International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 10 (2018) pp. 164-173

Received: 29 Jan, 2018; Accepted: 2 July, 2018; Published: 19 July, 2018

Selection of Performance Measures for Brainprint Authentication

Siaw-Hong Liew[1], *Yun-Huoy Choo[1], Yin Fen Low[2] and Zeratul Izzah Mohd Yusoh[1]

[1] Computational Intelligence and Technologies (CIT) Research Group
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
*siawhong.liew@gmail.com, {*huoy, zeratul}@utem.edu.my*

[2] Machine Learning and Signal Processing (MLSP) Research Group
Center for Telecommunication Research and Innovation (CeTRI)
Faculty of Electronics and Computer Engineering
Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
yinfen@utem.edu.my

***Abstract*: This paper aims to assess the performance of brainprint authentication by using Incremental Fuzzy-Rough Nearest Neighbour (IncFRNN) and Fuzzy-Rough Nearest Neighbour (FRNN) with different performance measures for brainprint authentication modelling. The proposed performance measures are accuracy, area under receiver operating characteristics (ROC) curve (AUC), precision, recall, f-measure, equal error rate (EER) and Cohen's Kappa. The selection of appropriate performance measures is utmost important in evaluating the classification performance especially the dataset with imbalance class distribution. It is to avoid misleading results. With the exception of accuracy measure, the experimental results showed the IncFRNN model achieved better classification results than the FRNN model. It is because of the dataset with imbalance class distribution. The excellent classification performance of FRNN model in accuracy is largely contributed by the true negative rate (TNR). Both the IncFRNN and the FRNN model gained high in accuracy but low in recall. Thus, the evaluation of the imbalance class distribution dataset like brainprint authentication modelling cannot depends only the accuracy measure. Further investigations will be conducted to improve the IncFRNN algorithm for brainprint authentication modelling.**

***Keywords*:** performance measures, brainprint authentication, incremental FRNN (IncFRNN)

I. Introduction

Person authentication using brain signals aims to accept or reject the identity that claimed by a particular individual, which is one-to-one matching. An authentication system is trying to match or compare the presented individual biometric against a biometric profile that already exists in the database. A biometric authentication system must have the seven specific characteristics such as uniqueness, universality, collectability, circumvention, permanence, performance and acceptability. Low intra-subject variability and high inter-subject are the main concern to be as perfect biometric trait.

The brainprint authentication system uses the distinct features that retrieved from the EEG signals to differentiate the user from impostors. The brain signals in the cortex are considered direct representation of the cerebral activities. The good thing in using brainprint authentication is the EEG signals are aliveness and relatively robust in certain situations. The human brain never rests, and the brain injury is very rare occurred. Besides, the EEG signals are unique. Every person has their own thinking towards different tasks. Thus, the EEG signals are almost impossible to be forged. Brainprint authentication is catching researchers' attention recently due to the uniqueness of EEG signals. Apart from that, the convenience of EEG signals acquisition through low-cost consumer grade EEG devices help in promoting the popularity of EEG related applications.

Visual evoked potential (VEP) is the brain electrical activities response to visual stimuli and recorded from the scalp. Most of the research works [1]–[9] on brainprint authentication are focused on VEP due to the signals are particularly strong and relatively clear response [8]. The characteristic of aliveness in EEG signals enhances its security over many commonly accepted authentication systems, such as fingerprint authentication [10]. Fingerprint can be easily forged by silicon and used by the impostor. It is definitely degraded the performance of biometric authentication. Hence, the feature of aliveness is important to prevent imitation as compared to the static physical characteristics. However, the EEG signals are highly uncertain, and easily affected by environmental as well as the electrophysiological noise.

Fuzzy-rough nearest neighbour (FRNN) is a hybridization technique which combines the strength of both fuzzy-rough sets, and the lazy learning from nearest neighbour approach. It is an extension of K-Nearest Neighbour (KNN) algorithm which employs fuzzy-rough set theory. In KNN, the nearest neighbours are calculated using the Euclidean distance method. Meanwhile, the FRNN is used the fuzzy similarity to

calculate the nearest neighbours. FRNN is proven performed well in several domains [11]–[15] due to it follows human decision making approach to solve the real world problems. However, the FRNN performed less promising in solving dynamic changing data such as EEG signals. Incremental learning is an alternative way and proven good in handling the dynamic data. Incremental learning able to learn new knowledge from time to time whenever the data is available. Nevertheless, the current FRNN model is not considered for incremental learning because it lacks of an update function to restructure the existing knowledge granules incrementally [16]–[18].

Several performance measures are used to judge the efficiency of experimental results to comprehend how good of the performance when the model is implemented in the particular domain. The most common used performance measures are accuracy and area under Receiver Operating Characteristics (ROC) curve (AUC). Accuracy measure is used to assess the agreement of a given measurement with the "correct" value of the parameter in a condition. Yet, the accuracy measure provides less meaningful information especially in the dataset with imbalance class distribution. It is because the accuracy measure does not take into account false positives and false negatives. In some cases, false positives can provide some useful information and have a certain tolerance. Therefore, the AUC is used rather than using accuracy measure only. Sensitivity and specificity are used in the AUC and it consider the indices of true positives and false positives [19]. Apart from accuracy and AUC measure, others performance measures such as recall, precision, f-measure, equal error rate (EER) and Cohen's Kappa are used to measure the authentication performance of the proposed models. Thus, the selection of performance measures is utmost important towards the different types of classification tasks like binary class or multi-class.

The rest of the paper is organized as follows: Section II provides the literature reviews on the EEG-based biometric authentication and the use of performance measures to evaluate the authentication performance. Section III outlines the experimentation which includes the data pre-processing steps, classification, and normality test and statistical test. Section IV describes on several types of performance measures such as accuracy, AUC, precision, recall, f-measure, EER and Cohen's Kappa. Section V portrays the results and discussions and finally, section VI draws the conclusion and the direction of future work.

II. Literature Review

EEG signals are noisy and contain large temporal variation between subjects and even within subjects. The reproducibility of an EEG signals for biometric have been conducted in several research [6], [20]. However, they were focus on the stability of EEG signals from session to session within 1 week to 6 months. Biometric authentication system is normally used for long periods and the individual EEG signals change according periods. EEG signals were analyzed on young and elderly adults in resting state and task conditions [21]. The research findings showed that the EEG signals of older subjects displayed smaller fluctuations than young subjects. Besides, no publication was found on the stability of EEG signals in different environments for person authentication purposes. In real world applications, a person usually records the EEG signals for registration in a quiet and focused conditions. However, identity verification attempts of are usually conducted in different locations, different environments and different emotional states [22]. The variation induces noise and uncertainty to the authentication process. Hence, EEG signals analysis becomes a challenging field in knowledge discovery and machine learning.

One solution to overcome this solution is by using incremental learning. The main advantage in using an incremental learning model is it provides the ability to learn new knowledge from time to time whenever it is available. Incremental learning plays vital role for the real-world applications because it is not compulsory to consider a sufficient set of data in the early stage, but the learning process is ongoing from time to time. Incremental learning is a machine learning process with updating the data continuously in the existing training pool. It is adaptable to the change of knowledge granules based on the new learning examples.

Other than the selection of classifiers, the selection of performance measures is also crucial to evaluate the performance of biometric authentication modelling. The dataset for biometric authentication is normally imbalance class distribution especially when the number of users increased. With imbalanced class distribution in the dataset, the accuracy often fails to learn anything useful on the minority class due to the dominant effect from the majority class. For example, a problem with 99% of the data belongs to one class and only 1% of the data belongs to rare class. A classifier can probably achieve 99% of accuracy easily but fail to classify the rare data correctly. From here, we can see that the accuracy measure can produce misleading results on the dataset with imbalance class distribution. Thus, it is utmost important to select the appropriate performance measures when the dataset with imbalanced class distribution in order to avoid misleading results.

In practice, two errors can be commonly found during the authentication task, i.e. the false acceptance and the false rejection. False rejection is falsely rejected the claim from the genuine user while false acceptance is falsely accepted the claim from the impostor as genuine user. Generally, most of the existing biometric research works aim to improve the recognition algorithm and enhance the performance from the perspective of equal error rate (EER), without considering the other evaluation aspects.

Several biometric competitions have been done to examine the recognition rate of the biometric authentication systems. The EER was used in dynamic signature-based competition that organized in the International Conference on Biometric Authentication (ICBA) in 2004 [23]. Other than accuracy measure, the ROC area is also a frequently used to measure the efficiency of the biometric system. Meanwhile, the second frequent used is the Detection Error Trade-off (DET) curve. The DET curve is the plot of false non-match rate (FNMR) versus false match rate (FMR) in a logarithmic chart [24].

Other than that, the performance measures used for the fingerprint-based verification competitions were the AUC,

verification time, the distribution of user and impostor scores, average and maximum template size, average enrolment and failure-to-enroll rate (FTE). The competitions were organized alternate years from 2000 to 2006. The AUC is also used to evaluate face recognition vendor test and iris challenge evaluation that organized by the National Institute of Standards and Technology (NIST) [25]. Furthermore, several performance measures such as the AUC, the distribution of user and impostor scores and the failure-to-acquire rate (FTA) are used for keystroke dynamics algorithms. From the past literature, the AUC and the EER are the most frequently used to assess the performance of the biometric authentication systems.

III. Experimentation

EEG signals classification is challenging because it is changing from time to time, high dimensionality and consist a very low signal-to-noise ratio [26]. Thus, incremental learning approach plays significant role to capture the weaknesses facing by the EEG signals.

A. Data Pre-processing and Data Preparation

EEG signals was collected from a group of 37 healthy subjects, which consists of 18 males and 19 females. Their ages are between 22 to 29 years old in developing the case study for brainprint authentication modelling. The subjects are having normal or corrected normal vision. The ethical approval and the experimental design have been granted by the Medical Research and Ethics Committee (MREC) from Ministry of Health Malaysia.

Every subject is required to read the participant information sheet in understanding the experiment procedures. The subjects were requested to sign the consent form before the EEG signals recording session start. The subject was sat on a back rested chair and provide maximum comfort before the EEG recording start. It is to reduce the possible movements or artifacts during the recording session. The distance between the computer screen and subject's eyes was 1 meter. All the visual stimuli with the size of 700 x 525 pixels and presented on a white background at the center of computer screen. The size of the computer screen is 15.6 inches.

The Inter-Stimulus Interval (ISI) for each trial was set to 1.5 seconds. The picture was remained on the computer screen for 1 second and then followed by 1.5 seconds of white-blank screen as illustrated in Figure 1. Each subject was completed with 120 trials. In total of 120 trials, 60 trials were the selected password picture and the other 60 trials were the pictures randomly selected from the picture set excluding the password picture selected by the subject. The subjects were asked to recognize their selected password picture from a random set of pictures shown on screen. The subjects were asked to think "yes" when their password pictures appeared; and not to perform any action when the password picture was not displayed. The sampling rate used is 256 Hz.

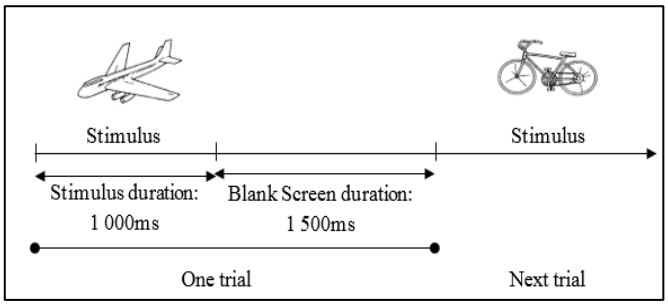

Figure 1. Visual Stimulus Presentation

The experiment was conducted in two different environment conditions: (1) a quiet condition, and (2) noisy condition with induced office noise sound effect played through an audio speaker. The purpose is to simulate a real world environment, that resulted in different elicitation of EEG signals across different individuals. Instead of using all the 64 electrodes, only eight electrodes [7] located at occipital area (as shown in Figure 2) were used to record the VEP signals. It is because the visual cortex is located at the occipital area.

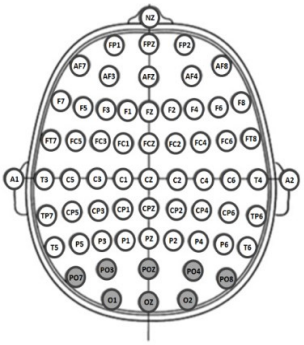

Figure 2. Eight VEP Electrodes Placement

The raw EEG signals are low signal-to-noise ratio and noisy. Thus, the preprocessing steps such as filtration, segmentation and artefact rejection are important before performing further analysis. The recorded EEG signals were filtered in the range of 1 Hz to 30 Hz by using bandpass filtering with Finite-duration Impulse Response (FIR) type. Feature extraction is a compulsory process to retrieve the important features and characteristics from EEG signals. Six feature extraction methods, as described in [27] (i.e. wavelet packet decomposition (WPD), coherence, cross-correlation, mutual information, Hjorth parameter and mean of amplitude) were selected from the literature reviews.

Nevertheless, feature selection is an optional process, but it is important when the feature vectors are large. It is to reduce the dimension of the feature vectors by selecting only the significant feature vectors and eliminates the redundant or useless feature vectors. The WPD method induced large number of feature vectors. In this paper, the important feature vectors from WPD were selected by Correlation-based Feature Selection (CFS) [28]. Only the selected feature vectors will be combined with the other extracted feature vectors.

The pre-processed EEG dataset was arranged into three use cases. It is to evaluate the ability of incremental learning in

handling the non-stationary signals. Thus, the three use cases are described as follows:

Use Case	Description	Aim
Perfect Situation	Data collected in the quiet environment only were used in train and test sets.	Baseline testing
Regular Situation	Data collected in the quiet environment and the noisy environment were used in train and test sets.	Capability testing
Challenging Situation	Data collected in quiet environment and noisy environment were used in test set; but only the data collected in quiet environment were used in train set.	Competence testing

Table 1. Data Arrangement for 3 Different Use Cases.

B. Classification

In this paper, IncFRNN and FRNN techniques were used to perform brainprint authentication modelling. It is a binary class problem with the output class yes or no. Both IncFRNN and FRNN techniques can be found in Waikato Environment for Knowledge Analysis (WEKA) data mining tool. The significant and selected features were split into train and test set by using 10-folds cross validation (CV). It is to prevent the biased evaluation of the classifiers. The designed 10-folds CV here is divided the data into 10% for train set and 90% for test set because the incremental learning able to update training pool from time to time rather than to have a full training data in the early stage of the learning process.

1) Incremental Fuzzy-Rough Nearest Neighbour (IncFRNN) [9], [29]

Incremental Fuzzy-Rough Nearest Neighbour (IncFRNN) [9] is an enhanced version from the original FRNN introduced by Jensen and Cornelis [30] by introducing a heuristic update method and the window size threshold. The main purpose to introduce the update method is that it can update incrementally the knowledge granules by inserting and deleting the object. Therefore, the knowledge granules are keep updated and the object changes over the time while the number of attributes remains the same.

The new object is insert selectively into the existing training pool whenever there is availability of new variant test object. By doing this, the knowledge granules able to capture the new characteristics that represent the individual biometric identity for the authentication process. However, insert the new object continuously results to an increasing on the size of training pool. Consequently, the window size threshold is set to control the size of the training pool for the IncFRNN algorithm.

In IncFRNN algorithm, similarity between the two objects is the main concern in order to delete the object from the training pool. This is because the lower and upper approximation are constructed by the nearest neighbours as described in the FRNN algorithm. The highest value of similarity is used to quantify the class decision for the test object. Hence, the enhancement of the similarity value can further increase the classification results. An object will be deleted if and only if the number of training objects is greater than the window size threshold and the window size threshold is greater than 0. A frequency counter is introduced to track the number of usage for the objects in the nearest neighbour pool. Hence, the IncFRNN algorithm will only deletes the object with the lowest frequency usage and must be within the same class label. Furthermore, the First-In-First-Out (FIFO) strategy is implemented in the IncFRNN algorithm if and only if the counters of the frequency usage for the training objects are same.

In summary, the IncFRNN algorithm preserves all the representative objects and removes the insignificant objects in the training pool. From the perspective of brainprint authentication, the new individual characteristics of EEG signals will be added into the knowledge granules by inserting the object. At the same time, the old and rarely used of the individual EEG signals characteristics will be removed by deleting the object. It is because the characteristics are less meaningful to be used as the identity for the particular individual. In summary, this heuristic update method is vital to obtain better classification results for the performance of brainprint authentication modelling.

2) Fuzzy-Rough Nearest Neighbour (FRNN) [30]

Fuzzy-Rough Nearest Neighbour (FRNN) was introduced by Jensen and Cornelis [30] in 2011, is a hybrid model with the combination of the fuzzy set, rough set and nearest neighbour classification approach. In the FRNN algorithm, the lower and upper approximations are constructed by the nearest neighbours to assign the decision class to the test object. The details of FRNN algorithm can be found in [30]. The FRNN algorithm calculates the similarity between the two objects and finally classify the test object into the most possible decision class. FRNN classifies the test object based on single nearest neighbour with the highest similarity measure. Therefore, the value of k does not affect the classification performance. The FRNN technique captures the uncertainty by using the fuzzy-rough approximations. The construction of the fuzzy lower and upper approximations is to avoid the use of fuzzy logical connectives altogether. The connectives here are very important in developing the fuzzy-rough set theory.

C. Experiment Setting

Various experimental parameters must be set to perform classification. First and foremost, the k-value should be always in odd number and always chosen between 3 and 10 [31]. It is crucial to set the k-value into odd number because the test object can be easily classified [32]. Thus, the optimal value of k had been investigated for the brainprint authentication. In this study, the k-value used for both IncFRNN and FRNN techniques is 5. Other than that, only IncFRNN technique need to set the window size threshold

because of the presence of incremental learning approach. The window size threshold with 0 signifies that unlimited number of training objects.

D. Normality Test and Statistical Test

First and foremost, a normality test, Anderson Darling test is used to test the distribution of the data. It computes the critical values for the specific distribution. The benefit of the Anderson Darling test is that it allows a more sensitive test and the drawback is the critical value must be calculated for each distribution. The Anderson Darling test is calculated as:

$$W_n^2 = n\int_{-\infty}^{\infty}[F_n(x) - F^*(x)]^2\psi\,(F^*(x)dF^*(x) \quad (1)$$

where, $\psi =$ non-negative weight function which can be defined from

$$\psi = F^*(x)(1 - F^*(x))^{-1} \quad (2)$$

The normality test of the data must be carried out before performing a statistical test. A statistical test is performed to determine the confidence level of the dataset that can be in reaching conclusions. Paired sample t-test will be chosen as the parametric test if and only if the data are normally distributed while Wilcoxon signed-rank test will be chosen as the non-parametric test if and only if the data that are not normally distributed. This is a very important point because the statistical test of non-parametric test is less accurate as compared to the parametric test when the data are normally distributed. The statistical tests were tested using IBM SPSS Statistics 22.0.

Paired sample t-test is a statistical test that compares the scores of mean within the same group of two different occasions. In order to calculate the mean of the difference between samples ($\overline{D}$) and the between population means (μ_D), the standard error of the differences ($S_D/\sqrt{N}$) is then takes into account. Thus, the equation is calculated as follow:

$$t = \frac{\overline{D} - \mu_D}{S_D/\sqrt{N}} \quad (3)$$

Wilcoxon signed-rank test is a frequently used for non-parametric test, which is the alternate method for paired sample t-test. It is using the sign and the magnitude of the rank of the differences. In statistical test, the null hypothesis is rejected if and only if the p-value is less than 0.05 and it means that there is significantly different between the two samples. On the contrary, the null hypothesis is accepted if and only if the p-value is larger than 0.05 and it means that there is no significant different between the two samples.

IV. Performance Measurement

Several types of performance measures can be used to evaluate the results of the techniques implemented. Examples of binary class performance measures are accuracy, area under ROC curve (AUC), precision, recall, f-measure, equal error rate (EER) and Cohen's Kappa. In general, a binary classification task takes four possible outcomes and it is known as confusion matrix (as shown in the Table 2). The correctness of a classification can be assessed by computing the number of accurately predicted class examples (true positives (TP)), the number of accurately predicted class examples that do not belong to the class (true negatives (TN)), the number of examples that inaccurately predicted to the class (false positives (FP)) and the number of examples that were not predicted as class examples (false negatives (FN)).

		Actual Class	
		Yes	***No***
Predicted Class	***Yes***	TP	FP
	No	FN	TN

Table 2. Different Outcomes for Binary Class Prediction.

A. Accuracy

Accuracy is widely used to evaluate the performance of classifiers. It is used to measure how good is a binary classification that correctly classified test objects. It is also a measure of the agreement with the correct value of the parameter under certain conditions. However, the accuracy can be misleading when the portions of the class distribution are huge different [33]. The range of accuracy is between 0 and 1; the higher the accuracy value indicates the perfection of the classification results. The accuracy is calculated as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

B. Area Under ROC Curve (AUC)

Area under ROC curve (AUC) is one of the frequent used measures for binary classification, which relies to specificity and sensitivity. AUC encapsulates a single point on the reception operating characteristic curve. It shows how the number of accurately predicted positive examples varies with the number of inaccurately predicted negative examples. As compare to accuracy measure, the AUC is proven to provide more discriminating value and statistically reliable. The AUC performs well and is frequently employed as a general metric of detection performance. ROC analysis had become a standard evaluation for signal processing and medical area. The AUC is calculated by simple trapezoidal integration as:

$$AUC = \sum_n P_{DET}(\tau_n)\Delta P_{FA}(\tau_n) + \frac{1}{2}\Delta P_{DET}(\tau_n) * \Delta P_{FA}(\tau_n) \quad (5)$$

where, $\Delta P_{DET}(\tau_n) = -\big(P_{DET}(\tau_n) - P_{DET}(\tau_{n-1})\big)$ and $\Delta P_{FA}(\tau_n) = \big(P_{FA}(\tau_n) - P_{FA}(\tau_{n-1})\big)$.

The range of the AUC measure is between 0 and 1. The higher the AUC, the better the classification performance. The AUC measure is interpreted as in Table 3.

AUC Measure	**Performance**
0.90 – 1.00	Excellent
0.80 – 0.90	Good
0.70 – 0.80	Fair
0.60 – 0.70	Poor
0.50 – 0.60	Fail
0.00	Incorrectly Classify

Table 3. Interpretation of AUC Measure.

C. Precision

In a binary classification task, the precision denotes the number of examples accurately predicted as belonging to the positive class divided by the total number of examples predicted as belonging to the positive class, which is the summation of TP and FP. Therefore, the precision is calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

D. Recall

In a binary classification task, the recall denotes the number of accurately predicted positive examples divided by the total number of positive examples in the dataset. The recall is also known as true positive rate (TPR). The higher the value of recall, the better the classification performance. The recall is calculated as:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

E. F-Measure

F-measure is the combination of precision and recall and is approximately the average of both the measures when there are close. It is also defined as the weighted harmonic mean. The f-measure is calculated as:

$$F - Measure = \frac{2*(Precision*Recall)}{Precision+Recall} \quad (8)$$

F. Equal Error Rate (EER)

Equal error rate (EER) is also known as crossover error rate, commonly used to assess the performance of biometric authentication system. The EER is used to predetermine the threshold values for its false acceptance rate (FAR) and its false rejection rate (FRR). The EER is obtained from the ROC plot by taking the crossover point where the value of FAR and the FRR are equal. FAR is the percentage of the system incorrectly classified an impostor due to incorrectly matching the biometric input with the existing template. Meanwhile, the FRR is the percentage of the system that incorrectly reject the access to a client due to failing to match the biometric input with the existing template. The lower the value of EER, the higher the reliability of the biometric system. The EER is calculated as:

$$EER = \frac{FAR+FRR}{2} \quad (9)$$

G. Cohen's Kappa

Cohen's Kappa is used to assess the inter-rater reliability for qualitative objects. It is usually believed to be a more powerful measure compared to the simple percentage compliance calculation. Cohen's Kappa is rated in the range from 0 to 1. The greater the value of the Cohen's Kappa, the better the reliability. In general, the performance is rated as satisfactory when the value of Cohen's Kappa is greater than 0.70 [34]. The Cohen's Kappa is calculated as:

$$Cohen's\ Kappa = \frac{P_0-P_e}{1-P_e} \quad (10)$$

where, P_0 is the relative observed agreement among raters, and P_e is the hypothetical probability of chance agreement.

V. Results and Discussion

Table 4 shows the average experimental results for the dataset in perfect, regular and challenging situations respectively. The evaluations of the performance are based on accuracy, AUC, precision, recall, f-measure, EER and Cohen's Kappa. A statistical test was performed to test the significance difference between the two classifiers for different use cases with 95% confidence level. Beforehand, normality test on the classification results must be performed in the earlier stage. It is a crucial step in choosing between parametric and non-parametric test. The experimental results are analyzed in three different perspectives, such as analysis by performance measures, analysis by subjects and analysis by environmental situations.

A. Analysis by Performance Measures

Based on the experimental results, the performance measures of AUC, recall, f-measure, EER and Cohen's Kappa of the IncFRNN model are significantly higher than the FRNN model. On the other hand, only the performance measures of accuracy and precision of the IncFRNN model are significantly lower than the FRNN model.

Use Case	Classifier	Accuracy	AUC	Precision	Recall	F-Measure	EER	Cohen's Kappa
Perfect Situation	IncFRNN	95.08	*0.8843	0.3329	*0.5289	*0.3883	*0.2543	*0.3671
	FRNN	*96.40	0.7918	*0.3893	0.3329	0.3325	0.3428	0.3156
Regular Situation	IncFRNN	94.16	*0.8798	0.2675	*0.5218	*0.3447	*0.2644	*0.3200
	FRNN	*96.22	0.7703	*0.3430	0.2991	0.2980	0.3601	0.2800
Challenging Situation	IncFRNN	94.39	*0.8842	0.2813	*0.5385	*0.3603	*0.2532	*0.3358
	FRNN	*96.30	0.7678	*0.3681	0.3138	0.3179	0.3526	0.3011

*indicates significantly better classification result

Table 4. Experimental Results using IncFRNN and FRNN Classifier.

From the perspective of accuracy measure, the FRNN model achieved 96.40%, 96.22% and 96.30%, which is 1.32%, 2.06% and 1.91% higher than the IncFRNN model for perfect, regular and challenging situations respectively. Even though the accuracy of the FRNN model is higher than the IncFRNN model but it is less reliable due to the data with imbalance class distribution. It is because of the accuracy paradox. When TP is less than FP, the accuracy measure will be increased when the classification rule changes to "negative" output. On the contrary, when TN is less than FN, the accuracy will also increase when the classification rule changes to "positive" output. Thus, the accuracy is not an appropriate performance measure for biometric authentication modelling as the authentication modelling is seriously having the problem of imbalance class distribution. The accuracy measure was affected by the majority class distribution.

Therefore, the evaluation on AUC, precision and recall are more significant than accuracy for the biometric authentication modelling. For the AUC measure, it is not generally relied on the correct prediction ratio only, and yet it considered the specificity and sensitivity. Generally, the AUC is calculated by taking into account the TPR and FPR. The larger value of FPR leads to lower value in AUC. Thus, it is more reliable as compared to accuracy measure. In this experiment, the AUC of the IncFRNN model are 0.8843, 0.8798 and 0.8842 whereas the AUC of FRNN model are only 0.7918, 0.7703 and 0.7678 for the perfect, regular and challenging situations respectively. The AUC measure showed the authentication performance of IncFRNN model is performed better than the FRNN model. Apart from that, when comparing between the dataset in different situations, the authentication performance of IncFRNN also outperformed than FRNN model. The AUC measure of IncFRNN model does not have much different when it applied to different situations. On the other hand, the FRNN model showed huge difference between the perfect, regular and challenging situations. These experimental results showed that IncFRNN model is performed better in handling the non-stationary EEG signals.

Recall and precision are inversely related. In Table 4, the recall of IncFRNN model is higher than the FRNN model while the precision is vice versa. Recall is also known as TPR. Thus, the higher the value of recall, the better the performance of the model. The recall of the IncFRNN model for user class are 0.5289, 0.5218 and 0.5385 for the perfect, regular and challenging situations respectively. While, the FRNN model only achieved 0.3329, 0.2991 and 0.3138 in recall measure, which is a very poor classification result. From the recall measure, we can see that the high accuracy of FRNN model were contributed by the impostor class. It biased to majority class distribution. Besides that, the FRNN model is also unable to handle the uncertainty when the EEG signals were recorded in regular situation. The EEG signals influenced by the simulate environmental noise. Although the IncFRNN model gained higher than the FRNN model in terms of recall, but it is not high enough to be rated as excellent performance.

In terms of EER, the IncFRNN model gained lower value than the FRNN model, which means the IncFRNN model is performed better. The lower the EER, the better the classification performance for the biometric authentication modelling. The lowest EER achieved by IncFRNN model is 0.2532 for the challenging situation while a slightly difference with 0.2543 for the perfect situation. In addition, the IncFRNN model is also performed better than the FRNN model in Cohen's Kappa measure. The IncFRNN model is performed slightly better than the FRNN model in terms of Cohen's Kappa. However, the IncFRNN model is unable to achieve satisfactory performance.

B. Analysis by Subjects

Figure 3 and Figure 4 show the False Acceptance Rate (FAR) of IncFRNN and FRNN model respectively in perfect, regular and challenging situation for 37 subjects. FAR is the measure that the biometric authentication systems incorrectly accept the access attempt by impostors. Thus, the lower the FAR, the better the performance of the biometric authentication systems.

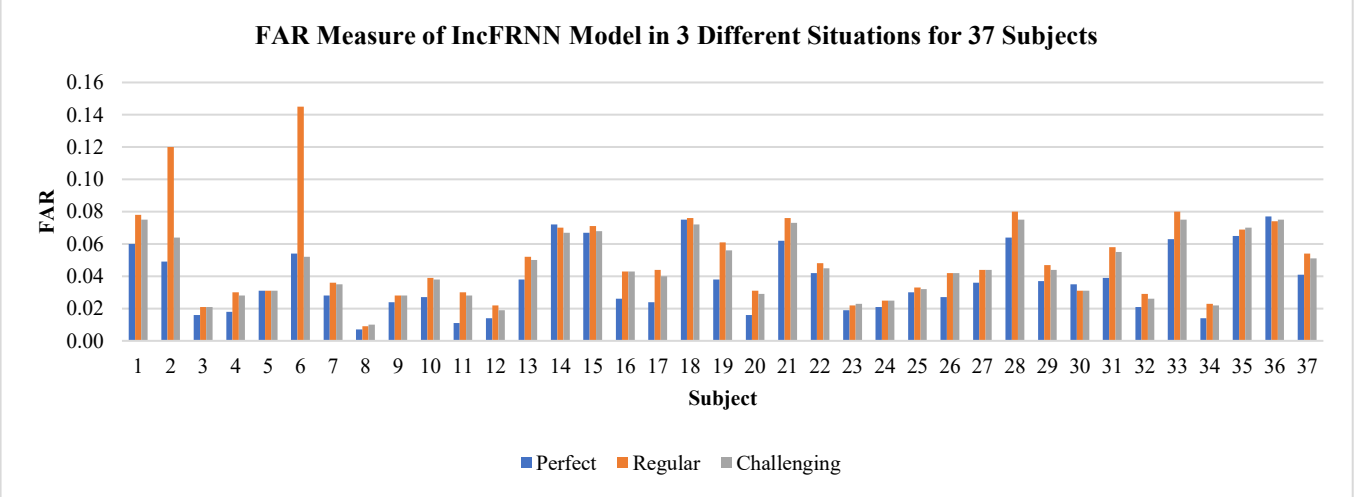

Figure 3. FAR Measure of IncFRNN Model in 3 Different Situations for 37 Subjects

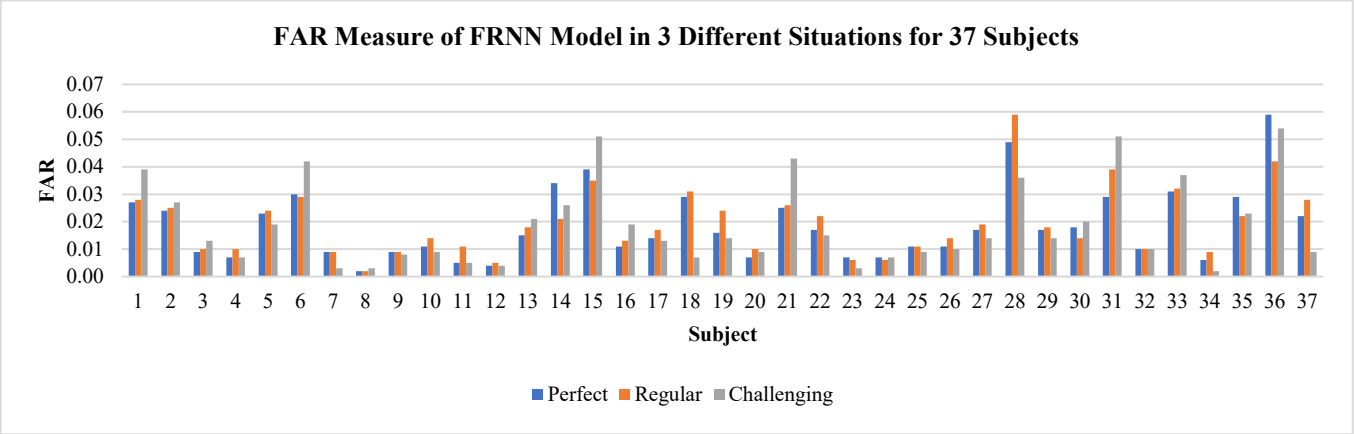

Figure 4. FAR Measure of FRNN Model in 3 Different Situations for 37 Subjects

From Figure 3 and Figure 4, subject 8 performed the best with the lowest FAR. Subject 8 gained 0.007, 0.009 and 0.010 in perfect, regular and challenging situation respectively by using IncFRNN model. Meanwhile, subject 8 only achieved 0.002, 0.002 and 0.003 in the 3 different situations by using FRNN model.

C. Analysis by Environmental Situations

Figure 5 shows the comparison of FAR between IncFRNN and FRNN model in perfect, regular and challenging situations. In overall, the average FAR from 37 subjects are low. Among the 3 situations, both IncFRNN and FRNN model gained the highest FAR in regular situation. This may due to the training objects in noisy environment cannot show the individual characteristics. Nevertheless, the FAR of FRNN model is lower than the IncFRNN model in 3 situations. Thus, we can conclude that the FRNN model performed better than IncFRNN model in term of FAR.

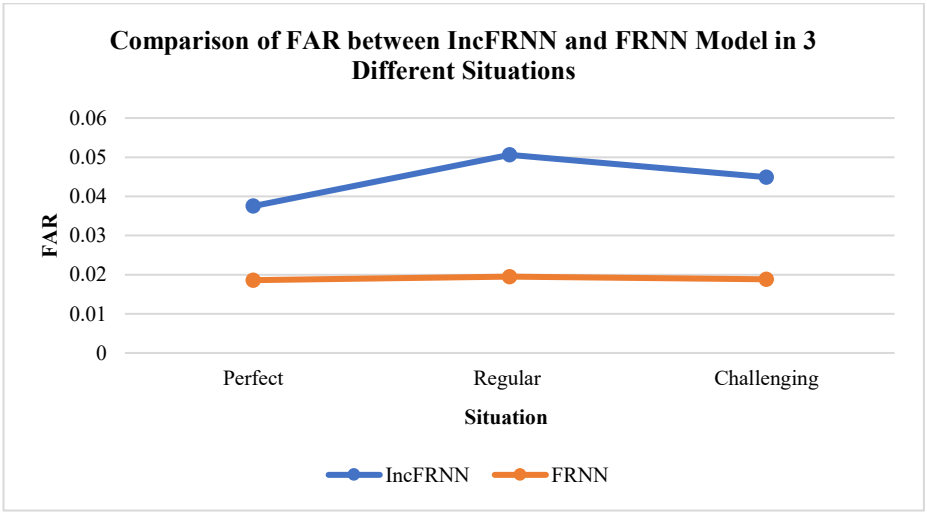

Figure 5. FAR Measure of FRNN Model in 3 Different Situations for 37 Subjects

VI. Conclusion

In this paper, we have highlighted the importance of choosing the appropriate performance measures to evaluate the classification efficiency of the IncFRNN and FRNN model. Among all the performance measures, accuracy should not be considered especially the dataset with imbalance class distribution. Even though the FAR measure of FRNN model is performed better than IncFRNN model, but the recall measure of user class in IncFRNN model is performed better than FRNN model. The recall measure is more important than the FAR measure because it measures the biometric authentication systems that are correctly accept the access attempt by the users. Meanwhile, the AUC measure accesses the overall classification performance for users and impostors class. In overall, the IncFRNN model is worked better than the FRNN model for biometric authentication modelling. But, the recall of the IncFRNN model is still considered low for the datasets in perfect, regular and challenging situations. A reliable performance of machine learning algorithms always has high recall. Besides, the EER is utmost important to evaluate the performance of biometric authentication modelling. However, the EER should be getting lower to get the excellent performance. Further investigation on the IncFRNN algorithm will be carried out to boost the efficiency of IncFRNN model for biometric authentication modelling.

Acknowledgment

The authors would like to express their appreciation to Universiti Teknikal Malaysia Melaka (UTeM) for providing the UTeM Zamalah scheme scholarship. Besides, the authors would also like to thank UTeM for the research facilities support.

References

[1] R. Palaniappan and K. V. R. Ravi, "Improving Visual Evoked Potential Feature Classification for Person Recognition using PCA and Normalization," *Pattern Recognit. Lett.*, vol. 27, pp. 726–733, 2006.

[2] C. N. Gupta, R. Palaniappan, and V. Orschot, *Biometric Paradigm Using Visual Evoked Potential*. Idea Group Inc., 2007.

[3] K. Malinka, "Usability of Visual Evoked Potentials as Behavioral Characteristics for Biometric Authentication," in *2009 Fourth International Conference on Internet Monitoring and Protection*, 2009, pp. 84–89.

[4] A. Zuquete, B. Quintela, and J. P. Silva Cunha, "Biometric Authentication using Brain Responses to Visual Stimuli," in *International Conference on Bio-inspired Systems and Signal Processing*, 2010, pp. 103–112.

[5] S. H. S.-H. Liew, Y. H. Y.-H. Choo, and Y. F. Y. F. Low, "Fuzzy-Rough Nearest Neighbour Classifier for Person Authentication using EEG Signals," in *Proceedings of 2013 International Conference on Fuzzy Theory and Its Application*, 2013, pp. 316–321.

[6] B. C. Armstrong, M. V Ruiz-Blondet, N. Khalifian, K. J. Kurtz, Z. Jin, and S. Laszlo, "Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics," *Neurocomputing*, vol. 166, pp. 59–67, 2015.

[7] S. H. Liew, Y. H. Choo, and Y. F. Low, "Identifying Visual Evoked Potential (VEP) Electrodes Setting for Person Authentication," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, pp. 85–99, 2015.

[8] I. B. Barbosa, K. Vilhelmsen, A. Van Der Meer, V. Der Weel, and T. Theoharis, "EEG Biometrics : On the Use of Occipital Cortex Based Features from Visual Evoked Potentials," in *Norsk Informatikkonferanse (NIK)*, 2015.

[9] S. H. Liew, Y. H. Choo, Z. I. Mohd Yusoh, and Y. F. Low, "Incrementing FRNN Model with Simple Heuristic Update for Brainwaves Person Authentication," in *IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2016, pp. 115–120.

[10] C. He, "Person Authentication Using EEG Brainwave Signals," The University of British Columbia, 2007.

[11] Q. Hu, D. Yu, and Z. Xie, "Neighborhood Classifiers," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 866–876, 2008.

[12] A. E. Hassanien, A. Abraham, J. F. Peters, G. Schaefer, and C. Henry, "Rough Sets and Near Sets in Medical Imaging: A Review," *Inf. Technol. Biomed. IEEE*, vol. 13, no. 6, pp. 955–968, 2009.

[13] T. Boongoen and Q. Shen, "Nearest-Neighbor Guided Evaluation of Data Reliability and its Applications," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 40, no. 6, pp. 1622–1633, 2010.

[14] N. Mac Parthaláin and R. Jensen, "Fuzzy-Rough Approaches for Mammographic Risk Analysis," *Intell. Data Anal.*, vol. 14, no. 2, pp. 225–244, 2010.

[15] P. Maji, "Fuzzy-Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data," *IEEE Trans. Syst. Man, Cybern. Soc.*, vol. 41, no. 1, pp. 222–233, Feb. 2011.

[16] X. Geng and K. Smith-Miles, "Incremental Learning," in *Encyclopedia of Biometrics*, Springer US, 2015, pp. 912–917.

[17] G. Charles Ditzler, "Incremental Learning of Concept Drift from Imbalanced Data," 2011.

[18] J. Read, A. Bifet, B. Pfahringer, and G. Holmes, "Batch-Incremental versus Instance-Incremental Learning in Dynamic and Evolving Data," in *Advances in Intelligent Data Analysis XI*, Springer Berlin Heidelberg, 2012, pp. 313–323.

[19] C. E. Metz, "Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems," *J. Am. Coll. Radiol.*, vol. 3, no. 6, pp. 413–422, 2006.

[20] H. J. Lee, H. S. Kim, and K. S. Park, "A Study on the Reproducibility of Biometric Authentication Based on Electroencephalogram (EEG)," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 13–16.

[21] R. Sleimen-Malkoun, D. Perdikis, V. Müller, J.-L. Blanc, R. Huys, J.-J. Temprado, and V. K. Jirsa, "Brain Dynamics of Aging: Multiscale Variability of EEG Signals at Rest and during an Auditory Oddball Task," *Eneuro*, vol. 2, no. 3, p. ENEURO.0067-14.2015, 2015.

[22] Đ. H. Gia, L. Khải, and Đ. Thị, "EEG Signals For Authentication In Security Systems," *J. Sci. Technol. Inf. Secur.*, no. 03, pp. 17–32, 2016.

[23] "International Conference on Biometric Authentication," 2004. [Online]. Available: http://www4.comp.polyu.edu.hk/~icba/icba2004/. [Accessed: 28-Dec-2017].

[24] M. Gamassi, M. Lazzaroni, M. Misino, V. Piuri, D. Sana, and F. Scotti, "Accuracy and Performance of Biometric Systems," in *Proceedings of the 21st IEEE Instrumentation and Measurement Technology Conference*, 2004, vol. 1, pp. 510–515.

[25] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "Facial Recognition Vendor Test 2006 and Iris Challenge Evaluation 2006 Large-Scale Results," 2007.

[26] K. Hassani and W. Lee, "An Incremental Framework for Classification of EEG Signals Using Quantum Particle Swarm Optimization," in *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2014, pp. 40–45.

[27] S. H. Liew, Y. H. Choo, and Y. F. Low, "Comparing Features Extraction Methods for Person Authentication Using EEG Signals," in *Pattern Analysis, Intelligent Security and the Internet of Things*, 2015, pp. 225–235.

[28] M. A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," in *Proceeding ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 359–366.

[29] S.-H. Liew, Y.-H. Choo, Y. F. Low, and Z. I. Mohd Yusoh, "EEG-based biometric authentication modelling using incremental fuzzy-rough nearest neighbour technique," *IET Biometrics*, vol. 7, no. 2, pp. 145–152, 2018.

[30] R. Jensen and C. Cornelis, "Fuzzy-Rough Nearest Neighbour Classification and Prediction," *Theor. Comput. Sci.*, vol. 412, no. 42, pp. 5871–5884, Sep. 2011.

[31] M. Chih-Min, Y. Wei-Shui, and C. Bor-Wen, "How the Parameters of K-Nearest Neighbor Algorithm Impact on the Best Classification Accuracy-In case of Parkinson Dataset," *J. Appl. Sci.*, vol. 14, no. 2, pp. 171–176, 2014.

[32] A. B. Hassanat, M. A. Abbadi, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," *Int. J. Comput. Sci. Inf. Secur.*, vol. 12, no. 8, pp. 33–39, 2014.

[33] L. a. Jeni, J. F. Cohn, and F. De La Torre, "Facing Imbalanced Data--Recommendations for the Use of Performance Metrics," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 245–251.

[34] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

Author Biographies

Liew Siaw Hong was born in Kuching, Sarawak, Malaysia. She received her Bachelor degree of Computer Science in Artificial Intelligence in 2014 and Master degree of Science in Information and Communication Technology in 2016 from the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). She is currently pursuing her PhD in the same university. Her research interests involve biometrics, computational intelligence and biomedical signal processing.

Choo Yun Huoy is an associate professor in the Department of Intelligent Computing and Analytics, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM) where she has been a faculty member since 2002. She completed her PhD in Management and Science from National University of Malaysia (UKM) specializing in Data Mining. She obtained her Bachelor degree of Science and Computer with Education majoring in Mathematics in 2000 and Master degree in Information Technology in 2002 at University of Technology Malaysia (UTM). Her research interests involve fundamental studies of rough set theory, fuzzy sets theory, association rules mining, and feature selection, besides the application of data science and data mining in different domains including machine failure analysis in manufacturing for Industry 4.0, person authentication using bio signal, muscle endurance analysis, personalized itinerary and route planning.

Yin Fen Low is an associate professor in Department of Computer Engineering, Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka (UTeM) where she has been a faculty member since 2003. She completed her PhD in EEG signal processing from Saarland University, Germany. She received the Bachelor degree in Electrical Engineering from University of Technology Malaysia (UTM) in 2001 and Master in Engineering Science from University of Multimedia (MMU) in 2003. Her research interests involve biomedical signal processing, Brain Computer Interface (BCI), neurofeedback and machine learning in cognitive science and engineering.

Zeratul Izzah Mohd Yusoh is a senior lecturer in the Department of Intelligent Computing and Analytics, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM) where she has been a faculty member since 2003. She completed her PhD in Queensland University of Technology, Australia. Her PhD work is focusing on SaaS Resource Management System in Cloud Computing using Evolutionary Computation. She did her Master at University of Edinburgh, Scotland and her undergrad study is at Universiti Teknologi Malaysia (UTM), Johor. She is also a Certified IT Professional (Malaysia) and certified EMC Data Science and Big Data Analytics Associate. Her research interests lie in the area of intelligent systems, with a focus on developing soft computing solutions to improve the system's quality. In recent years, she has focused on evolutionary computation techniques, Software as a Service (SaaS) and Cloud computing area.