Classification of Infarction using Random Forest

S H Rukmawan¹, F R Aszhari¹, Z Rustam¹ and J Pandelaki²

¹Department of Mathematics, University of Indonesia, Depok 16424, Indonesia
²Medical Department of Radiology, RSUPN Dr. Cipto Mangunkusumo, Kota Jakarta Pusat, DKI Jakarta 10430, Indonesia

*e-mail: shafirahanif@sci.ui.ac.id

Abstract. Stroke is a condition caused by disruption in the blood supply to the brain. When the flow of blood is decreasing and resulting dead brain tissue that is called an infarction. If this condition not treated immediately and don’t get the right treatment will cause the death of the brain. Therefore, the classification of infarction is important to help increase the life expectancy of the patients. In this study, we are using infarction data from the Department of Radiology at Dr. Cipto Mangunkusumo Hospital and propose a random forest method to help the health sector for detecting infarction quickly and accurately. The best result by using the random forest method is 94.44 percent with 65 percent as training data.

Keyword: Classification, infarction, random forest

1. Introduction

Stroke is the leading cause of long-term disability that cause of the death. In US, approximately 795,000 stroke events happen in each year [1]. In 2010, the absolute numbers of people with first stroke is 16.9 million patients, stroke survivors are 33 million patient, stroke-related deaths are 5.9 million, and disability-adjusted life-year is 102 million [2]. According to World Health Organization, one of the top ten causes of death worldwide is stroke, with approximately 5 million deaths in 2016 [3].

Stroke a condition caused by disruption brain function that occurs due to problems with blood vessels in the brain. Stroke are classified into two types, as ischemic stroke and hemorrhagic stroke [1]. Ischemic stroke is a condition whereby the blood supply to the brain is reduced or disrupted to a blockage of blood vessel [4]. If it is not treated immediately will cause the death of the brain tissue, that is called an infarction. Meanwhile, hemorrhagic stroke is a caused by an increase in blood pressure and it causes ruptured blood vessel in the brain [4].

Risk factor of stroke divided into modified and non-modified or genetic risk factors. The modified risk factors of stroke are important to reduce the risk factor of stroke by early identification. The modified risk factors of stroke such as hypertension, hyperlipidemia, diabetes, alcohol consumption, smoking [1]. The genetic or non-modified risk factors of strokes are age, sex, race, etc. Moreover, risk factors divided into short-term risks or triggers, intermediate-term risk factors, and long-term risk factors for stroke [1].

The diagnose of infarction early is important to increase the life expectancy of the patients. A cerebral infarction in patients can be seen in the brain by CT scan [5]. However, the result of the CT scan not enough to diagnose cerebral infarction. Therefore, machine learning is needed to help the health sector for classified or diagnosed infarction in patient using features and labels from the CT scan result.
Machine learning is algorithm that provides the system to automatically learn and improve from trial or experience without being explicitly programmed. Machine learning has several types of learning, one of them is supervised learning. Supervised learning usually used for prediction and classification. In this paper, Random forest was proposed to classify infarction data patient from CT scan result. Random forest is one of method of supervised learning that produces a highly accurate classifier, so we can classify data set accurately.

Infarction has been classified using several machine learning methods, they are using SVM with hybrid preprocessing method [4], and cubic Support Vector Machine (SVM) with gaussian Support Vector Machine (SVM) [5]. The other diseases that has been classified using supervised learning are cancer data using SVM and features selection [6], osteoarthritis using random forest [7], chronic kidney disease using SVM and random forest method [8], prostate cancer by feature selection using random forest [9], soft tissue tumor using stochastic SVM [10], brain cancer using SVM-RFE [11], and others disease.

2. Materials and methods

2.1 Dataset

In this paper, the dataset was obtained from Department of Radiology at Dr. Cipto Mangunkusumo Hospital. This infarction dataset consists of 206 data patients with 103 patients are infarction and 103 patients are non-infarction. The infarction dataset represented by 7 features and class (infarction and non-infarction). The features of dataset are shown in Table 1.

| Features Name | Features description           |
|---------------|--------------------------------|
| Area          | The size of the area from the infarction point |
| Maximum       | Maximum value of infarction    |
| Minimum       | Minimum value of infarction    |
| Sum           | Total amount value of infarction |
| Average       | Average value of infarction    |
| SD            | Standard error value of infarction |
| Length        | Length of infarction point     |

2.2 Random Forest (RF)

Random Forest has some advantages, there are best in accuracy, can handle input variable without deletion, effective for estimate missing data, etc. [12]. Random forest is supervised learning that build by several decision trees. In classification problems, ntree is the number of decision tree will be grown in random forest method [7]. Algorithm 1 show random forest method for classification [13].

**Algorithm 1 Random Forest Classifier Algorithm**

1. Draw bootstrap samples from the data
2. Grow a decision tree from each bootstrap data set.
3. Repeat step 1 and 2, N times.
4. At each node:
   5. Construct v as subset of V
   6. Split best features in v
   7. Aggregate information from the number of trees for new data prediction
   8. Calculate the accuracy
According to the algorithm above, this algorithm started by input training set, features, and number of
trees. After that, draw bootstrap samples with replacement from data and grow decision trees from
each bootstrap data set. Repeat the step until $N$ times. At each node in the trees are constructed as
subset of features $V$ and split the best features in $v$. After decision trees are grown, aggregate
information from $ntrees$ such as majority voting for classification [13,14]. The result of this algorithm
is the accuracy of the program.

2.3 Confusion Matrix
The confusion matrix is needed to evaluate performance of the method and show rate of data classified
and misclassified in the program. Table 2 show the confusion matrix.

| Class   | Predict | Actual  |
|---------|---------|---------|
|         | Positive| Negative|
| Positive| TP      | FP      |
| Negative| FN      | TN      |

The four condition from confusion matrix table is describe below [15]:
True Positive (TP): number of infarction samples correctly classified as infarction.
False Positive (FP): number of non-infarction samples incorrectly classified as infarction.
False Negative (FN): number of infarction samples incorrectly classified as non-infarction.
True Negative (TN): number of non-infarction samples correctly classified as non-infarction.

According to confusion matrix table, we can evaluated the result of the program by calculate accuracy,
precision, and recall. Accuracy is the average number of samples correctly predicted by classifiers.
Precision is the ratio of actual positive sample that is predicted correctly with the overall positive class
sample. Moreover, Recall is ratio of actual positive sample that is predicted correctly with the overall
actual positive sample. Table 3 show the formula to evaluate the result of the program.

| Parameters | Formula |
|------------|---------|
| Accuracy   | $\frac{TP + TN}{TP + TN + FP + FN} \times 100\%$ |
| Precision  | $\frac{TP}{TP + FP} \times 100\%$ |
| Recall     | $\frac{TP}{TP + FN} \times 100\%$ |

3. Experimental results
The data of 206 infarction patients were divided into multiple five in range 50 to 85 percent as training
data. The result of the random forest using $ntree = 500$ decision trees, 7 features, and class (infarction
and non-infarction) is shown in Table 4.
### Table 4. The result of Random Forest

| Training Data (%) | Accuracy (%) | Recall (%) | Precision (%) |
|-------------------|--------------|------------|---------------|
| 50                | 91.18        | 96.08      | 87.50         |
| 55                | 89.13        | 95.65      | 84.62         |
| 60                | 91.46        | 95.12      | 88.64         |
| 65                | 81.94        | 97.22      | 74.47         |
| 70                | 85.00        | 96.67      | 78.38         |
| 75                | 90.00        | 100        | 83.33         |
| 80                | 95.00        | 100        | 90.91         |
| 85                | **96.67**    | **100**    | **93.75**     |
| 90                | 95.00        | 100        | 90.91         |

According to random forest algorithm’s result, the best value of accuracy, recall, and precision are 96.67 percent, 100 percent, and 93.75 percent, respectively. It is obtained by using 85 percent as training data. Moreover, by using training data is multiple five in range 75 to 90 percent, the value of recall is 100 percent. The average of random forest results are 90.60 percent, 97.86 percent, and 85.83 percent for accuracy, recall, and precision, consecutively. The running time of the random forest is shown in figure below.

![The Running Time](image)

**Figure 1.** The running time of random forest method

Based on Figure 1, The average of running time by using multiple five in range 50 to 90 percent as training data is 3.39 seconds and we can know by using more training data, the running time of the program to show the result will be longer.

### 4. Conclusion

In this paper, Random forest has been used to classify infarction data patient. The best results of the random forest method are using 85 percent as training data. The best value of accuracy, recall and precision are 96.67 percent, 100 percent, and 93.75 percent, consecutively. Moreover, the running of using 85 percent as training data is 4.10 seconds, and the average of running time by using multiple five in range 50 to 90 percent training data is 3.39 seconds.

We hope the next research can develop or modify this method to give the best result for predict or classify the other diseases. Moreover, we hope this paper can help to classify infarction in patient quickly and accurately.
Acknowledgment
This research supported financially by the Ministry of Research, Technology, and Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) with PTUPT 2019 research grant scheme. The authors want to thank Dr. dr. Jacob Pandelaki Sp. Rad(K) and the other people who help the author to finish this paper.

References
[1] Boehme A K, Esenwa C, and Elkind M S V 2018 Stroke Risk Factors, Genetics, and Prevention Circ Res. 120(3) pp. 472–95
[2] Feigin V L, et al. 2014 Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010 Lancet 383(9913) pp. 245–54
[3] World Health Organization 2018 The top 10 causes of death Accessed on: 6 October 2019 https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
[4] Rustam Z, Utami D A, Hidayat R, Pandelaki J, and Nugroho W A 2019 Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infarction Datasets International journal on Advance Science Engineering Information Technology (IJASEIT) 9 (2) pp.685-91
[5] Bagasta A R, Rustam Z, Pandelaki J, and Nugroho W A 2019 Comparison of Cubic SVM with Gaussian SVM: Classification of Infarction for detecting Ischemic Stroke IOP Conf. Ser.: Mater. Sci. Eng. 546 05201
[6] Nadira T and Rustam Z 2018 Classification of Cancer Data Using Support Vector Machines with Features Selection on Global Artificial Bee Colony AIP Conf. Proc. 2023 020205
[7] Aprilliani U and Rustam Z 2018 Osteoarthritis Disease Prediction Based on Random Forest ICACIS Conf. pp 237-40
[8] Rustam Z, Sudarsono E, Sarwinda D 2019 Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD) IOP Conf. Ser.: Mater. Sci. Eng. 546 052066
[9] Huljanah M, Rustam Z, Utama S, and Siswantining T 2019 Feature Selection using Random Forest Classifier for Predicting Prostate Cancer IOP Conf. Ser.: Mater. Sci. Eng. 546 052031
[10] Zahras D, Rustam Z, and Sarwinda D 2019 Soft Tissue Tumor Classification using Stochastic Support Vector Machine IOP Conf. Ser.: Mater. Sci. Eng. 546 052089
[11] Panca V, and Rustam Z 2017 Application of machine learning on brain cancer multiclass classification AIP Conference Proceedings 1862 030133
[12] Hu X, and Zhao Y 2018 Block Ciphers Classification Based on Random Forest J. Phys.: Conf. Ser. 1168 032015
[13] Nguyen C, Wang Y, and Nguyen H 2013 Random Forest Classifier Combined with Feature Selection for Breast Cancer Diagnosis and Prognostic JbiSE, 6 pp 551-60.
[14] Breiman L 2001 Random Forests Accessed on: 20 August 2019 https://www.stat.berkeley.edu/
[15] Rustam Z and Hartini S 2019 Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel IOP Conf. Ser.: Mater. Sci. Eng. 546 052067