A COMPARATIVE WORK OF INCREMENTAL LEARNING AND ENSEMBLE LEARNING FOR BRAINPRINT IDENTIFICATION

Siaw-Hong Liew[1], Yun-Huoy Choo[2], Yin Fen Low[3] and Fadilla 'Atyka Nor Rashid[4]
[1]Faculty of Computer Science and Information Technology Universiti Malaysia Sarawak (UNIMAS), Kota Samarahan, Sarawak, Malaysia
[2]Faculty of Information and Communication Technology, Malaysia
[3]Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Melaka, Malaysia
[4]Faculty of Information Science and Technology Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia
E-Mail: shliew@unimas.my

ABSTRACT

Electroencephalogram (EEG) signals are nonstationary and vary across time. The static learning model requires large training data to ensure sufficient knowledge acquisition to build a robust model. However, it is very challenging to achieve complete concept learning due to the behavioural changes in model learning. This issue is particularly critical in brainprint identification, where data acquisition in a short time cannot ensure sufficient training data for comprehensive model learning. Thus, dynamic learning, i.e., incremental learning and ensemble learning, presents a better solution for encapsulating EEG signal changes and variations. Both incremental and ensemble learning follow different approaches to manag the concept learning. Incremental learning merges new variations of EEG signals into the existing learning model over time, while ensemble learning uses multiple models for prediction. Nevertheless, limited research works were reported on comparing these two learning methods to prove the efficiency in handling nonstationary data for brainprint identification. Thus, this paper aims to compare incremental learning and ensemble learning for brainprint identification modelling. Incremental Fuzzy-Rough nearest Neighbour (IncFRNN) and Random Forest are selected to represent incremental learning and ensemble learning, respectively. Accuracy, area under the ROC curve (AUC) and F-measure were used to evaluate the classification performance. The experimental results proved that incremental learning outperformed ensemble learning when the training data were limited. The classification results of IncFRNN model were recorded at 0.9160, 0.9827 and 0.9169 while the Random Forest model only yielded 0.8113, 0.9709, and 0.9169 in accuracy, AUC, and F-measure, respectively. The ongoing learning process in incremental learning helps to capture the new changes in EEG signals and improve the classification performance.

Keywords: incremental learning, ensemble learning, electroencephalogram (EEG) signals, brainprint identification.

Manuscript Received 17 February 2023; Revised 21 July 2023; Published 13 August 2023

1. INTRODUCTION

With the advancement of non-invasive Brain-Computer Interface (BCI), Electroencephalogram (EEG) signals have grown into a popular topic in a variety of fields of study due to their high time resolution, low cost and portability [1]. EEG signals are being used as a biometric trait for authentication and identification and have been highlighted recently. Brainprint identification uses EEG signals to identify an individual among a group of persons who are being evaluated (one-to-N matching). In recent years, brainprint identification is catching researchers' attention [2]-[6] corresponding to the rise of security. EEG signal is private and provides uniqueness. Everyone has diverse mental reactions towards different stimuli. EEG signal is outstanding because it is covered in the brain and physically invisible. Other biometric traits, for example, fingerprint, palm print, or face, are effortlessly reachable by physical sensors on the body surface [7]. These are simply violated and inclined to be imitations by third parties. For example, an artificial fingerprint can be made from silicone, gel, or rubber. However, the EEG signal is difficult to replicate at different locations and at different times.

The main challenges in EEG signals classification are the low signal-to-noise ratio (SNR), and nonstationary characteristics within or between persons, where the EEG signals of the same person vary across time. Correspondingly, our brain is easily affected by emotions, moods, feelings, and other surrounding environmental factors [8]. As a result, a classifier might be trained on a limited amount of training data [9]. Generally speaking, a static learning model requires a large or full amount of training data to ensure sufficient knowledge acquisition to build a robust learning model. Static or traditional approaches will become useless for learning new information. The issue will grow difficult if the previously seen data is no longer available when the new data arrives. It is due to the static or traditional approach required to combine the old and new data to retrain the classifier, which is very impractical. In brainprint identification modelling, it is very challenging to achieve complete concept learning due to emotional and behavioural changes in the model learning. This issue is particularly critical in the case of brainprint identification, where data acquisition in a short time cannot ensure sufficient training data for comprehensive model learning [9]. To address the above-mentioned issues, dynamic learning, i.e., incremental learning, and ensemble learning, presents a better solution in encapsulating the changes and variations in EEG signals. However, limited research works were reported on the comparison between these two learning methods to prove the efficiency in handling nonstationary data for brainprint identification modelling. It is important to identify the right learning approach that can perform well despite the limited training data.

Both incremental and ensemble learning follow different approaches to manage the concept learning. Incremental learning trains the model wisely to conquer the drawbacks of static learning. Incremental learning is an effective dynamic data mining technology that can gain information from the current data more quickly based on prior knowledge from previous data. It tends to modify the existing concept when the variation of knowledge is presented [10], [11]. From the perspective of brainprint identification modelling, new variations of EEG signals were captured in the identification system by incremental learning, which then evolved to what had been learnt from the new examples [12]. In contrast, ensemble learning aims to form a variety of sub-concepts hoping that these sub-concepts can generalize knowledge variation as a whole [13]. On the other hand, ensemble learning utilizes multiple learning algorithms to retrieve additional information from the existing data to improve the prediction [14]. Dynamic ensemble learning divides the data into small data pieces. After that, the classifiers train on each piece of the data independently. Eventually, it generates heuristic rules for combining various classifiers into a single super classifier. As a result, the dynamic ensemble learning model can handle a growing amount of data. Both incremental and ensemble learning offer interesting solutions for brainprint identification.

Incremental Fuzzy-Rough Nearest Neighbour (IncFRNN), K-Nearest Neighbour (KNN), and Incremental Support Vector Machine (SVM) are examples of incremental learning while the Random Forest is an example of ensemble learning. IncFRNN utilizes a heuristic update method to gradually reshape and reform the personalized knowledge granules. It captures the new changes and variations by inserting unseen and representative data. It has been used in [12], [15] for brainprint authentication, and achieved good classification results. The accuracy and AUC for the brainprint authentication were 95.08% and 0.8843, respectively [12]. The classification results are much better than incremental KNN. It might be due to the imbalanced dataset and the First-In-First-Out (FIFO) update strategy in the KNN algorithm. Incremental SVM is slightly more complex than the KNN. The classifier's training time in large datasets is affected by the parameter selection and the algorithmic complexity. Incremental SVM was used to classify the discrimination of movement imagery EEG signals in Brain-Computer Interface (BCI) to cope with EEG dynamic variations [16]. It also implemented the FIFO update strategy to remove certain historical examples consecutively and replenish certain recently acquired new objects. This experiment achieved 80% in terms of the classification rate.

On the other hand, Random Forest includes and utilizes many or all the features to develop multiple decision trees. Thus, it will help to limit the errors due to bias and variance. Random Forest was used to identify the EEG data from a group of 40 participants for the human mental state [17]. The classification accuracy achieved 75%. Besides that, Random Forest and SVM have been used for brainprint identification in eyes-open and eyes-closed conditions [18]. The experiment proved that the Random Forest performed better than SVM. The identification rates in the eyes open condition were recorded at 98.16% for Random Forest and 97.64% for SVM. For the eyes closed condition, the identification rates were recorded at 97.30% and 96.02% for Random Forest and SVM, respectively.

This paper presents a few sections as follows: Section 2 illustrates the experimental materials and methods, including data acquisition, data pre-processing, feature extraction, and feature selection. Meanwhile, Section 3 describes the machine learning approach, such as incremental learning and ensemble learning. Besides, Section 4 depicts the experimental results and validation tests with discussion. Section 5 concludes the overall works in the conclusion term and hints at the future work direction.

2. MATERIALS AND METHODS

In this study, the brainprint identification modelling comprises four main steps (as shown in Figure-1). The experimentation starts with the data acquisition, then followed by the data pre-processing step. It is an essential step to avoid misleading the acquired data. Next, the feature extraction stage is required to retrieve the unique characteristics and meaningful data from the EEG signals to represent an individual. Without meaningful data, a machine learning model will be impractical. Besides that, feature selection is also required to minimize the dimension of feature vectors before fitting into the classifiers. The classification labels are according to the number of subjects.

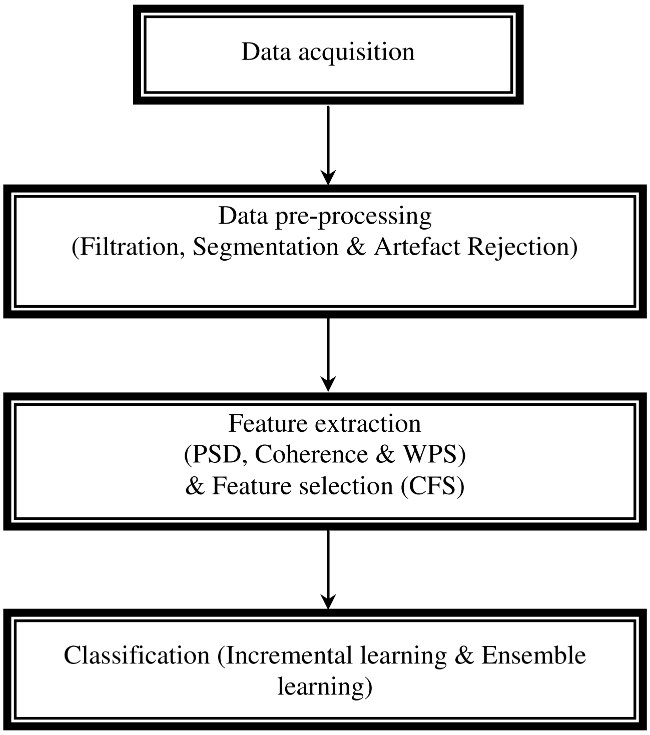
Data acquisition
Data pre-processing
(Filtration, Segmentation & Artefact Rejection)
Feature extraction
(PSD, Coherence & WPS)
& Feature selection (CFS)
Classification (Incremental learning & Ensemble learning)

Figure-1. Process of the experiment.

2.1 Data Acquisition

A new EEG dataset is gathered from 10 healthy participants (7 males and 3 females), ranging from 26 to 32 years old. The data acquisition was carried out at Universiti Teknikal Malaysia Melaka (UTeM) in a controlled quiet research laboratory to maintain the participant's focus and reduce the signal noise. The experimental design and ethics were endorsed by the Medical Research and Ethics Committee (MREC) from the Ministry of Health Malaysia. Before their involvement, the participants were granted an informed consent form before starting the experiment. All the participants are right-handed and had normal vision or rectified normal vision.

Each participant is compulsory to understand and follow the experimental procedures stated in the participant information sheets. The participant is sitting on a reclining armrest chair to provide a comfortable condition during EEG signal recording. It is to avoid any possible artefacts during the recording process, as the artefacts might cause the data to be misinterpreted.

Each participant is required to recognize the selected password image and click on the mouse as soon as the password image is displayed. For each session, a total of 150 trials were recorded, including 60 trials of the selected password image, and 90 trials with a random image from 260 black and white images prohibited the password image. Next, the 150 trials were shown arbitrarily to the participant. The visual stimuli are presented in the computer screen centre with a white background and a resolution of 700 x 525 pixels. The image will be shown for 1 second only, followed by 1.5 seconds for Inter-Stimulus Interval (ISI). The experimental paradigm is interpreted in Figure-2.

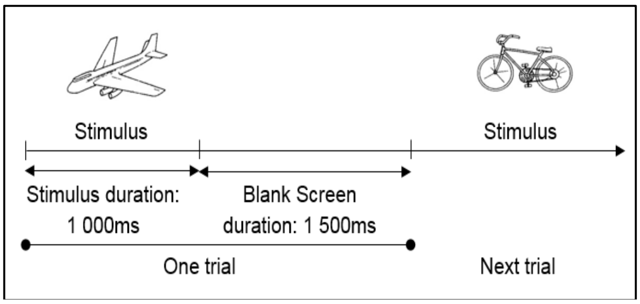

Figure-2. EEG data acquisition experiment paradigm.

The EEG signal acquisition is done by a non-invasive method using Twente Medical Systems International (TMSi) porti system (as shown in Figure-3). It is a multifunctional 32-channel stationary and ambulatory system equipped with both unipolar and bipolar electrophysical inputs designed for physiological research. TMSi porti system comes with water-based electrodes and is convenient to use. The shielded cables for EEG electrodes and ground electrodes were used to achieve low impedance (< 1 kΩ) [19].

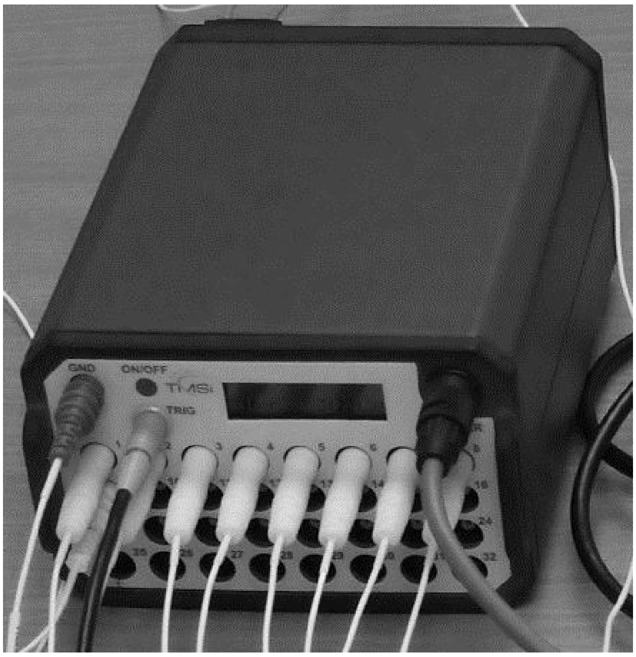

Figure-3. TMSi porti system.

The most commonly used system for research purposes is the International 10-20 electrodes positioning, which comprises 21 electrodes (as shown in Figure-4). All the electrodes in the experiment were referred to the right earlobe (A2) and grounded on the participant's right hand. The sample rate was set to 512 Hz. To reduce the modelling complexity, only five electrodes (O1, OZ, O2, T5, and T6) were chosen and used in this experiment. They are selected based on their importance in visual and audio tasks [20]. The occipital and the temporal electrodes are the dominant electrodes for visual and auditory respectively.

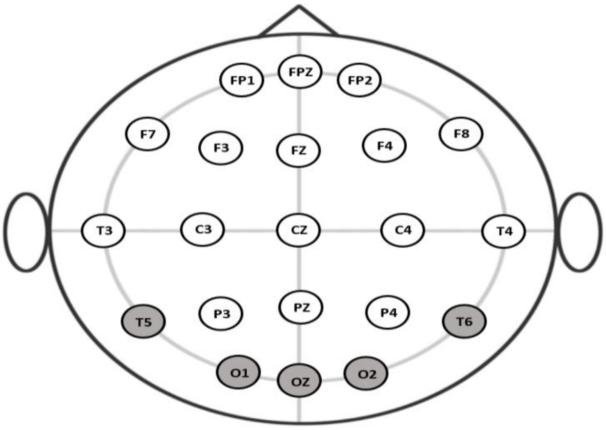

Figure-4. International 10-20 electrode placements.

2.2 Data Preprocessing and Data Preparation

Filtering, segmentation, and artefact rejection are the three important processes in stimulus-locked EEG data. A bandpass filter, Finite-duration Impulse Response (FIR) is used to filter the acquired EEG signals with the cut-off frequencies of 8-12 Hz to obtain alpha band signals. Furthermore, artefact rejection is used to abolish the EEG signals responses caused by excessive body motions or other artefacts with an amplitude of more than 100 µV.

2.3 Feature Extraction and Feature Selection

Feature extraction is a necessary procedure to extract the representative characteristics from EEG signals in achieving robust classification results. Power Spectral Density (PSD), coherence and Wavelet Phase Stability (WPS) were used in this study. PSD is defined as the distribution of signal strength in the frequency domain [6]. It retrieves the correlation information between the measured signals from several electrode channels [21]. The PSD is calculated as follows:

$$P_x(k) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n)e^{-j\left(\frac{2\pi}{N}\right)nk}\right|^2 \qquad (1)$$

where, $k = 0, 1, 2, \ldots, N-1$.

Besides that, WPS is also good at extracting meaningful information from EEG signals. It uses wavelet-based measures to compute the phase information. In signal processing, the phase implies more meaningful information than the amplitude [22]. The formula of WPS is defined as follows:

$$\Gamma_{s,\tau}^{m}(\mathcal{F}) = \frac{1}{m}\left|\sum_{n=1}^{m} e^{larg\left(\left(\mathcal{W}_{\psi}f_m\right)(s,\tau)\right)}\right| \qquad (2)$$

where, $m = 1, \ldots, M$ and $\Gamma_{s,\tau}^{m}(\mathcal{F})$ measures the mean of the degree of clustering of the angular distribution for certain s and τ for M trials.

Furthermore, coherence is also widely applied in brainprint identification model. It utilizes the degree of linear correlation between two signals [23]. The coherence in the EEG signals provides an important estimation of functional interactions between the neural systems operating in each frequency band [24]. The value of coherence is ranging from 0 to 1. The higher the value of the coherence, the higher the linear dependence between the two signals. The expression of coherence is given as follows:

$$C_{xy}(f) = \frac{\left|P_{xy}(f)\right|^2}{P_{xx}(f)P_{yy}(f)} \qquad (3)$$

After extracting the features, the Correlation-based Feature Selection (CFS) method is used to reduce the size of feature vectors without degrading the classification performance [25]. CFS could be a basic and correlated-based filter algorithm that applies to discrete and continuous problems [25]. The CFS algorithm assesses the feature subset based on the correlation-based heuristic merit. It determines the feature's effectiveness through the inter-correlation between the features. The equation used to filter out the redundant, irrelevant, and noisy features is expressed as follows:

$$F_s = \frac{k\overline{r_{zn}}}{\sqrt{k+k(k-1)\overline{r_{nn}}}} \qquad (4)$$

where F_s is the evaluation of a subset of S consisting of k features, $\overline{r_{zn}}$ is the average correlation value between features and class labels, and $\overline{r_{nn}}$ is the average correlation value between two features.

2.4 Performance Measures and Validation Test

In this study, accuracy, area under Receiver Operating Characteristics (ROC) curve (AUC), and f-measure are used to evaluate the classification efficiency for the brainprint identification modelling. Accuracy is the calculation used to decide which model is the most excellent at recognizing relationships and trends between variables in the training dataset. The better a model can generalize to testing data, the better predictions and insights it can produce. The accuracy is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (5)$$

where TP is a true positive, TN is a true negative, FP is a false positive and FN is false negatives.

AUC is also frequently used in evaluating classification performance. It captures a single point on the reception operating characteristic curve. The higher the value of AUC, the better the classification performance. The AUC is calculated by simple trapezoidal integration:

$$A\mathrm{UC} = \sum\nolimits_{\mathrm{n}} \mathrm{P}_{\mathrm{DET}}(\tau_{\mathrm{n}})\Delta\mathrm{P}_{\mathrm{FA}}(\tau_{\mathrm{n}}) + \frac{1}{2}\Delta\mathrm{P}_{\mathrm{DET}}(\tau_{\mathrm{n}}) * \Delta\mathrm{P}_{\mathrm{FA}}(\tau_{\mathrm{n}}) \qquad (6)$$

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

In addition, 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:

$$F - Measure = \frac{2*(Precision*Recall)}{Precision+Recall} \quad (7)$$

where, $recall = TP/(TP+FN)$ and $precision = TP/(TP+FP)$.

Before performing the validation test, an Anderson-Darling test was performed in MATLAB to check the normality distribution of the results. The Anderson-Darling test computes the key values for a specific distribution. It appears to be more viable in finding variations in the distribution tail. The distribution tail is very crucial, especially in the analysis of capability. Next, paired sample t-test, was used to test the significant difference in the comparison of approaches.

3. MACHINE LEARNING APPROACH

Incremental learning and ensemble learning are two basic ways of learning from dynamic data or big stream data. Incremental learning fits a machine learning model where the learning process occurs as the new examples appear, and after that evolves to what has been learned from the unused cases. The goal of incremental learning is for the learning model to adapt to new data without neglecting its existing knowledge. In contrast, ensemble learning utilizes numerous base learners and merges their predictions. The fundamental concept of dynamic ensemble learning is splitting large data into small data pieces and training the classifiers independently on each piece of data. The key difference between incremental learning and ensemble learning is that ensemble learning may dispose of outdated training data, whereas incremental learning may not [14]. In this section, Incremental Fuzzy-Rough Nearest Neighbour (IncFRNN) is used for incremental learning, and Random Forest is used for ensemble learning.

3.1 Incremental Learning

Incremental Fuzzy-Rough Nearest Neighbour (IncFRNN) is one of the incremental learning methods. IncFRNN was introduced by Liew *et al.* [12], [15], by employing a heuristic update method to revise the personalized knowledge granules incrementally. The heuristic update method implements the variation of an object, the addition of an object and the deletion of an object for the incremental update strategy.

In the IncFRNN algorithm, the new object is added selectively to the training pool when the learning model meets with test objects. The insertion of an object occurs when the test object is inaccurately classified. The principle concept of this condition is that the current knowledge granules are unable to predict the new test object. Hence, this modification will help the identification process if the model meets with another similar test object in the future.

However, continuously inserting the object would increase the number of training data and lead to high computational complexity. Thus, the IncFRNN algorithm updates the training pool selectively. Instead of inserting all the new test objects, the IncFRNN algorithm selects the representative test objects added to the training pool. Besides that, the IncFRNN algorithm limits the number of training objects by using the window size threshold. It is an optional process. The deletion of an object is executed when the number of training objects is larger than the window size threshold. In the IncFRNN algorithm, a frequency counter is introduced to track the usage frequency for every nearest neighbours. This will delete the training object with the lowest count. Moreover, the IncFRNN algorithm pursues the First-In-First-Out (FIFO) strategy, where the frequency counter for the training objects is equal. In summary, the incremental update strategies in the IncFRNN algorithm would keep all the distinct objects while abolishing the trivial objects.

IncFRNN algorithm calculates the nearest neighbours by using similarity, as shown in Equation (8).

$$R_a(m,n) = 1 - \frac{|a(m)-a(n)|}{|a_{max}-a_{min}|} \quad (8)$$

where, $R_a(m,n)$ is the degree to which objects m and n are similar in attribute a, a_{max} and a_{min} are the maximal and minimal occurring values of that attribute.

3.2 Ensemble Learning

Random Forest is one of the ensemble learning methods. Random Forest creates multiple decision trees during the training phase and outputs the average prediction of individual trees [17]. It is an expansion of bagging for decision trees. The fundamental concept of the bagging method is that a mixture of learning models would improve the overall classification results. However, a drawback of bagged decision trees is that the decision trees are built using a greedy algorithm that chooses the best-split point in the tree-building process at every step [26]. Therefore, the resulting trees eventually look very comparative, which diminishes the fluctuation of predictions from all the bags.

The advantage of the Random Forest is that it is straightforward to compute the relative significance of each feature on the prediction. Random Forest generates forests with a random number of trees. The aggregation of many decision trees can decrease the chances of overfitting and therefore helps to yield useful results. The decision tree classifier uses information gain (refer to Equation (9)) and Gini index (refer to Equation (11)) to split the criteria [27]. Given a training set, $D = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, x_i is a K-dimensional feature vector of the ith sample and y_i is the label of x_i. The whole training set D is used to train a decision tree. A is the number of classes in the initial training set while kth dimensional feature has M possible values. The proportion of the jth class data in D is represented by p_j. D^m is the set of samples with the mth possible value in feature k. Thus, the information gain and Gini Index are calculated as follows:

$$\mathrm{Gain(D,k)} = \mathrm{Entropy(D)} - \sum_{m=1}^{M} \frac{|D^m|}{D} \mathrm{Entropy}(D^m) \quad (9)$$

$$Entropy(D) = -\sum_{j=1}^{A} p_j \, log \, p_j \quad (10)$$

$$GiniIndex(D,k) = \sum_{m=1}^{M} \frac{|D^m|}{D} Gini(D^m) \quad (11)$$

$$Gini(D) = -\sum_{j=1}^{A} p_j(1-p_j) \quad (12)$$

4. RESULTS AND DISCUSSIONS

This section presents and discusses the classification performance of incremental learning and ensemble learning for brainprint identification modelling. There are a few experimental parameters that are required to be set before performing the classification. Firstly, the value of k is set to 5. The value of k should be always an odd number to ease the classification [28]. In this study, the value of the window size threshold is set to 0, which denotes an unlimited number of training objects. The classification performance is assessed based on accuracy, area under the ROC curve (AUC), and F-measure. The data was split into three sets: 10% for training data and 90% for testing data, 25% for training data and 75% for testing data, 50% for training data and 50% for testing data. The goal is to investigate the classification performance between incremental learning and ensemble learning with different amounts of training data. Moreover, a statistical test was performed to test the significant difference between incremental learning and ensemble learning with 95% confidence level.

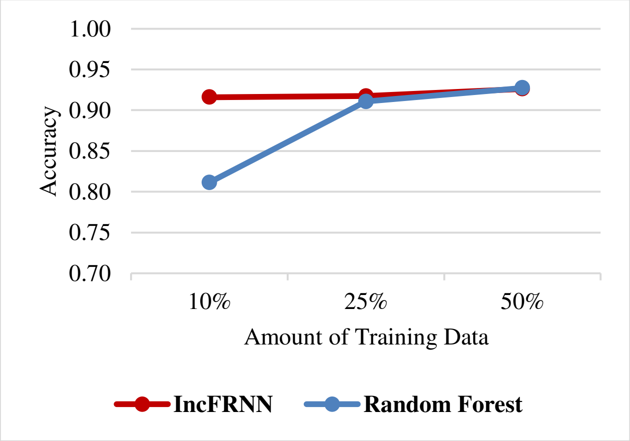

Figure-5. Comparison of accuracy between IncFRNN and Random Forest with different amounts of training data.

Based on Figure-5, we can observe that the IncFRNN technique maintained a good classification accuracy from a small training dataset to a large training dataset. Meanwhile, Random Forest showed a huge difference when using a limited amount of training data. The classification performance has been further improved when more data was used for the training phase. With the 10% of training data, the accuracy achieved by the IncFRNN technique was 0.9160. The accuracy was slightly increased to 0.9174 when the amount of training dataset increased to 25%. An accuracy of 0.9261 was yielded by the IncFRNN technique when 50% of the data was used for training.

On the other hand, Random Forest gained 0.8113 only when using 10% of training data. The accuracy increased by 12.25% when the amount of training data increased to 25%. The accuracy of Random Forest increased slightly to 0.9273 when using 50% of the training data. The experiment results show that incremental learning can perform well, even though with limited training data.

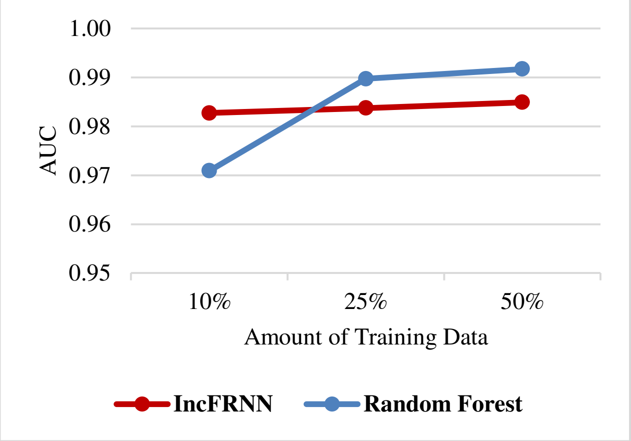

Figure-6. Comparison of AUC between IncFRNN and Random Forest with different amounts of training data.

Based on Figure-6, we can observe that the overall classification performance in terms of AUC was excellent, with a value of AUC of more than 0.95. The IncFRNN showed minor differences from small amounts to large amounts of training data. The IncFRNN achieved the AUC of 0.9827, 0.9837, and 0.9849 when using 10%, 25%, and 50% of the training data, respectively. Nevertheless, the AUC of Random Forest showed a huge difference when the training data size increased from 10% to 25%. With 10% of training data, the AUC only recorded at 0.9709; then, it increased to 0.9897 when using 25% of training data. It was higher than the AUC of IncFRNN. In addition, the Random Forest also performed better than IncFRNN when using 50% of the training data. The AUC of Random Forest was getting higher when the size of the training data increased. Apart from AUC, the classification performance is further analyzed in F-measure.

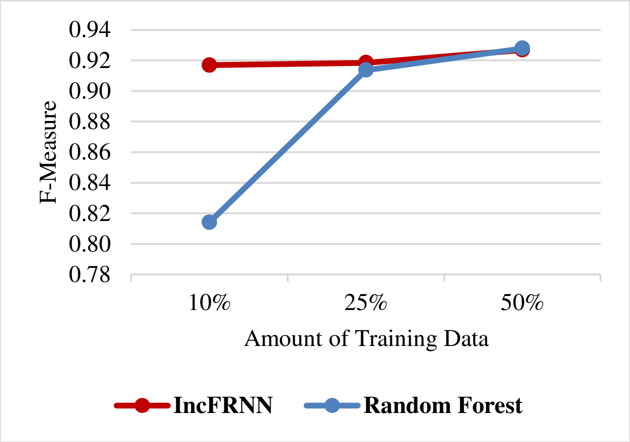

Figure-7. Comparison of F-measure between IncFRNN and Random Forest with different amounts of training data.

Based on Figure-7, we can observe that the overall classification performance of the IncFRNN classifier was higher than the Random Forest classifier in terms of F-measure. Conceptually, F-measure is not as simple to understand as accuracy, but F-measure is typically more useful than accuracy. Inspecting the results, the F-measure of IncFRNN was recorded at 0.9169, 0.9184, and 0.9268 when using 10%, 25%, and 50% of training data, respectively. On the other hand, the F-measure of Random Forest gained 0.8141 when using 10% of the training data. The F-measure of Random Forest increased by 12.22%, with the value 0.9136 when using 25% of the training data. The Random Forest showed a slight increment when using 50% of the training data, which was recorded as 0.9278. The result shows that the Random Forest worked well if the amount of training data is sufficient. However, it is troublesome to get sufficient training data during the initial stage of brainprint identification modelling.

Table-1. Validation tests for the comparison between IncFRNN and random forest with different amount of training data.

Performance Measure	Classifier	10%	*p*-value	25%	*p*-value	50%	*p*-value
Accuracy	IncFRNN	0.9160	0.001	0.9174	0.271	0.9261	0.842
	Random Forest	0.8113		0.9107		0.9273	
AUC	IncFRNN	0.9827	0.014	0.9837	0.014	0.9849	0.002
	Random Forest	0.9709		0.9897		0.9917	
F-Measure	IncFRNN	0.9169	0.000	0.9184	0.525	0.9268	0.897
	Random Forest	0.8141		0.9136		0.9278	

The classification results are further analyzed with validation tests. Paired sample t-test was carried out to test the significant difference between incremental learning and ensemble learning with different amounts of training data for brainprint identification modelling. Based on Table-1, all the comparisons between IncFRNN and Random Forest showed a significant difference when using 10% of the training data. The p-value of the comparison between accuracy, AUC, and F-measure were 0.001, 0.014, and 0.000, respectively. With 10% of training data, the classification performance of IncFRNN was significantly better than the Random Forest.

With 25% of the training data, only the AUC showed a significantly different between the two learning models. The p-value was recorded at 0.014. Since the AUC of Random Forest was higher than the IncFRNN, thus, it can be concluded that the Random Forest performed better than the IncFRNN. The p-value of accuracy and F-measure were 0.271 and 0.525, respectively, which are greater than 0.05. This means that the classification performance of IncFRNN and Random Forest did not show significantly different. The classification performance of Random Forest improved if the amount of training data increased.

The classification performance is further analyzed with 50% of the training data. Based on Table-1, the overall classification performance of the Random Forest was slightly better than the IncFRNN. Among the three evaluation metrics, only the AUC showed significantly different in the comparison. The p-value was 0.002, which was less than 0.05. By comparing the AUC between the two learning models, it can be concluded that Random Forest performed better than IncFRNN.

This experiment proved that incremental learning is more suitable if the amount of training data is limited. It is because incremental learning can include new examples and revise the knowledge granules according to the changes in EEG signals. In EEG data acquisition, it is difficult to get a full training dataset in the early stage. Thus, it is crucial to update the knowledge granules from time to time. On the other hand, ensemble learning utilizes numerous base learners and merges their predictions to improve predictive performance. Ensemble learning is unable to update knowledge granules incrementally. Thus, ensemble learning can only perform well if the training data are sufficient.

5. CONCLUSIONS

In conclusion, we have studied the classification performance of incremental learning and ensemble learning with different amounts of training and testing data for brainprint identification modelling. The proposed models were tested on the EEG data of 10 participants. Our experimental results and validation tests revealed that incremental learning can often contribute to the highest properties of accuracy, AUC, and F-measure. These findings suggested that incremental learning was more suitable for brainprint identification modelling, especially with a limited amount of training data. As compared to ensemble learning, incremental learning was able to capture new examples and included them in the training pool for future prediction. Hence, personalized knowledge granules are kept updated. In other words, incremental learning is a good dynamic learning model with the ability to update the knowledge granules whenever the new data arrived. Future works can be done by integrating incremental learning and ensemble learning to improve the classification performance for brainprint identification modelling.

ACKNOWLEDGEMENT

The authors would like to express their appreciation to Universiti Teknikal Malaysia Melaka (UTeM) and the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS) for providing the research facilities support and funding.

REFERENCES

[1] H. Banville and T. H. Falk. 2016. Recent Advances and Open Challenges in Hybrid Brain-Computer Interfacing: A Technological Review of Non-Invasive Human Research. Brain Comput. Interfaces. 3(1): 9-46.

[2] S. Zhang, L. Sun, X. Mao, C. Hu, and P. Liu. 2021. Review on EEG-Based Authentication Technology. Comput. Intell. Neurosci., 2021: 1-20, doi: 10.1155/2021/5229576.

[3] O. Landau, R. Puzisand N. Nissim. 2020. Mind your mind: EEG-based brain-computer interfaces and their security in cyberspace. ACM Comput. Surv., 53(1): 1-38, doi: 10.1145/3372043.

[4] J. Ortega, K. Martín-Chinea, J. F. Gómez-González, and E. Pereda. 2020. Biometric Person Authentication Using a Wireless EEG Device. In International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning, pp. 615-620. doi: 10.1007/978-3-030-36778-7_67.

[5] A. Jalaly Bidgoly, H. Jalaly Bidgoly, and Z. Arezoumand. 2020. A survey on methods and challenges in EEG based authentication. Comput. Secur., 93: 1-16, doi: 10.1016/j.cose.2020.101788.

[6] Q. Gui, M. V. Ruiz-Blondet, S. Laszlo and Z. Jin. 2019. A survey on brain biometrics. ACM Comput. Surv., 51(6): 1-38, doi: 10.1145/3230632.

[7] C. Wang, Y. Wang, Y. Chen, H. Liu and J. Liu. 2020. User authentication on mobile devices: Approaches, threats and trends. Comput. Networks, vol. 170, doi: 10.1016/j.comnet.2020.107118.

[8] C. M. Tyng, H. U. Amin, M. N. M. Saad and A. S. Malik. 2017. The influences of emotion on learning and memory. Front. Psychol., 8(AUG): 1-22, doi: 10.3389/fpsyg.2017.01454.

[9] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk and J. Faubert. 2019. Deep learning-based electroencephalography analysis: A systematic review. J. Neural Eng., 16(5): 1-37, doi: 10.1088/1741-2552/ab260c.

[10] X. Wang and Y. Xing. 2019. An online support vector machine for the open-ended environment. Expert Syst. Appl., 120: 72-86, doi: 10.1016/j.eswa.2018.10.027.

[11] J. Hu *et al*. 2021. An Integrated Classification Model for Incremental Learning. Multimed. Tools Appl., 80(11): 17275-17290, doi: 10.1007/s11042-020-10070-w.

[12] S. H. Liew, Y. H. Choo, Y. F. Low and Z. I. Mohd Yusoh. 2018. EEG-based biometric authentication modelling using incremental fuzzy-rough nearest neighbour technique. IET Biometrics, 7(2): 145-152, doi: 10.1049/iet-bmt.2017.0044.

[13] O. Sagi and L. Rokach. 2018. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 8(4): 1-18, doi: 10.1002/widm.1249.

[14] W. Zang, P. Zhang, C. Zhou and L. Guo. 2014. Comparative study between incremental and ensemble learning on data streams: Case study. J. Big Data, 1(5): 1-16, doi: 10.1186/2196-1115-1-5.

[15] S. H. Liew, Y. H. Choo, Z. I. Mohd Yusoh and Y. F. Low. 2016. Incrementing FRNN Model with Simple Heuristic Update for Brainwaves Person Authentication. In IEEE EMBS Conference on

Biomedical Engineering and Sciences (IECBES). pp. 115-120.

[16] X. Zheng, B. Yang, X. Li, P. Zan and Z. Dong. 2010. Classifying EEG Using Incremental Support Vector Machine in BCIs. pp. 604-610.

[17] D. R. Edla, K. Mangalorekar, G. Dhavalikar and S. Dodia. 2018. Classification of EEG data for human mental state analysis using Random Forest Classifier. In Procedia Computer Science, 132: 1523-1532. doi: 10.1016/j.procs.2018.05.116.

[18] B. Kaur and D. Singh. 2017. Neuro signals: A future biomertic approach towards user identification. In Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering, pp. 112-117. doi: 10.1109/CONFLUENCE.2017.7943133.

[19] V. Mihajlovic, G. G. Molina and J. Peuscher. 2011. To What Extent Can Dry and Water-Based EEG Electrodes Replace Conductive Gel Ones? A Steady State Visual Evoked Potential Brain-Computer Interface Case Study. In International Conference on Biomedical Engineering, Venice, Italy. pp. 14-26.

[20] S. A. F. Stehlin, X. P. Nguyen and M. H. Niemz. 2018. EEG with a reduced number of electrodes: Where to detect and how to improve visually, auditory and somatosensory evoked potentials. Biocybern. Biomed. Eng., 38(3): 700-707, doi: 10.1016/j.bbe.2018.06.001.

[21] Z. Y. Ong, A. Saidatul and Z. Ibrahim. 2018. Power Spectral Density Analysis for Human EEG-based Biometric Identification. In 2018 International Conference on Computational Approach in Smart Systems Design and Applications, ICASSDA 2018, pp. 3-8. doi: 10.1109/ICASSDA.2018.8477604.

[22] Y. F. Low and D. J. Strauss. 2011. A performance study of the wavelet-phase stability (WPS) in auditory selective attention. Brain Res. Bull., 86: 110-117, doi: 10.1016/j.brainresbull.2011.06.012.

[23] G. Safont, A. Salazar, A. Soriano and L. Vergara. 2012. Combination of multiple detectors for EEG based biometric identification/authentication. In 2012 IEEE International Carnahan Conference on Security Technology (ICCST). pp. 230-236.

[24] R. Srinivasan, W. R. Winter, J. Ding and P. L. Nunez. 2007. EEG and MEG coherence: Measures of functional connectivity at distinct spatial scales of neocortical dynamics. J. Neurosci. Methods, 166(1): 41-52, doi: 10.1016/j.jneumeth.2007.06.026.

[25] M. A. Hall. 2000. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. In Proceeding ICML ’00 Proceedings of the Seventeenth International Conference on Machine Learning. pp. 359-366.

[26] N. Venkatesan and G. Priya. 2015. A Study of Random Forest Algorithm with implemetation using Weka. Int. J. Innov. Res. Comput. Sci. Eng., 1(6): 156-162, [Online]. Available: www.ioirp.com

[27] C. Hu, Y. Chen, L. Hu and X. Peng. 2018. A novel random forests based class incremental learning method for activity recognition. Pattern Recognit., 78: 277-290, doi: 10.1016/j.patcog.2018.01.025.

[28] A. B. Hassanat, M. A. Abbadi and A. A. Alhasanat. 2014. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. Int. J. Comput. Sci. Inf. Secur. 12(8): 33-39.