* is part of artificial intelligence, which is divided into supervised and unsupervised learning. Machine learning is the design of algorithms that computers use to perform certain tasks without using explicit instructions by using patterns and interventions instead [7]. Meanwhile, deep learning is a branch of machine learning consisting of a high-level abstraction modeling algorithm in data using a set of non-linear transformation functions arranged in layers and depth [8].

In 2019, Clerigues et al. [9] presented and evaluated an automation method for segmenting the nucleus of acute stroke lesions on CT scan images using Convolutional Neural Networks (CNN). Barros, et al. [10] automatically segmented subarachnoid hemorrhage and Chin, et al. [4] detected ischemic stroke using CNN. From their second research, CNN produced good work performance with an accuracy of more than 90%. From the literature above, the contribution of deep learning, namely CNN in detecting ischemic strokes, results in high accuracy in early detection. However, in its use, it requires a device that is high cost and time-consuming. Some studies combine the two approaches with differentiating methods for feature extraction and classification. Classified based on the results of feature extraction. Rajini and Bhavani [11] detected ischemic stroke with segmentation and texture features that consist of several stages, namely, pre-processing, feature extraction with Gray Level Co-Occurrence Matrix (GLCM), and classification with Support Vector Machines (98% accuracy), KNN (97% accuracy), K-Means (96% accuracy). Maier et al. [12] found that the feature extraction used is intensity, the weighted local mean, 2D center distance, and the local histogram. These are are then classified by using KNN, Gaussian Naive Bayes (GNB), Generalized Linear Models (GLM), Gradient Boosting, AdaBoost, Extra Trees, Random Forests (RF), CNN. RF and CNN have a better performance, the dice coefficient of 0.67 ± 0.18. Ho, et al. [13] used a machine learning approach to classify ischemic stroke onset time (TSS) with image data. Feature extraction used, namely;

descriptive statistics in the region of interest and morphological features, then classified into logistic regression (LR), gradient boosted regression tree (GBRT), support vector machine (SVM), stepwise multilinear regression (SMR) and RF. The cross-validation results show that the author's best classifier reaches the area under the curve of 0.765, with a sensitivity of 0.788 and a negative predictive value of 0.609, namely with LR.

Lee, et al. [14] compared several machine learning to identify stroke within 4.5 hours. Feature extraction used are: Intensity (4), Gradient (4), GLCM (21), Gray-Level Run Length Matrix (11), Local Binary Pattern (4) and classification by logistic regression, SVM and RF with a specificity of 82.6% and logistic regression sensitivity and RF 75.8% and SVM 72.7%. Machine learning algorithms are shown to have significantly greater sensitivity than human readers. Another study published this year is the automation of segmentation and classification of strokes with expectation-maximization and RF by Subudhi, et al. [15]. The authors detected ischemic stroke using a diffusion-weighted image sequence (DWI) of MR images. The part of the brain affected by a stroke is segmented using the expectation-maximization algorithm. The segmented area is then further processed with the Darwinian Particle Swarm Optimization (FODPSO) fractional technique to improve detection accuracy. A total of 192 MRI scans were considered for evaluation. Different morphological and statistical features were extracted from segmented lesions to form feature sets which were then classified by SVM and RF. The proposed system efficiently detects stroke lesions with an accuracy of 93.4% using RF, which is better than SVM results.

CNN is a development method of Neural Networks (NN), namely by adding a layer after the image input called a convolutional layer. The CNN method consists of two stages, namely feature extraction (convolutional layer and pooling layer) and trainable classifier (fully connected neural networks) [16]. The convolutional layer and pooling layer will produce an input layer for the next process, namely the NN's fully connected layer. In CNN, the NN method plays a role in the data classification process based on the extraction results from image data to numeric.

Apart from NN, other machine learning methods can be used for the classification process. In machine learning, random forests are a method that is widely used for classification in the analysis of medical images and other data in the health sector. Based on the research results, random forests are considered an effective method because they produce more accurate accuracy than other classification methods [16]. In 2019, Rustam and Saragih predicted schizophrenia with random forests; the results obtained are 100% accuracy, with a percentage of 40% training data [17]. The percentage of training data used in Rustam and Saragih's research proves that random forests still produce optimal accuracy with a small percentage of training data. In contrast to NN, the performance is more optimal if the available training data increases [8]. Another advantage of random forests compared to NN is that they are easy to interpret both structurally and predictably, unlike NN, which is a black box method or difficult to interpret.

Based on RF ability in the classification process in ischemic stroke data [12]–[15], [18] and on data other health; prediction of diabetes mellitus with an accuracy of 80.8% [19], prediction of prostate cancer with an accuracy of 87% [20], and prediction of osteoarthritis with an accuracy of 86.96% [21], so in this study, the authors will provide an update in the detection of ischemic stroke based on patient CT scan by replacing the role of NN in CNN with random forests. Thus, after feature extraction on CNN, the fully connected layer on CNN is completely replaced by random forests in classifying data. In this study, the classification of ischemic strokes is carried out to categorize ischemic strokes into acute, sub-acute, and chronic, which are more crucial. Hence, it is expected that doctors and medical personnel in radiology can be assisted in diagnosing ischemic stroke patients quickly and accurately in determining the type of care to be given to the patient. This research is organized as follows: Section the background of research, Section 2 the material and method used. Section 3 discussed the result and analysis, and finally, conclusions are included in Section 4.

II. MATERIALS AND METHOD

A. Data

The dataset used in this research is the image data from the brain CT scan results of ischemic stroke patients obtained from the Department of Radiology, Cipto Mangunkusumo National General Hospital (RSCM), Indonesia. The data used were 92 images consisting of 48 images with no density changes and 44 images with density change in ischemic stroke patients. The image data used has dimensions of 512×512 pixels and the type of file is jpg. An example of image data used in this research is shown in Figure 1 and Figure 2.
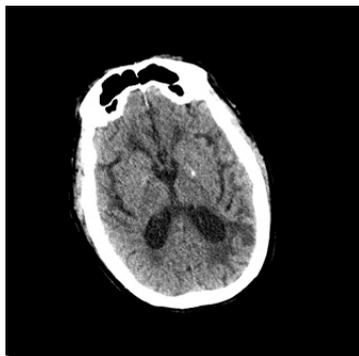

Fig. 1 There is a change in density

Fig. 2 There is no change in density

B. Convolutional Neural Networks

Convolutional Neural Networks is a development method of Neural Networks (NN), namely by adding a layer after the image input called a convolution layer. The CNN method consists of two stages, namely feature extraction (convolutional layer and pooling layer) and trainable classifier (fully connected neural networks) [16].

CNN uses convolutional, unification, and non-linearity layers such as tanh, sigmoid, and Rectified Linear Unit (ReLU) [22]. The advantage of the CNN method is that it can be used to detect and recognize objects not just in the middle of an image.

In this research, we built the model architecture that consists of the convolutional layer, activation function, pooling layer, and normalization layer by using Keras, as the python deep learning library, the details were given in Figure 1. As the activation function, ReLU is used to faster computations compared to other activation functions such as tanh and sigmoid functions [23]. Then, to reduce the size of input images, we used Maximum Pooling.

```
(1) INPUT: 152 x 152 x 1
(2) Conv2D: 3x3 size, 32 filters, 1 stride
(3) ReLU: max (0, hθ(x))
(4) MaxPooling: 2x2 size, 1 stride
(5) BatchNormalization
(6) Conv2D: 3x3 size, 64 filters, 1 stride
(7) ReLU: max (0, hθ(x))
(8) MaxPooling: 2x2 size, 1 stride
(9) BatchNormalization
(10) Conv2D: 3x3 size, 96 filters, 1 stride
(11) ReLU: max (0, hθ(x))
(12) MaxPooling: 2x2 size, 1 stride
(13) BatchNormalization
(14) Conv2D: 3x3 size, 64 filters, 1 stride
(15) ReLU: max (0, hθ(x))
(16) MaxPooling: 2x2 size, 1 stride
(17) BatchNormalization
(18) Conv2D: 3x3 size, 32 filters, 1 stride
(19) ReLU: max (0, hθ(x))
(20) MaxPooling: 2x2 size, 1 stride
(21) Dropout: rate=0.2
(22) Flatten output: 1 x 6273
```

Fig. 3 CNN architecture for feature extraction

Preventing model diverge is important, so we could go to a normalization layer. Last is we drop it out with a rate of 0.2, which means 80% of the information on every image is kept. The resulting neuron from the dropout layer is then flattened and form a vector with length 6273 for every image.

C. Random Forest

Definition *Random forests are a classification method consisting of clusters of classification trees $\{h(x, S^b), b = 1, ..., B\}$ where $\{S^b\}$ are independent random vectors and identical distributions and each tree gives a vote for the most popular class on input x* [24].

This method is an extension of the Classification and Regression Tree (CART) method with applying the bagging

(bootstrap aggregating) method, a random selection of features, and voting in determining the classification results. The classification results are determined from the most votes on the voting of each tree formed. The bootstrap method is used in a tree formation. Bootstrapping is a resampling method with returns. The basic idea is to replicate a set of observational data by drawing random samples with returns, each sample having the same size for the amount of observed data [25]. The number of bootstrap samples generated will determine the number of trees in the model. After the bootstrap method, each tree (bootstrap sample) is formed with the following rules:

- If there are M input variables, then the number of m predictor variables at each node satisfies $m \leq M$
- The variable m is chosen randomly from M
- The selection of the best predictor variable from m is selected by calculating the measure of purity (Gini or entropy). The best split on m is used to separate the nodes.
- The amount of m is kept constant during the growth of forests.
- Each tree is formed to the maximum extent possible, without pruning.

The choice of the best predictor variable in question is the predictor variable that provides more decision-making information. In tree formation, these variables are at the higher node level. The number of predictor variables that are usually used is $m = \sqrt{M}$, however, the best setting of the number of predictor variables selected randomly and the number of trees depends on the data [18]. The more trees that are formed and used in the decision-making process, the more robust the results will be [15].

D. Proposed Method

The flowchart of the proposed method is illustrated in Fig. 4. Developing an efficient method can help the doctor/radiologist diagnose ischemic stroke accurately and appropriately. The proposed method has four stages, pre-processing, feature extraction, classification, and evaluation model. The pre-processing is by changing the color into greyscale and resized to the same size 512×512 pixels. In the second stage, the CT scan images can be extracted to numeric; the input is passed to a bunch of CNN layers. After that, the result of feature extraction is classified by using RF. The proposed method's performance is determined with a confusion matrix; it shows the number of correct and incorrect predictions made by the model compared to the actual data [26]. Table I shows us the confusion matrix used in this research.

TABLE I
CONFUSION MATRIX

		Predicted Class	
		Density Change	No Density Change
Actual Class	Density Change	TP	FN
	No Density Change	FP	TN

However, there are four parts to this confusion matrix:
- TP: True Positive, observed in instances where density change is detected as density change

- TN: True Negative is when no density change is detected as no density change
- FP: False Positive is seen in cases where no density change is detected as density change
- FN: False Negative is when density change is detected as no density change.

Based on the value of TP, TN, FP, and FN from the confusion matrix, the valued accuracy is obtained with the following formula:

$$\text{Accuracy} = \frac{TP+TN}{TP + TN + FP + FN}$$
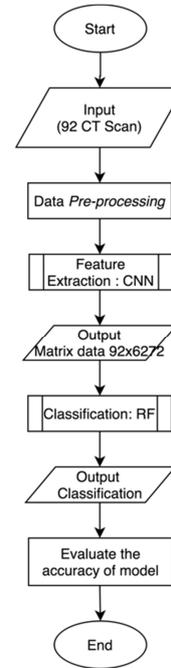

Fig. 4 Proposed Method

III. RESULT AND DISCUSSION

CT scan images were processed on CNN; the result was matrix data; after that we added a label in each row: 1 for there is a change density data or 0 for there is no change density data. Table I shows us the matrix input that we used as the new input for classification by using RF.

TABLE II
MATRIX INPUT

Data	y_1	y_2	…	y_{6272}	y_{6273}	y
x_1	5,96E-07	4,77E-07	…	7,15E-07	1,53E+00	1
x_2	9,53E-07	1,31E-06	…	2,62E+00	2,00E+00	1
…	…	…	…	…	…	
x_{91}	0,00E+00	7,15E-07	…	2,60E+00	2,04E+00	0
x_{92}	0,00E+00	8,34E-07	…	2,03E+00	2,30E+00	0

In this research, the algorithm was run ten times; repetition was carried out because of the random element in the method of random forests. Improving the accuracy is done by tuning the hyperparameters [27] is tracked for the combination of hyperparameters. There are two hyperparameters combined; the number of trees and the criterion. The number of trees used was 50, 100, 150, 200, 250, and 350. The criterion is a function that measures the equality of a split; the Gini index and entropy criterion for the information gain. The library that we used in this research is *scikit-learn*. The accuracy of training for

classification using random forest with Gini and/or entropy as a criterion correctly classified the ischemic stroke data, with a 100% accuracy for all percentage of training data set. Then, we will show the accuracy of the test by using random forests. Table III shows us the performance of stroke ischemic classification with Gini as a criterion, while entropy as a criterion in Table IV.

TABLE III
THE PERFORMANCE OF STROKE ISCHEMIC CLASSIFICATION USING RANDOM FORESTS WITH GINI AS A CRITERION

Percentage of Data Testing (%)	Number of Tree					
	50	100	150	200	250	350
	Accuracy of testing (%)					
10	90	100	100	100	100	100
20	89.473	89.473	94.736	94.736	94.736	94.736
30	85.714	85.714	89.285	92.857	92.857	92.857
40	84.211	81.578	92.105	86.842	92.103	92.105
50	85.106	85.106	85.106	87.234	89.316	89.361
60	83.928	82.142	85.714	87.5	89.285	89.285
70	83.333	80.303	83.333	84.848	86.363	87.878
80	73.333	81.333	84	85.333	89.333	89.333

TABLE IV
THE PERFORMANCE OF STROKE ISCHEMIC CLASSIFICATION USING RANDOM FORESTS WITH ENTROPY AS A CRITERION

Percentage of Data Testing (%)	Number of Tree					
	50	100	150	200	250	350
	Accuracy of testing (%)					
10	90	100	100	100	100	100
20	78.947	84.211	89.473	94.736	89.473	94.376
30	75	85.714	85.714	92.857	92.857	92.857
40	73.684	81.578	84.211	92.105	89.473	94.736
50	74.468	76.595	78.723	89.361	87.234	91.489
60	73.214	78.571	75	89.285	87.5	94.642
70	74.242	78.787	77.272	86.363	84.848	92.424
80	73.333	74.666	74.666	85.333	86.667	90.666

Table III shows us the highest accuracy of testing is obtained when the percentage of testing data set is 10% with a number of trees 100-350, with the accuracy is 100%. From Table IV, the entropy in the accuracy of testing data provides the slightly same results with Gini. The highest accuracy also provides when the percentage of testing data is 10%, is obtained 100% accuracy. Based on Table III and IV, we know the more trees, the accuracy of testing will increase, although the increase in not too significant and slightly the same. Then, the percentage of data used for testing greatly affects the performance of the model. The accuracy of testing is 100% in almost all the combinations of models using 10% of the testing data set, which means the training data set used is 90%, so the model learns more. Then, the use of Gini and entropy does not show much difference in the value of accuracy.

IV. CONCLUSION

In this research, we used the novelty method by changing the Neural Network function in Convolutional Neural Network (CNN) with random forests for classifying the stroke ischemic data based on CT scan. The CT scan images were obtained from the Department of Radiology, Cipto Mangunkusumo National General Hospital (RSCM), Indonesia. Every image in this dataset is passed through the convolutional layers on CNN. After that, we used the result of feature extraction to the random forest algorithm for classifying the CT scan image into density change and no density change. According to our experiments, our proposed method achieved the best performance with 100 percent accuracy when the percentage data usage of testing dataset is 10% and the number of trees used 100-350 trees with criterion used Gini or entropy. This result proves that our proposed method has the promised accuracy for stroke ischemic detection from CT scan images.

REFERENCES

[1] J. Stein, *Essentials of Physical Medicine and Rehabilitation Chapter 159 Stroke*, Philadelphia: Elsevier, 2020.

[2] Heniwati, and H. Thabrany, "Perbandingan Klaim Penyakit Katastropik Peserta Jaminan Kesehatan Nasional di Provinsi DKI Jakarta dan Nusa Tenggara Timur tahun 2014", *Jurnal Ekonomi Kesehatan Indonesia*, 2006, vol. 1, no. 2.

[3] Ministry of Health of the Republic of Indonesia. (2019). [Online]. Available: http://p2ptm.kemkes.go.id/artikel-sehat/hari-stroke-sedunia-2019-otak-sehat-sdm-unggul

[4] C. Chin, et. al, "An Automated Early Ischemic Stroke Detection System Using CNN Deep Learning Algorithm," *IEEE 8th*

International Conference on Awareness Science and Technology (iCAST), pp. 368-372, 2017.

[5] Ministry of Health of the Republic of Indonesia. (2017). [Online]. Available: http://p2ptm.kemkes.go.id/artikel-sehat/germas-cegah-stroke.

[6] American Stroke Association, *Let's talk about Ischemic Stroke*, Dallas: American Heart Association, 2020.

[7] J. G. Lee, et. al, "Deep Learning in Medical Imaging: General Review," *Korean Journal of Radiology,* vol. 18, no. 4, pp. 570-584, 2017.

[8] I. Goodfellow, et. al, *Deep Learning*, England: The MIT Press Cambridge, 2016.

[9] A. Clerigues, et al. "Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks," *Computers in Biology and Medicine,* vol. 115, 2019.

[10] R. S. Baros, et. al. "Automated segmentation of subarachnoid hemorrhages with convolutional neural networks," *Informatics in Medicine Unlocked*, vol. 19, 2020.

[11] N. H. Rajini, and R. Bhavani, "Computer aided detection of ischemic stroke using segmentation and texture features," *Measurement*, vol. 46, no. 6, 1865-1874, 2013.

[12] O. Maier, et. al, "Classifier for ischemic stroke lesion segmentation: A comparison study," *PloS One*, vol. 10, no. 12, 2015.

[13] K. C. Ho, et. al, "A machine learning approach for classifying ischemic stroke onset time from imaging," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1666-1676, 2019.

[14] H. Lee, et. al, "Machine learning approach to identify stroke within 4.5 hours," *Stroke*, vol. 51, no. 3, pp. 860-866, 2020.

[15] A. Subudhi, et. al, "Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 277-289, 2020.

[16] G. Cao, et. al, "A hybrid Cnn-Rf method for electron microscopy images segmentation," *Journal Biometrics Biomaterials and Tissue Engineering*, vol. 18, no. 114, 2013.

[17] Z. Rustam and G. S. Saragih, "Prediction schizophrenia using random forest," *TELKOMNIKA Telecommunication, Computing, Electronics and Control,* vol. 18, no 3, pp. 1433-1438, 2019.

[18] A. Vrtkova, "Predicting clinical status of patients after an acute ischemic stroke using random forests," *The International Conference on Information and Digital Technologies*, 2017.

[19] Q. Zou, "Predicting diabetes mellitus with machine learning techniques," *Frontier in Genetics*. Available: 9.doi:10.3389/fgene.2018.00515.

[20] M. Huljanah, et. al, "Feature selection using random forest classifier for predicting prostate cancer," *IOP Conference Series: Materials science and Engineering*, 546, 052031. Available: doi:10.1088/1757-899x/546/5/052031.

[21] U. Apriliani and Z. Rustam, "Osteoarthritis disease prediction based on random forest," *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. Available: doi:10.1109/icacsis.2018.8618166.

[22] A. F. M. Agarap, "An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification," *arXiv preprint arXiv*. 1712.03541v2, February 2019.

[23] L. Guifang, S. Wei, "Research on convolutional neural network based on improved Relu piecewise activation function," *Science Direct. Elsevier*., vol. 13 pp 977-984, 2018.

[24] L. Breiman, "Random forest", *Machine learning,* vol. 45, pp. 5-32, 2001.

[25] T. Hastie, et. al, *The Element of Statistical Learning Data Mining, Inference, and Prediction*, California: Springer, 2008.

[26] K. M. Ting, "Confusion Matrix," in Encyclopedia of Machine Learning and Data Mining, C. Sammut and G. I. Webb Eds. Boston, MA: Springer US, 2017, pp. 260-260.

[27] Z. Rustam and G. S. Saragih, "Predict bank financial failures using random forest," *IEEE*. Available: doi: 10.1109/IWBIS.2018.8471718.2018