Classification Tree Performance Analysis on ICA-based Functional Near-infrared Spectroscopy Signals

Airita Fajarnarita Sumantri[1], Raditiana Patmasari[2], Nur Ibrahim[3]
School of Electrical Engineering,Telkom University
Bandung, Indonesia
[1]airitafs@gmail.com, [2]raditiana@telkomuniversity.ac.id, [3]nuribrahim@telkomuniversity.ac.id

***Abstract*—This paper proposes a method to classify between the clean and the contaminated signal by motion artifact (MA) signal on functional near-infrared spectroscopy (fNIRS) signals, by extracting the signal features based on independent component analysis (ICA) and statistical models using classification tree as the classifier. The extracted features such as kurtosis, skewness, mean, variance, standard deviation, interquartile range, and weight vector are used in a classification tree as the prediction model for class classification. The result of this paper is to acknowledge the performance of classification tree to classify the fNIRS signals, which results in 88.9% accuracy, 81% sensitivity, 100% specificity, and 0.83 value of area under convergence (AUC).**

***Index Terms*—fNIRS, ICA, feature extraction, classification tree**

I. Introduction

FNIRS is one of the brain imaging methods to detect and monitor the brain activity, the detected data in the form of signal usually contains an unwanted signal or noise. Noise can be in various forms, one of the examples is motion artifact that is contained in the fNIRS data which is discussed in this paper. FNIRS data are chosen since it is one of the non-invasive methods which is widely used to detect the brain activity by looking at the oxy-Hemoglobin (oxy-Hb) and deoxy-Hemoglobin (deoxy-Hb) levels [1]. The fNIRS data that is used in this paper is obtained from physionet and contains the data from nine subjects [1]. This paper aims to use ICA as the method to extract the features of the contaminated and the clean signals. Following the obtained features, classification tree is used to process those features to be trained as the parameters for classification. In order to know if classification tree is feasible to use for classification, we analyze the accuracy result and other parameters such as sensitivity, specificity, and AUC.

II. Methodology

This research aims to obtain the parameters to classify the types of fNIRS signals. ICA and statistical methods are used in this paper since these methods is used to extract features which is needed to classify the fNIRS signals classes.

A. ICA

Independent component analysis (ICA) is an algorithm to separate between two or more independent signals from a mixing signal, also means separating a different source signals [2].

$$x = As \tag{1}$$

where x is the observed variables, A is the mixing matrix, and s is the independent components [2].

B. Feature Extraction

In this paper, we are doing feature extraction to classify between two types of fNIRS signals, and since we are using the ICA method that works together with statistical model, therefore the features we are looking for are based on these two models. There are seven features which we are extracting from the data as mentioned below:

1) Kurtosis
   Kurtosis is the outliers occurring in the data, high kurtosis results in the appearance of the outliers, and vice versa for low kurtosis resulting where a lack of outliers occur.

$$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} \tag{2}$$

   where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points [3].
2) Skewness
   Skewness is a parameter to measure the symmetry value of the data.

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3/N}{s^3} \tag{3}$$

   where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points [3].
3) Mean
   Mean is the average value of data obtained from the total value of data divided by the number of data.
4) Variance
   Variance is the measurement of a spread data sets.

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} \tag{4}$$

   where σ^2 is the variance, μ is the mean, and N is the number of data points [4].
5) Standard Deviation
   Standard deviation is the square root of the variance.

6) Interquartile Range (IQR)
IQR value obtained when the data set divided into quartiles, where the average value of data measured in each quartile.

$$IQR = Q3 - Q1 \tag{5}$$

where $Q3$ is the third quartile and $Q1$ is the first quartile.

7) Weight Vector
Weight vector is a variable used as the multiplication vector which gives weight in the process of separating signals from the mixed signal.

C. Classification Tree

Classification tree or decision tree is one of the methods for classification in machine learning. The root-node represents the fNIRS signal, the sub-nodes represents features, the edge node represents the result from feature extraction, and lastly the label represents contaminated (MA) class and clean class [5].

D. Performance Analysis

To achieve the result of this research, we are analyzing these four parameters as the results from classification.

1) Accuracy: Accuracy here means the level of percentage that represents the performance of classification tree in classifying the classes of fNIRS signals, the higher the percentage shows that the classification tree is classifying correctly.

2) Sensitivity and Specificity: Sensitivity in this paper refers to the ability of the method to correctly identify that a data is classified as their class, the true positive rate of the data. On the contrary, specificity is the ability to identify that the data is not classified as their class.

$$sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$specificity = \frac{TN}{TN + FP} \tag{7}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is the false negative, all this values are obtained from the confusion matrix. The confusion matrix is a matrix of prediction and the actual class of classification, from this we can calculate the true and false objects [6].

3) AUC: Area under convergence (AUC) is acquired from the range of convergence (ROC) curve. ROC curve is used to visualize the method performance on classification. The curve represents the true positive rate on the y-axis and false negative rate on the x-axis, both of this value are obtained from the confusion matrix. The perfect AUC value is 1, and it shows that both of the classes are perfectly separated [7].

III. Result and Analysis

In this section, we are analyze the performance parameters which are accuracy, sensitivity, specificity, and AUC. The total data which we classify are 36 data, consists of 18 contaminated data and 18 clean data obtained from physionet.

Accuracy and AUC value we can directly obtain from the computation result, while for sensitivity and specificity we need to calculate the value from equation 6 and 7 with the information obtained from confusion matrix as shown in table I.

TABLE I: Results Performance of Classification

TP	TN	FP	FN
18	14	0	4

TABLE II: Results Performance of Classification

Accuracy	Sensitivity	Specificity	AUC
88.9%	81%	100%	0.83

This table II shows the performance results of classification tree in classifying the classes of contaminated and clean fNIRS signals. The accuracy and sensitivity does not reach 100% since there are four variables that failed to be classified, due to this case, the AUC value also less than 1.

IV. Conclusion

This paper concludes the analysis of the classification tree performance to classify between two classes in fNIRS signals using the features extracted by applying the ICA and statistical models. From the classification process using classification tree, we achieved a high accuracy level that is 88.9%, this value shows that the classification tree is feasible enough for classification.

References

[1] K. T. Sweeney, H. Ayaz, T. E. Ward, M. Izzetoglu, S. F. McLoone, and B. Onaral, "A methodology for validating artifact removal techniques for physiological signals," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 5, pp. 918–926, 2012.

[2] C. Klaussner, "Independent Component Analysis for Feature Extraction Independent Component Analysis (ICA)," 2013.

[3] *. Measures of Skewness and Kurtosis. [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

[4] D. M. Lane. Measures of Variability. [Online]. Available: onlinestatbook.com/2/summarizing_distributions/variability.html

[5] B. L. Aurelian, "An information entropy based splitting criterion better for the Data Mining Decision Tree algorithms," *2018 22nd International Conference on System Theory, Control and Computing (ICSTCC)*, vol. 2, no. 1, pp. 535–540, 2018.

[6] I. Gazalba, N. Gayatri, and I. Reza, "Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification," *International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) Comparative*, pp. 294–298, 2017.

[7] D. M. J. Tax and R. P. W. Duin, "Linear model combining by optimizing the Area under the ROC curve," *The 18th International Conference on Pattern Recognition (ICPR'06)*, pp. 18–21, 2006.