*Asia-pacific Journal of Convergent Research Interchange*
Vol.6, No.10, October 31 (2020), pp.41-63
*http://dx.doi.org/10.47116/apjcri.2020.10.04*

# Thick Data Analytics through Ensemble Techniques: Identifying Personalized EEG Biometrics based on Eye State Prediction

Tejas Wadiwala[1], Jinan Fiaidhi[2], Sabah Mohammed[3]

## Abstract

Thick data analytics are being pursued to break the barriers of using the big data predictive analytics for small datasets. The main objective of this paper is to improve the performance of the EEG for biometric authentication using eye blinking brain signals through the use of ensembles techniques. Biometric identification differs largly from the other EEG eye movement analytics applications such as detecting epileptic seizure, identification of stress feature or detecting driving drowsiness as it requires high model rubstness and accuracy. A perfect biometric should be unique, universal and permanent over time. Previous analytical approaches on eye movement failed to show the reliability of the the brain signals to distinguish individuals based on the properties of eye-movements seen as time-signals and for this reason the eye movement have not been considered as a possible solution for a biometric system. This paper's primary focus is on the use of ensemble methods to secure the robustness of the person identification from the EEG eye movement waves. Our approach is a multitier one and it start with training notable binary classification models for biometic identification using eye movement. The training tier is followed by ensemble learning (boosting, bagging, and stacking algorithms) to narrow the differences of accuracy gap among classifiers. The classifier's robustness has been measured with the help of variety of accuracy measures including the Matthews correlation coefficient (MCC). The third tier is guage the person prediction model stability using the AUROC (Area Under the Receiver Operating Characteristics) metric. The results obtained in this study proves that it is possible to use an eye tracking based biometric for detection of person identity with reasonably high sensitivity and specificity.

Keywords: EEG Brain Waves, EEG Eye Movement Tracking, Biometrics, Ensemble Learning, Thick Data Analytics

## 1. Introduction

Cryptographic frameworks presented several challegenes and issues related to authentication of users over the networks as most of these mechanisms attempt to use some kind of key sharing

Received(May 10, 2020), Review Result(1st: June 27, 2020, 2nd: August 13, 2020), Accepted(September 25, 2020)

1) (Professor) Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
email: tejas@lakeheadu.ca
2) (Professor) Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
email: jfiaidhi@lakeheadu.ca
3) (Professor, Corresponding Author) Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
email: mohammed@lakeheadu.ca

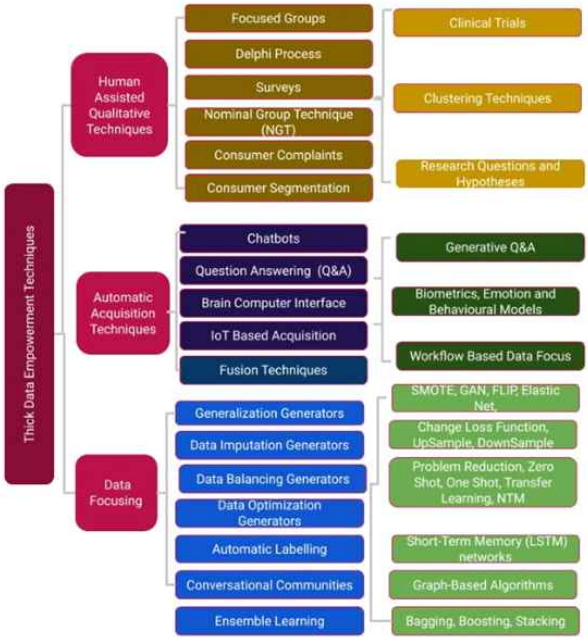
*ISSN: 2508-9080 APJCRI*
*Copyright ⓒ 2020 FuCoS*

for ensuring security of the data. However, biometric authentication is presenting rather simpler model for people identification using some kind of biological markers uniquely associated with people such as fingerprint and eye iris patterns which are mainly based on the physical characteristics of a person. However, traditional biometrics considers the static credentials is a person's body in which it does not matter if the person is unconscious. Therefore, such static biometrics poses a threat to security for identifying living persons correctly. There has been, however, a new trend arising in recent years in the biometric identification research domain, which considers the dynamic characteristics and that are capable of utilizing data directly resulting from the individual's brain activity. Behavioural features are what these components are called; the main ones include gait, speech, keystroke dynamics, and most newly eye movement characteristics and features using signals acquired from EEG, including eye state detection[1]. It isn't easy and technically not possible to correctly replicate such complicated behavioural human characteristics for the primary purpose of forging it. Since the eye movement characteristics are mainly produced with the help of brain waves, it becomes a more robust approach that can be used to identify a person, or it can be used for security purposes.

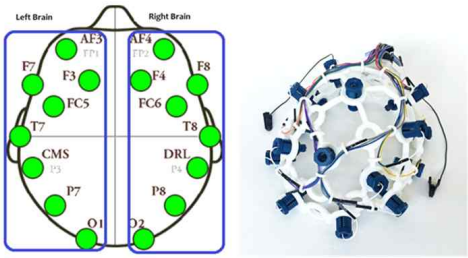
Thick data analytics, however, is empowering big data and machine learning at scale with more qualitative informative materials, tools or techniques that help investigators gather granular, specific knowledge about their target users and learning goals. Thick data can be used to reveal people's underlying stories, emotions and behavioural models by using ethnographic, anthropological, qualitative, conversational communities and practices as well as authomatic data labeling methods[2-4]. [Fig. 1] illustrate some of the notable thick data analytics techniques.

With the advent of brain-computer interface technology (BCI), signals from brain waves can be recorded. There are three types of BCI devices: non-invasive, partially-invasive and invasive. These devices capture EEG recordings in various EEG frequencies; the raw recordings are then converted into distinct waves having different frequencies using a method known as Fast Fourier Transform. Brainwaves are categorized into four main types: beta waves (frequency: 14Hz to 30Hz), alpha waves (frequency: 7Hz to 13Hz), theta waves (frequency: 4Hz to 7Hz), delta waves (frequency range up to 4Hz). There is an increased need to accurately predicting the acquired signals for a particular output. Measuring these waves is not that complicated as brain signal sensors like the Cyton Biosensing Board, OpenBCI Scalp and OpenBCI UI available from the OpenBCI.com. [Fig. 2] illustrate the OpenBCI tools for measuring brain waives.

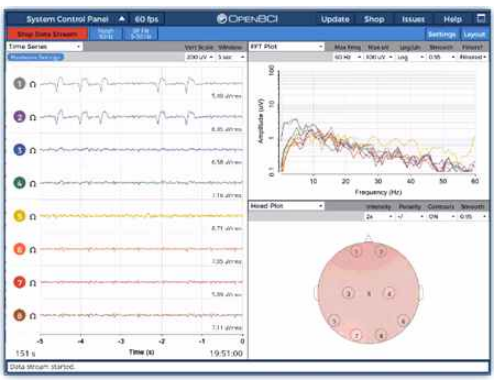

*Copyright ⓒ 2020 FuCoS*




[Fig. 1] Thick Data Analytics Technques




(a) OpenBCI Electrorodes and Scalp




(b) OpenBCI User Interface

[Fig. 2] OpenBCI Scalop and User Interface


*Copyright ⓒ 2020 FuCoS*

Eyes are one of the essential symbolic sensory organs. Through the eyes, it is possible for a human being to interact with the environment. The other qualities which the eyes are commonly associated with are intelligence, vigilance, light, truth and moral conscience. Eye movements like open and closed eyes open a new door for human identification. These movements are in the form of signals and are captured through the help of Emotiv EEG Neuroheadset. The technique used for capturing these signals is non-invasive. Through an EEG headset which we can detect the electrical activity that is going on in the brain through small, metal discs (electrodes) attached to your scalp. In our experiments, these electrodes are attached at 14 different points on the scalp, which represent the different areas of the scalp. This headset's readings have been made available on the UCI Machine Learning Repository by Oliver Roesler[5]. These readings are then used for performing analytics on it. Firstly, the data would be cleaned, i.e., preprocessing the data would be done using data balance check, checking for missing values, and maintaining authenticity and ambiguity. It would then be passed to the predictive machine learning model that would help classify the data, i.e., predicting whether the eyes are opened or closed.

## 2. Related Research on Brain Wave Analytics for Eye Movement

Pawel Kaprowski implemented a technique of performing individual identification based on the characteristics of eye movement. They have used this method for measuring human eyes reaction for visual stimulation. An eye-tracking method is shown employing which it is easy to identify people and has significant advantages. Also, it can be combined with other camera-based techniques for more secure identification. Eye physiology has been briefly discussed, like the retina has dense clustering at its center rather than its periphery. The entire eye movement system is being controlled by the oculomotor system which in turn is controlled by the brain. Recording the movements of the eye of an individual for a predefined time is called a probe, and the person's eye movements along-with where the person is looking at should be recorded in the system. Some considerations have to be kept while recording a probe, i.e., for getting a similar kind of result a person should be looking at the same sight, but the problem is that it becomes tedious for the person to look at the same sight and the brain learns movements called as learning effect and acts otherwise after several recurrences of the same image. The authors have also mentioned different ways of recording probes, like a visual task, and text reading task. Looking at different images would eliminate this effect, but it would provide different probes and would be difficult to train a system using that. The next


*Copyright ⓒ 2020 FuCoS*

way to eliminate this effect would be a wandering point, where the eye has to track the point on the screen, the drawback of this method is that the will of the person is completely ignored. But this is the best method and looking at the same stimulation again and again would make the result of the probe converge at a point which can then be used for training then model. A jumping point stimulation was selected, and the probe should not last for more than 10 seconds but should have as many point position changes as possible. Different methods like Karhunen-Loeve transform, Fourier transform, cepstrum, or wavelets were tried for feature extraction. Of these, cepstrum was selected because of its victory in voice identification. For classification, four different algorithms were presented in the paper, including a k-nearest neighbour, a C4.5 decision tree, Naïve Bayes, and support vector machines. The best performing model was the k-nearest neighbour algorithm, where k was set to 3[6].

Chiara Galdi and Michele Nappi have discussed different approaches that help in eye movement analysis and which have been developed over the years. However, the main focus of the paper lies in biometric recognition. There are two main categories through which a person can be identified: physical - DNA, fingerprints, face, iris, and behavioural - signature dynamics, gaze, keystroke, and gait. Eye movements can be classified into two types: saccades - eye movements that occur quickly whose duration lasts less than 100 ms and fixation - the period in which eye movements are fixed at a particular point is usually for a period if 100-600 ms. The authors have described various eye-tracking devices that have been used for recording eye movements. With the advancements in technology over the years, four categories of methods have been developed for eye-tracking that include photo-oculography (POG), electrooculography (EOG), video-oculography (VOG), combined pupil-corneal reflection, and scleral contact lens/search coil. The approaches that are mentioned above help in measuring the eye location relative to the head. However, the blended practice of pupillary and corneal impressions disambiguates head movement and eye rotations. The author has also described how different authors have developed different eye-tracking technologies over the years. Various different kinds of stimuli have been developed, which is used for training the data. The most employed stimuli are the jumping point, which was also used in [6] for experimentation. Inattentive optic movements are mainly exploited for analysis of eye movements, as described by the author in the paper. The static cross stimulus is where a cross that is small is at the screen's center. In this, saccades originating from the observer's eyes are recorded, and it has been successful and better than jumping point and can help in classifying, which can have an accuracy of 90%. The author has also said a different approach to trigger eye movements would need an object in motion. The stimulus is usually obtained by


*Copyright ⓒ 2020 FuCoS*

displaying the object in motion, i.e., it can be a circle, a cross, or any other geometrical pattern on a screen. The stimulus itself defines the gaze path, and this is what the above stimuli follow. Another kind of stimuli can be a grey-scaled or coloured image for eye movement analysis. It can be of two types: task-based observation (where the viewer is given a task like finding an object on the screen, seeing a photo, make a decision by pressing the yes/no button) and free viewing (where the viewer can look at the illustrated picture for a planned time). The author has also described that in some works video can also be used as a stimuli in which along-with saccadic movements, the fixation points are also analyzed. The author has then described head-mounted eye-tracker based approaches, remote eye-tracker based approaches, and movement of the eye for demographic categorization. The zenith goal of a recognition system is to perform identification quickly, transparently, and accurately[7].

Evgeniy R. Abdulin and Oleg V. Komogortsev have presented a paper in which they have performed person verification by utilizing the eye movement biometrics model. While reading text, there are a series of saccades and fixations that occur, i.e., eyes can only see a restricted portion of text on a screen. The authors have also explained the metrics they have used, i.e., TRM metrics: Hop-related metrics, Pass-related metrics, and Fixation-related metrics. For experimentation purposes, an EyeLink 1000 eye-tracker was used for recording the eye movements. To eliminate the learning effect, as explained in [6], the authors had given a hard reading stimulus to the participants. The data was collected from 103 participants, 51 males and 52 females. Two kinds of tests were conducted, a same-day test where the data was recorded on the same day and had very low average calibration error; and template ageing test where the same 103 subjects came for the test precisely after one month to see the effect of ageing on biometric identification. It was found that saccadic behaviour is not affected by ageing, which provides the highest biometric accuracy. The collected data was then passed to a velocity threshold algorithm (I-VT) for employing classification. The outliers were then removed from the data, where the outliers were defined that were three standard deviations away from the mean. The proposed technique did not exceed an existing movement of eye driven biometric method, in the same day, it outperformed the disorganized way in a scenario where the tunnel of time might alter the properties if captured eye movements. Therefore, the proposed system's primary usage would be in actual life where the time period between enrollment and the subsequent uses of the biometric system might be large[8].

In this paper, Tobias Bornakke et al. have presented the importance of methodological specificity supported by practical cases to complement insights of big and thick data. The principal methodology introduced in the article is about combining analytical ideas. The


*Copyright ⓒ 2020 FuCoS*

combining usually lies upon the benefaction from the two or more input spaces containing thick data and big data analytics, which administer some conceptual connections in a generic term. One of the paper's essential contributions is analytical objectives floored in big and thick data, conceptually combining acumens based on profoundly complex datasets. The authors have even emphasized that when thick data and big data are blended, it unfolds such insights that would lead to the creation of novel results. Also, trading with substantial datasets requires unique expertise in that domain in the same way ethnography and microanalysis skills are required for thick data. One should always try to acknowledge the differences in both the domain, exchanging the focus towards analytical results of different methods during the blending processes. The authors have finally concluded that for developing novel insights from big data and thick data, scholars from both the domain are needed[9].

Jinan Fiaidhi et al. have presented the increasing standard of thick data analytics and concentrated on valuable identification of communication centers. It has been shown in the paper how these conversation structures have an extensive influence on how to understand the social network's outcome. The document has also given light on how big data only is not able to solve challenging business problems. It was explained by giving an example of the company Lego, how Lego utilized the insights provided by thick data by employing conventional construction pieces and concentrate less on action stuff and figures. Usually, thick data allows the emotions and context of analyzed subjects to be divulged, while big data requires an algorithmic model for its prediction. Big data is referred to as quantitative data in this paper, as normally, it is present in large numbers. On the other hand, the data that gives contextual information is referred to as thick data; even though it is small, it significantly impacts the decision. In the paper, it has also been mentioned that big data becomes stale after a period of time, and new data needs to be added for its continuous authenticity. Inventing insights from social media conversations over thick data, thick data crawler, and thick data indicators and human conversation detectors have also been explained. According to Fiaidhi, "Meaningful data tells a more complete story - ultimately getting closed to the "why""[4].

Jinan Fiaidhi has presented a visionary insight-driven learning (IDL) based approach for gaining meaningful insights from thick data and transferring it to the healthcare provider so that they can make use of those insights for the betterment of patients. The primary focus of this article is on social media data, particularly Twitter data, as it is omnipresent and has geolocation focus. The paper has described a comprehensive view of the identification of useful patient insights by mining relevant Twitter data over patient pain points using qualitative IDLs. Community detection algorithms have been used for identifying conversation networks. It is


*Copyright ⓒ 2020 FuCoS*

also mentioned in this paper the failure rate of companies using big data analytics is quite high with the metrics at the cost of seizing the customer's viewpoint data. The first five IDL methods can help in providing meaningful thick data analytics to the healthcare providers. The remaining, i.e., IDL 6, is used for training the machine learning algorithm from the inferred thick data observations, and IDL 7 is used if there are any more conversations or detected, then the machine learning models are updated accordingly[10].

In [11], S. H. Adil et al. have performed a comprehensive analysis of data that is collected through an EEG device. Electroencephalography is a monitoring system to film the electrical activity of the brain. The data that is used in this paper is collected through a 14 channel EEG device, precisely using Emotiv EEG Neuroheadset. Classification is then performed on the collected Eye State Data and the results obtained are compared and contrasted with [12], that is the base paper. The proposed methodology by the authors seems to have surpassed the classification accuracy of [12]. The following are the steps in which the classification accuracy was predicted, firstly the data collection - it is a step where data is collected through an EEG device, then splitting the dataset into a training set (70%) and testing set (30%). For building the grouping model utilizing KNN, the preparation information is then utilized in K-fold (i.e., parting the preparation information into k segment, each time k-1 parcel is utilized for preparing and 1 segment is utilized for approval). The creators have then applied distinctive boundary setups to construct the characterization model utilizing the preparation set. Initially, arrangement is performed utilizing the prepared KNN model on test dataset. Furthermore, the expectation precision is estimated on the test dataset. Finally, select the model precisely KNN model with the parameter that gives the best classification accuracy. The authors have even given the python code for their implementation in their paper. The author also mentioned that most of the researchers customarily adopt complicated algorithms in search of high-grade results even without acknowledging simple algorithms. This paper suggested that simple algorithms sometimes do give better classification accuracy than complicated algorithms.

## 2.1 The Dataset

Certainly we can start collecting brain signals using the OpenBCI setup. However, such data was collected by Oliver Roesler and David Suendermann[12] to distinguishing eye state, i.e., to check whether the eye is open or shut (EEG Eye State Dataset). It is a time-series dataset that is freely available to download from the UCI Machine Learning Repository[5]. The dataset characteristics are multivariate, sequential, and time-series. The attribute features are real


*Copyright ⓒ 2020 FuCoS*

numbers and integers. All the information is recorded by Emotiv EEG Neuroheadset from one nonstop EEG estimation. The duration of each continuous EEG measurement is 117 seconds. The truth labels were then added manually after analyzing the video frames as the eye state was distinguished through a camera during the EEG estimation. The dataset contains the data ordered in sequential request with the primary estimated an incentive at the head of the information. [Fig. 1] represents the placement of Emotiv EEG Neuroheadset[13], on the scalp at the 14 different electrode positions as shown. The dataset attributes are AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4.

The ground truth is represented in the last column of the dataset Output= {0, 1} where:

0 - eye open state

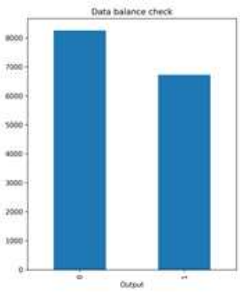
1 - eye closed state.

## 2.2 Using Basic Thick Data Focusing Techniques

Data focusing is an essential step in refining the data and making it closer for prediction. Usually, loosely controlled methods are data gathering, for instance, missing values, categorical data, see if the data are balanced, and many more. Therefore, to avoid misleading results, analyzing such kind of aberrations in the data becomes a crucial step[14]. Following are the dataset preprocessing techniques applied:

### 2.2.1 Balancing the Dataset

Before moving further with making a machine learning model for classification. First, check that the dataset is balanced. [Fig. 3] shows the visual representation of the dataset to see if there is any data imbalance. After visualizing, it can be clearly seen that data imbalance exists.

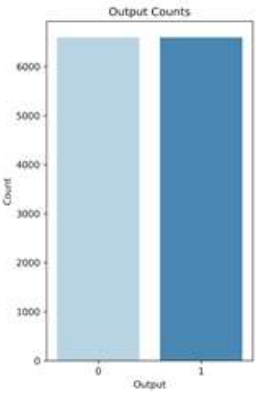



[Fig. 3] Imbalanced EEG Dataset

For balancing the data, the Synthetic Minority Oversampling Technique (SMOTE) is used. It is a special technique that was introduced by Chawla et al.[15] with the help of which new


*Copyright ⓒ 2020 FuCoS*

synthetic examples can be generated. In view of feature space similarities between the minority class's unique instances, this method creates artificial examples. It does this by presenting manufactured models along the lines of selected closest neighbors from the minority class and itself, by taking every case of the minority class. The distance is determined dependent on the Euclidean distance metric for numerical highlights[16]. SMOTE is used only on the training, and the testing set has the same data imbalance, [Fig. 4] shows the bar graph of labels 0 and 1 after oversampling, i.e., data is now balanced.




[Fig. 4] Balanced Data using SMOTE Algorithm

### 2.2.2 Checking for Missing Values

The next step in data preprocessing would be checking for missing values. If there are any missing values found, they have to be replaced by the row's mean because no values inside the data frame create noise. That is why it is better to replace them with the mean. By running the code to check for the missing values, it was discovered that there are no missing values in the dataset. The data is already free from the missing values.

### 2.2.3 Maintaining Authenticity and Ambiguity

For maintaining authenticity and ambiguity, we perform shuffle on our dataset to make sure that there is no order associated with our samples.

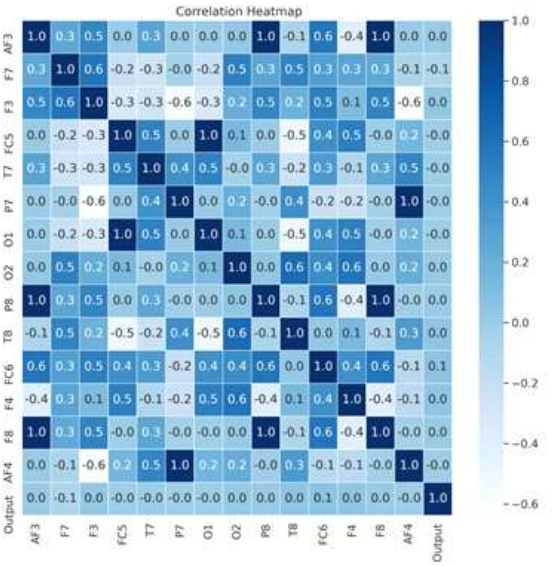
Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a way of interpreting datasets to summarize, often with visual methods, their main characteristics. It is primarily used to tell us what the data can tell us beyond the formal modelling or hypothesis testing task, not necessarily by making the use of a statistical model[17].


*Copyright ⓒ 2020 FuCoS*

Correlation Heatmap:

A correlation matrix is produced using python code, and a visualization heatmap was created for the correlation. [Fig. 5] represents the correlation heatmap.




[Fig. 5] The Correlation Heatmap

In a correlation matrix, the higher the value of correlation, the stronger the relationship would be, and stronger the relationship, it would be difficult to distinguish between those two features. From [Fig. 3] shown above, it seen that the correlation matrix is quite balanced, and there is no strong relationship that exists between the two features, except for a few.

## 3. Thick Data Analytics using Ensembles Learning

Once the data is cleaned and readily organized, it is now ready to be deployed in the proposed model. The overall methodology of the model has been defined in such a way that it would help in providing meaningful analysis once it has been employed on the machine learning classifiers. In turn, it would help in letting us know of the eye state, i.e., is the eye open or closed. In this paper, the focus has been given on ensemble methods for binary classification of EEG signals. [Fig. 6] shows the overall methodology that is being used in the paper. The capability to specifically differentiate observations is valuable for various applications that involve prediction, which can also be aligned with the medical field to extract the best results with the help of machine learning. Since the EEG signals (thick data) are being analyzed with the assistance of machine learning, if it can be amalgamated with big data,


*Copyright ⓒ 2020 FuCoS*

astonishing results can be produced.

An ensemble is a supervised machine learning algorithm. Ensemble techniques are learning algorithms built by a set of classifiers. With the help of that group, new information focuses on taking a combined(weighted) vote of their expectations[18]. The cleaned dataset has been passed on to all kinds of ensemble classifiers that are present in the python's scikit-learn library. The following is the underlying explanation for them:

**(1) Adaptive Boosting Classifier**

Adaptive Boosting Classifier is often referred to as AdaBoost. The final output of the boosted classifier is the yield of other learning calculations consolidated into a weighted total. It changes resulting weak learners for those occurrences misclassified by past classifiers. It is even sensitive to boisterous/noisy data and anomalies/outliers[3].

**(2) Bagging Classifier**

Bagging classifier is also called bootstrap aggregating, which is utilized to improve the precision and stability of the machine learning algorithm utilized for analytical classification. It helps to avoid overfitting and also helps in avoiding variance. Bagging is a remarkable instance of the model averaging approach[19].

**(3) Stacking Classifier**

In a stacking classifier, the predictions of multiple classifiers (commonly cited as level-one classifiers) are used as new features to train a meta-classifier for the final prediction[20]. In our case, one K-Nearest Neighbor classifier with k = 1 is used, with the meta-classifier being Extra Trees classifier.

The following is an overview of non-ensemble methods that were used for testing.

A. K-Nearest Neighbour

k-NN is a type of instance-based learning; the input usually comprises of k closest preparing models in the component space. The yield is a class membership, the plurality vote of its neighbours are used to classify an object, and the object is usually doled out to the class generally regular among its k closest neighbours[21].

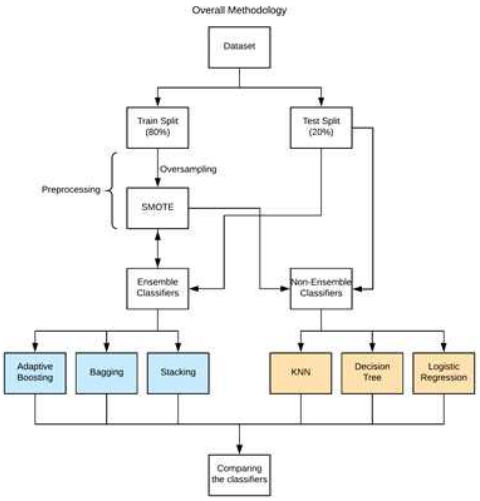
B. Decision Tree

A decision tree is a flowchart-like structure in which an inward hub constitutes a test. Each leaf hub constitutes a class name (choice taken subsequent to processing all the attributes), and each branch constitutes the result of the test. The way from the root to the leaf shows us the classification rules[22].

C. Logistic Regression


*Copyright ⓒ 2020 FuCoS*

It is a machine learning algorithm that is utilized for classification issues; it depends on the idea of likelihood and is a prescient investigation algorithm. It utilizes a more perplexing cost work known as the sigmoid function[23].

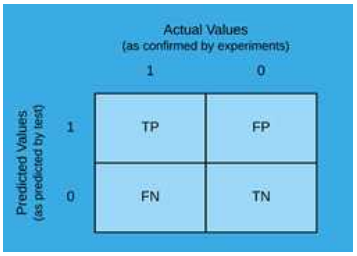



[Fig. 6] Our Overall Comparative Methodology

The classifiers mentioned in the Methodology section of this paper have been employed on the preprocessed dataset. These classifiers have been employed using the pre-existing python libraries present in scikit-learn. The accuracy score, confusion matrix, and classification report have been calculated for each one of them. The classification report includes the calculation of precision, recall, and f1 score.

### 3.1 Confusion Matrix:

A confusion matrix ciphers the performance of classification on test data of a classifier. It is generally a two-dimensional matrix and is represented in [Fig. 7]. It is a confusion matrix for two classes[24].




[Fig. 7] The Confusion Matrix


*Copyright ⓒ 2020 FuCoS*

There can be a number of classification performance parameters that can be derived based on this matrix.

Precision (Positive Predicted Value): It is the proportion effectively predicted positive perceptions to total predicted positive perceptions. It can be represented by an equation as follows:

$$\frac{TP}{TP+FP}$$

Recall (True Positive Rate): It is the proportion of accurately predicted positive observations to all the measurements of positive predicted actual values. It can be represented by an equation as follows:

$$\frac{TP}{TP+FN}$$

F1 Score: It considers both recall and precision. It can be represented by an equation as follows:

$$2*\frac{\mathrm{Precision}*Recall}{\mathrm{Precision}+Recall}$$

Accuracy: It is the proportion of accurately anticipated observations to their absolute observations. It can be represented by an equation as follows:

$$\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

### 3.2 Ensemble V/S Non-Ensemble Classifier Comparison

After employing the classifiers listed above, the methodology was evaluated using the performance metrics. [Table 1] shows the comparison of the four performance parameters when tested against the six classifiers. For a better understanding of the table, the ensemble classifiers are shown by cells in blue and non-ensemble are shown by cells in orange. All the values in the table can be converted to percentages. For example, 0.98 will be equal to 98%.

[Table 1] Ensemble V/S Non-ensemble Classifier Comparison

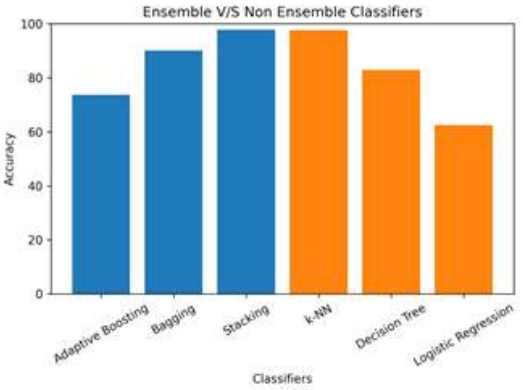
| Algorithm | Precision | Recall | F 1 Score | Accuracy | MCC |
|---|---|---|---|---|---|
| AdaBoost | 0.73 | 0.73 | 0.73 | 0.74 | 0.47 |
| Bagging | 0.90 | 0.90 | 0.90 | 0.90 | 0.808 |
| Stacking | 0.98 | 0.98 | 0.98 | 0.98 | 0.953 |


*Copyright ⓒ 2020 FuCoS*

| | | | | | |
|---|---|---|---|---|---|
| k-NN | 0.98 | 0.98 | 0.98 | 0.98 | 0.951 |
| Decision Tree | 0.83 | 0.83 | 0.83 | 0.83 | 0.653 |
| Logistic Regression | 0.62 | 0.62 | 0.62 | 0.63 | 0.248 |

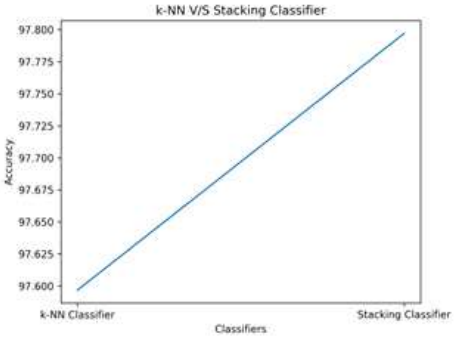
The above comparison concludes that the most efficient algorithm for classifying the eye state for thick data analysis is the Stacking classifier.

Due to the advancement in the data visualization techniques, different classifiers are compared based on their accuracy. [Fig. 8] represents the same. It can also be seen that the accuracies of k-NN and Stacking classifiers are almost identical.




[Fig. 8] Comparison of Ensemble V/S Non-Ensemble Classifiers based on Accuracy

To have a more unobstructed view of the accuracy, another graph has been plotted, which shows only the accuracy of the two classifiers. [Fig. 9] represents the same; it shows that k-NN classifier's accuracy is around 97.60%, and the Stacking classifier's accuracy is approximately 97.80%, where we can see the line going in a positive direction from 97.60 to 97.80.




[Fig. 9] Comparison of k-NN V/S Stacking Classifier

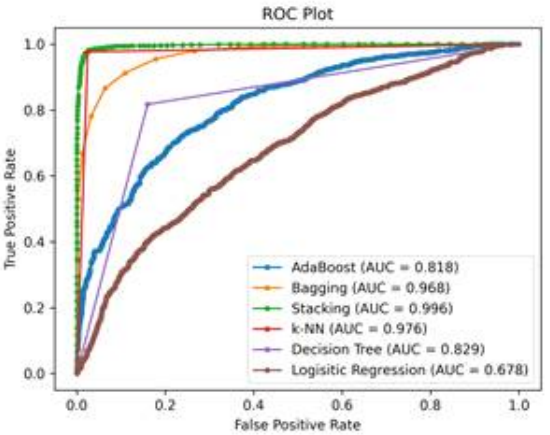

*Copyright ⓒ 2020 FuCoS*

### 3.3 Receiver Operating Characteristics (ROC)

An ROC curve is a graph that is plotted against True Positive Rate (TPR) and False Positive Rate (FPR). TPR's equation is represented above, and FPR can be represented as:

$$\frac{FP}{FP+TN}$$

It is a diagram indicating the exhibition of a classification model at all thresholds of classification. If the classification is lowered, it will classify more items as positive, increasing both False Positives and True Positives[25]. AUC represents the region under the ROC curve, which quantifies the whole two-dimensional territory underneath the entire ROC curve. [Fig. 10] below shows the AUC score comparison of all the classifiers used.




[Fig. 10] AUC for all the Classifiers

For a classification model, to perform better, its AUC score should be closer to 1, then it is correctly able to classify 1's as 1's and 0's as 0's. From the above figure, AUC for the stacking classifier is the highest, having a value of 0.996, which is the closest to 1, making it the best classification model for our dataset. The following are the next best classifiers after the chosen model (stacking classifier), i.e., the k-NN, Bagging, Decision Tree, AdaBoost, and Logistic Regression.

### 3.4 Matthews Correlation Coefficient (MCC):

One other type of metric that is used for calculating the measure of how the model is


*Copyright ⓒ 2020 FuCoS*

correctly classifying the values is MCC. This measure is also unaffected by the data imbalance issue. It is essentially a contingency matrix technique for ascertaining the Pearson product correlation coefficient among real and anticipated values. MCC is the only binary classifier rate that assures us a good score that should lead to +1 only if the binary classifier correctly predicts the positive data instances as positive and negative data instances as negative. The range for MCC is [-1, +1], where the extreme values -1 represent perfect misclassification and +1 perfect classification. It can be represented by an equation as follows:

$$\frac{TP\ .\ TN - FP\ .\ FN}{\sqrt{(TP+FP)\ .\ (TP+FN)\ .\ (TN+FP)\ .\ (TN+FN)}}$$

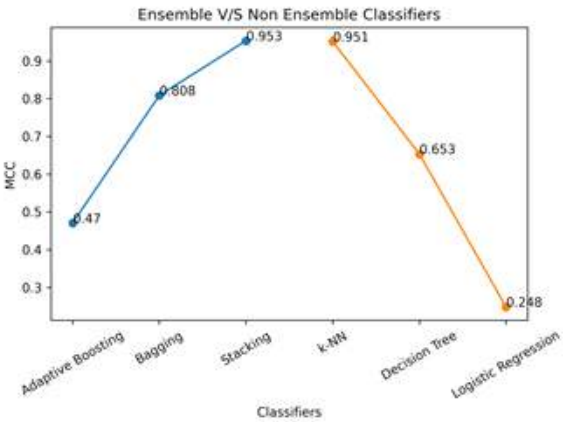
Since MCC is unaffected by data imbalance, we have calculated the MCC score before and after applying SMOTE. The reason behind this is to make sure that the accuracy metric score is not deceiving[26]. The following are the scores obtained:

Value of MCC before SMOTE: 0.953421.

Value of MCC after SMOTE: 0.953409.

These values confirm that MCC, in turn is also a valuable, valid metric that can be used to confirm our projected research that Ensemble classifiers perform better than Non-ensemble classifiers.

[Fig. 11] below show the ensemble classifiers and their MCC score in blue line and non-ensemble classifiers and their MCC score in orange line with their MCC value written on4 top of it. All the values are calculated after applying SMOTE.




[Fig. 11] Comparison of Ensemble V/S Non-Ensemble Classifiers based on MCC


*Copyright ⓒ 2020 FuCoS*

The above values confirm that Ensemble classifiers perform better than Non-ensemble classifiers based on their accuracy and MCC results obtained above.

### 3.5 Advantages of using Ensemble Classifiers

Ensemble methods have a plethora of benefits over using traditional classifiers; they are explained below.

The main advantage of ensemble classifiers is that it combines multiple inducers (classifiers) to make a decision, so it has the power of various classifiers in itself. In a way that combines multiple models, there is a probability that other inducers would likely compensate for a single inducer's errors. As a result, an ensemble would serve better by increasing the overall prediction performance, in comparison to the prediction performance, one would have got using a single inducer. This would be rather more beneficial in case the data size is small, i.e., thick data.

Whenever there is an available modest quantity of information, a solitary learning algorithm generally attempts to discover various hypotheses, which help predict the training data correctly but perform imperfectly when presented with unseen data. This is often the case of working with a single classifier. But, take an average of the different hypotheses would help in increasing the overall predictive performance. Also, there is always a need to decrease the local optima (it is a cost function that we have to minimize for a machine learning problem); however, single learners may lead nearby quests that may stall out in neighbourhood optima.

Ensemble methods likewise help in diminishing the scourge of dimensionality. At whatever point the number of features fed into an machine learning model is expanded, the pursuit space additionally builds exponentially and, along these lines, the likelihood of fitting the models that can't be generalized[27].

These methods have the advantage that they can be made to adapt any kind of change that may occur in the monitored data stream more precisely than single classifiers[28].

Ensemble methods also handle the class imbalance problem. There is always a possibility that when the dataset is imbalanced, the machine learning algorithm may build up an inclination for the majority class. Therefore, the ensemble method can come to the rescue if applied in a proper way. This is very clearly demonstrated in our experimentation above when we are taking into consideration the MCC. The value of MCC before applying SMOTE, i.e., when the class was imbalanced, is the same as MCC after applying SMOTE.

Thus, from the above research, we can also conclude that ensemble methods do a better task


*Copyright ⓒ 2020 FuCoS*

in serving thick data, especially in our case, i.e., eye state detection.

### 3.6 Boosting, Bagging, and Stacking – Which is better?

Boosting, bagging, and stacking have been explained above in the Methodology section.

Boosting is a two-step approach where subsets of the original dataset are used to produce a series of averagely performing models whose performance is boosted utilizing a cost function. Subset creation is not random in boosting; the performance thus depends on the previous models[28]. One of the drawbacks of this method is that it is bound to fix its predecessors' errors.

Bagging can be successfully used with extensive and high-dimensional data in datasets, such as the intrusion detection system, since it can be challenging to infer information from just one step of a single classifier[28]. In our case, the dataset is thick data, which corresponds to small data, which becomes a major disadvantage of bagging over stacking.

Stacking uses layers; therefore, it promotes diversity, in a way if different kind of models are used having different strategies, then there would be a variety of opinions, at different levels, models would then disagree with each other inducing a natural diversity, in turn producing a better result[29]. This works primarily in the case of time-series data, just as in our case.

In their paper, Iwan Syarif et al. have compared boosting, bagging, and stacking with single classifiers and have found out that stacking outperforms the others. It was the only method that helped in reducing the false positive rate by a significant amount[28]. This result is also in line with the research in our paper.

Eun Sung Lee has performed a similar comparison on ensemble classifiers (stacking) and the traditional classifier. According to his results, the stacking classifier has the highest predictive performance compared with various single learners[30]. The research presented in our paper also gives us similar results.

According to the explanation above and similar work that has been done in this area, we thus confirm that the stacking classifier outperforms the other two ensemble methods, especially in the case of thick data.

## 4. Comparison with Other Related Work

The following are the papers that have used the same dataset, which is used in this paper.

In their paper, Oliver Roesler et al.[12] mentioned K* instance-based method having the best


*Copyright ⓒ 2020 FuCoS*

classification error rate and mentioned that their accuracy was above 97%. Cameron R. Hamilton et al.[31] had reconstructed the workings of Oliver Roesler and found the accuracy to be 97.3%. They tried to improve the accuracy over the model prepared by Oliver Roesler, and the ensemble obtained the best result by combining the rotational forest that implements random forests as its base classifiers (RRF) with K* instance-based algorithm (RRF + K*). It was found to have an accuracy of 97.40. When compared the readings of both the paper to the current research, the accuracy of this paper still stands out to be 97.80%.

Ting Wang et al.[32], in their research, extracted the raw data features and then sorted them using the Incremental Attribute Learning feature ordering approach. They have also mentioned that the results of time-series data are highly dependent on their relationship with each other. In contrast, in our approach, we have maintained the authenticity and ambiguity of the data by shuffling.

In their paper, S. H. Adil et al.[11] have performed a comprehensive analysis of this EEG dataset and compared it with the result published by [12]. Their proposed KNN algorithm reported 97.5% accuracy, and the time taken was 3 minutes when the CPU was 2.2GHz Dual Core and RAM 8GB. As opposed to the methodology presented in this paper reported accuracy of 97.80%, and the time taken was 18 seconds when the CPU was 2.2 GHz Dual Core and RAM 13GB. [Table 2] below shows the comparison, as reported by [11] along-with our proposed methodology.

[Table 2] Time Performance Comparison WRT Different Algorithms

| | KNN [11] | K* [12] | Proposed Methodology |
|---|---|---|---|
| CPU | 2.2 GHz Dual Core | 2.4 GHz Dual Core | 2.2 GHz Dual Core |
| RAM | 8GB | 32GB | 13GB |
| Time | 3 minutes | 38 minutes | 18 seconds |
| Accuracy | 97.5% | 97% | 97.80% |

Efendi Nasibov et al.[33], in their paper, they have devised the Cxk-NN instance-based algorithm for their experimentation, which is an advanced version of a simple k-NN algorithm. Their results show that the success ratio of Cxk-NN is 0.969, having the value of k as 2, 3, 4, and that of simple k-NN is 0.967 having the value of k as 1. From the above results, it can be seen that the stacking classifier devised in this paper tops their performances.

The novel approach devised in this paper is in regard to the preprocessing stage - i.e., the application of SMOTE to the raw data to make it cleaner, that in turn, improves it over the


*Copyright ⓒ 2020 FuCoS*

other systems implemented. Application of SMOTE causes the removal of data imbalance, which would then help in deriving gripping analytics. This is wholly unique with respect to the papers [11], [12], [31], [32], and [33], who have tried to apply different strategies mentioned above along-with their results. In this paper, near investigation of traditional classifiers with respect to ensemble classifiers revealed that consistent higher accuracy is obtained through ensemble classifiers. Our paper still tops in terms of typical parameters like accuracy. This paper also exercises non-typical settings like MCC and finds the model stability using AUC[34][35].

## 5. Conclusions

Detecting the eye state has been a research area for a long time. With advancements in technology, brain readings are recorded through a process called EEG and eye state detection has been performed on that data. Preprocessing of the dataset is done by oversampling the data using SMOTE, which has been extensively explained. After this, ensemble and non-ensemble classifiers have been applied to this data. Ensemble classifiers include boosting, bagging, and stacking. Non-ensemble classifiers include k-NN, decision tree, and logistic regression. Analytics has been performed by comparing the classifiers mentioned above, with each other to find out which one of the classifiers stands out, and from which group. Various kinds of classification metrics, including MCC and AUC score, have been used to have a clearer picture. Empirical results revealed that a stacking classifier produced fewer false positives, which further helped to deduce better predictive performance compared with other classifiers. This can also be seen through the easily understandable visualizations in this paper that embark on ensemble classifiers' performance over thick data. Research performed to deduce which group of classifiers (ensemble or non-ensemble) is better tells us that ensemble has more advantages over single classifiers/learners. Analysis has also been performed to see which among the ensemble classifiers, i.e., boosting, bagging, and stacking is better. This extensive experimentation showcases that the stacking classifier clearly outperforms all the other classifiers by achieving the highest accuracy and MCC.

## 6. Acknowledgments

This research is part of the second author NSERC DDG Research Grant on Thick Data Analytics. It is also part of the MSc Project of the first author.


*Copyright ⓒ 2020 FuCoS*

## References

[1] Ioannis Rigas, George Economou, Spiros Fotopoulos, Biometric identification based on the eye movements and graph matching techniques, Pattern Recognition Letters, **(2012)**, Vol.33, No.6, pp.786-792, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2012.01.003.

[2] Tricia Wang, Why Big Data Needs Thick Data, https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7#:~:text=Thick%20Data%20reveals%20the%20social,isolates%20variables%20to%20identify%20patterns, May 12 (2020).

[3] AdaBoost, https://en.wikipedia.org/wiki/AdaBoost, May 30 (2020).

[4] J. Fiaidhi, S. Mohammed, Thick Data: A New Qualitative Analytics for Identifying Customer Insights, IT Professional, **(2019)**, Vol.21, No.3, pp.4-13.

[5] EEG Eye State Dataset, https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State, May 15 (2020).

[6] Kasprowski P., Ober J., Eye Movements in Biometrics, International Workshop on Biometric Authentication (BioAW), Lecture Notes in Computer Science, **(2004)**, Vol.3087, pp.248-258.

[7] Galdi Chiara, Nappi Michele, Eye Movement Analysis in Biometrics, Series in BioEngineering, **(2019)**, pp.171-183.

[8] R. Abdulin, O. V. Komogortsev, Person verification via eye movement-driven text reading model, 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), **(2015)**, September 8-11; Arlington, VA, pp.1-8, doi: 10.1109/BTAS.2015.7358786.

[9] Bornakke Tobias, Brian L. Due, Big-Thick Blending: A method for mixing analytical insights from big and thick data sources, Big Data & Society, **(2018)**, Vol.5, No.1, 2053951718765026.

[10] J. Fiaidhi, Envisioning Insight-Driven Learning Based on Thick Data Analytics With Focus on Healthcare, IEEE Access, **(2020)**, Vol.8, pp.114998-115004, doi: 10.1109/ACCESS.2020.2995763.

[11] S. H. Adil, M. Ebrahim, K. Raza, S. S. Azhar Ali, Prediction of Eye State Using KNN Algorithm, 2018 International Conference on Intelligent and Advanced System (ICIAS), **(2018)**, August 13-14; Kuala Lumpur, pp.1-5.

[12] O. Roesler, D. Suendermann, A First Step towards Eye State Prediction Using EEG, In Proceedings of the AIHLS, **(2013)**, Istanbul, Turkey, http://ehrai.com/su/pdf/aihls2013.pdf

[13] Emotiv Epoc +, https://www.emotiv.com/epoc/, May 15 (2020).

[14] Pyle Dorian, Data preparation for data mining, USA: Morgan Kaufmann Publishers, Inc, **(1999)**

[15] Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, **(2002)**, Vol.16, pp.321-357. https://arxiv.org/pdf/1106.1813.pdf

[16] Cost S., Salzberg S., A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, Machine Learning Journal, **(1993)**, Vol.10, No.1, pp.1213-1228.

[17] Chatfield, Chris, Problem solving: a statistician's guide, USA: CRC Press, **(1995)**

[18] Dietterich, Thomas G., Ensemble methods in machine learning, International workshop on multiple classifier



*Copyright ⓒ 2020 FuCoS*

systems, Springer, **(2000)**, June 21-23, Cagliari, Italy, pp.1-15.

[19] Bootstrap Aggregating, https://en.wikipedia.org/wiki/Bootstrap_aggregating, May 30 (2020).

[20] Frank Ceballos, Stacking Classifiers for Higher Predictive Performance, https://towardsdatascience.com/stacking-classifiers-for-higher-predictive-performance-566f963e4840, May 30 (2020).

[21] k-nearest neighbors algorithm, https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm, May 30 (2020).

[22] Decision tree, https://en.wikipedia.org/wiki/Decision_tree, May 30 (2020).

[23] Ayush Pant, Introduction to Logistic Regression, https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148, June 3 (2020).

[24] Claude Sammut, Geoffrey I. Webb, Encyclopedia of Machine Learning and Data Mining, USA Boston: Springer, https://doi.org/10.1007/978-1-4899-7687-1_50, **(2017)**

[25] Classification: ROC Curve and AUC, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, June 6 (2020).

[26] Chicco, D., Jurman, G., The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics, **(2020)**, Vol.21, No.6, https://doi.org/ 10.1186/s12864-019-6413-7

[27] Sagi Omer, Lior Rokach, Ensemble learning: A survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **(2018)**, Vol.8, No.4, e1249.

[28] Syarif I., Zaluska E., Prugel-Bennett A., Wills G., Application of Bagging, Boosting and Stacking to Intrusion Detection, Machine Learning and Data Mining in Pattern Recognition, MLDM, Lecture Notes in Computer Science, **(2012)**, Vol.7376, German Berlin: Springer, https://doi.org/ 10.1007/978-3-642-31537-4_46.

[29] Ribeiro, Matheus Henrique Dal Molin, Leandro dos Santos Coelho, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, Applied Soft Computing, **(2020)**, Vol.86, p.105837. https://doi.org/10.1016/j.asoc.2019.105837

[30] S. Lee, Exploring the Performance of Stacking Classifier to Predict Depression Among the Elderly, IEEE International Conference on Healthcare Informatics (ICHI), **(2017)**, August 23-26; Park City, UT, pp.13-20, doi: 10.1109/ICHI.2017.95.

[31] Hamilton, Cameron R., Shervin Shahryari, Khaled M. Rasheed, Eye state prediction from EEG data using boosted rotational forests, 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), **(2015)**, December 9-11; Miami, FL, USA, pp.429-432.

[32] Wang, Ting, Sheng-Uei Guan, Ka Lok Man, T. O. Ting, EEG eye state identification using incremental attribute learning with time-series classification, Mathematical Problems in Engineering, **(2014)**, Vol.2014, https://doi.org/10.1155/2014/365101

[33] Nasibov, Efendi, Alican Dogan, An efficient algoritm for classification of EEG eye state data, Global Journal of Information Technology: Emerging Technologies, **(2016)**, Vol.6, No.3, pp.158-165.

[34] Ho Tin Kam, Random decision forests, Proceedings of 3rd international conference on document analysis and recognition, **(1995)**, August 14-16, Montreal, Quebec, Canada, Canada, Vol.1, pp.278-282.

[35] sklearn.ensemble.StackingClassifier, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html, June 6 (2020).


*Copyright ⓒ 2020 FuCoS*