CHAPTER

A SYSTEM FOR THE ANALYSIS OF EEG DATA AND BRAIN STATE MODELING

30

S. Subedi*, Y. Li†, C. Early‡, A. Chan§, J. Garza*, G. Schreiber¶, Y. Chang*, H. Lin*

*Department of Computer Science and Engineering Technology, University of Houston-Downtown, Houston, TX, United States** *Department of Mathematics, Illinois State University, Normal, IL, United States†* *Department of Science and Engineering Technology, University of Houston-Clear Lake, Houston, TX, United States‡* *Department of Statistics, University of California, Los Angeles, CA, United States§* *Chevron-Phillips Chemical Company, Houston, TX, United States¶*

30.1 INTRODUCTION

Renowned scientist and philosopher Galvani was the first person to discover electrical activity in a living organism in the 18th century [1]. Later, the electrophysiologist Hans Berger successfully recorded electrical activity from the human brain using electroencephalography (EEG), which measures voltage oscillations due to ions flow in the neurons of the brain [1]. Today, EEG is one of the popular noninvasive techniques to record brain activity in clinical and research settings, and there is a wide range of applications for the analysis and interpretation of these measurements. The development of EEG devices, for example, EPOC from Emotiv (http://www.emotiv.com) and NeuroSky (http://www.neurosky.com), and increasing interest in EEG data analysis is evident. EEG data carries an immense potential in its usability in various areas including human-computer interaction, psychology, and neurological sciences. Therefore, it is a valuable endeavor to design an application that applies various analytical techniques to EEG data and predicts the state of the brain from which the data was acquired.

There are five major waves recorded by EEG devices (Table 30.1). Beta and gamma waves are linked with mechanism of consciousness, while alpha waves are associated with disengagement [2]. Similarly, inefficiency and daydreaming occurs during theta waves, and finally, delta waves are associated with low activity and sleeping [2]. There are numerous studies aimed at deciphering the complex relationship among consciousness of the brain, the underlying pattern of its activity, and the generation of waves, using mathematical models and computing technology. Yang et al. [3] have proposed some novel feature extraction methods using harmonic wavelet transform and bispectrum for EEG signals to be used in a brain-computer interface (BCI) system to classify left- and right-hand motor imagery. The experimental results have shown that the separation of the classes extracted by the proposed method achieved recognition accuracy of 90%. Similarly, in a different study, the spectrum analysis of brain waves using specific music stimulus has been successfully completed utilizing various statistical models [4]. The research group found that the upper alpha wave was entrained under the special

Emerging Trends in Applications and Infrastructures for Computational Biology, Bioinformatics, and Systems Biology.
http://dx.doi.org/10.1016/B978-0-12-804203-8.00030-4

Table 30.1 Major Brainwave Frequencies

Brainwave Type	Frequency Range (Hz)	Mental States and Conditions
Delta	0.1–3	Deep, dreamless sleep, non-REM sleep, unconscious
Theta	4–7	Intuitive, creative, recall, fantasy, imaginary, dream
Alpha	8–12	Relaxed, but not drowsy, tranquil, conscious
Low beta	12–15	Formerly SMR, relaxed yet focused, integrated
Midrange beta	16–20	Thinking, aware of self and surroundings
High beta	21–30	Alertness, agitation
Gamma	30–100	Peak focus, super consciousness

brainwave stimulus. This study showed the positive correlation between upper alpha wave generation and memory formation in the brain.

The EEG is also being used to develop innovative systems in healthcare and biomedical research. A recent study has been reported to discover links between emotional states of patients and their brain activity using machine learning algorithms [5]. The research group analyzed EEG data collected during various emotional states from 40 Parkinson's disease patients and healthy subjects using a bispectrum feature and concluded that the higher frequency bands such as alpha, beta, and gamma played an important role in determining emotional states compared to lower frequency bands, delta and theta. In a different study, Direito and group [6] have designed a model to identify the different states of the epileptic brain using topographic mapping relative to delta, theta, alpha, beta, and gamma frequencies. The method achieved 89% accuracy in predicting abnormal versus normal brain states. These studies have reported that variability in analysis occurs due to two major reasons: the first based on the feature extraction method implemented, and the second being the prediction of the model is directly proportional with the increase in the constant variables associated with the modeling equation. This, overall, underscores the complexity of applying mathematical models to a natural phenomenon such as brain activity [5,6].

Brain state modeling research can be divided into two major models: statistical models and micro models. For example, statistical models are built by applying statistical analysis to collected data from meditation practitioners, while micro models try to catch physiological features of the brain state under examination. The current literature shows that both methods are used in the study of complementary and alternative medicine, which includes meditation as one of the methods. One approach is to study finite differences within the minds of those practicing meditation, and those who do not. Such an endeavor is an avenue toward modeling a wide range of brain states [7]. Loizzo et al. [8] performed a 20-week contemplative self-healing program study, which showed that a contemplative self-healing program can be effective in significantly reducing distress and disability among the testers. Habermann et al. [9], on the other hand, performed a long-term (5–20 years) project to investigate the use of complementary and alternative medicine and its effects on the testers' health. Comparisons across different groups of people are also found. For example, in a 6-week mindfulness-based stress reduction program, subjects assigned to the program demonstrated significant improvements in psychological status and quality of life compared with usual care [10]. In another study a group of Qigong practitioners were compared to a control group and positive impact on the quality of life of cancer patients was observed using EEG technologies [11].

In this chapter we focus on statistical classification methods to build a model, and any new data can then be analyzed, compared to the model, and assigned to a particular class. Thus, we can see that this process involves two discrete phases: (1) data modeling and (2) data classification. It is then our proposal that through EEG data collection and machine learning techniques, it should be possible to implement a wide range of specific applications that can provide useful functions to the user. One conceived implementation is an embedded system that would be capable of performing a real-time analysis of EEG data, as it is collected. A device such as this, when given unclassified EEG data, would then be able to classify the brain state of the user into one of a number of different classes, depending on the particular models available to the system, and how these models were trained. As the data is classified, the system would provide an indication to the user of which brain state the incoming data most closely resembles.

30.2 SYSTEM FOR EEG DATA COLLECTION, STORAGE, AND VISUALIZATION

A platform for comprehensive EEG data storage, processing, and analysis is desirable to promote applications of using EEG tools in both physiological (eg, clinical uses, sleep evaluation, fatigue detection, etc.) and psychological (cognitive sciences, BCI, etc.) scopes. Such a platform consists of EEG data collection devices (viz., EEG headset), communication channels (eg, smart phones), an online database for EEG data storage and processing, a web interface for users to access stored EEG data and activate data analysis and classification algorithms, and a forum for users to collaborate with each other while using the system. Fig. 30.1 shows a general outline of the proposed system. A valuable aspect of web-based systems is that most users find a well-designed system to be easy to use without a steep learning curve [12], primarily because of the ubiquity and widespread use of modern web-based applications. Ultimately, this means that nontechnical users can more easily focus on the analysis of the data without spending too much time learning how to use the system.

The various technologies involved in using the web for visualization purposes have been analyzed before. The Holmberg group performed a study on interactive web-based visualization in which they developed a framework to categorize different web-based technologies for 2D and 3D visualization [13]. DHTML/AJAX, or Dynamic Hypertext Markup Language with Asynchronous JavaScript and XML, consists of a combination of HTML and JavaScript and performs well under most conditions,

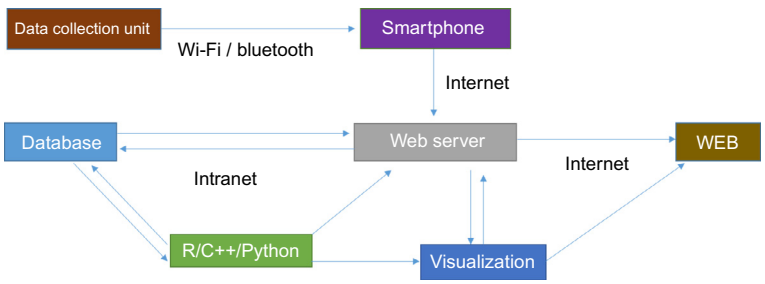

FIGURE 30.1

EEG data analysis system architecture.

with certain disadvantages related to limited communication with servers at the time of this writing. Another point made by Holmberg et al. is the popular use of this technology in commonly used platforms such as Google Maps, Facebook, and many more. Furthermore, a distinct advantage of the use of DHTML/AJAX is the fact that it is not a plugin that users must install but is rather built natively into the web browser. This means that clients do not need to install any special software to run the visualization software under this implementation. Aside from Adobe's Flash, JavaScript is the only widely used and highly compatible solution for executing client-side code, such as the code that allows a static HTML page to spring to life.

Another study, done by the Poliakov group found that a major advantage, among others, in a server-client system setup is that the client's hardware does not need to be particularly powerful as most of the processing and analysis of large data is done by the server [14]. This is desirable for our purposes, partly because servers tend to be more powerful than personal computers. This affords a decrease in execution time, as well as time spent transferring data. Another important factor with regard to server hardware is the increasingly common inclusion of more than one CPU. This makes it possible to execute data processing methods in parallel, consequently reducing the overall processing time of large amounts of data. This becomes crucial in the execution of machine learning algorithms, some of which can require large amounts of processing time, depending on the amount of data. The choice to implement our algorithms on the server side means that the majority of the work is done by the server, reducing both the load on the client as well as the amount of data that must be transferred between the server and the client.

JSON (JavaScript Object Notation) has been shown to be a viable way to transmit data from a server to a client's browser. Many common programming languages aside from JavaScript currently offer support for the JSON format. This provides convenience when developing using a combination of languages, such as with the use of PHP on the server side and JavaScript on the client side. Web-based systems have also shown that it is not only possible, but also desirable to display multiple records of data together, allowing users to better compare interpersonal differences and similarities between different records [14]. This provides a deeper level of analysis than what is possible when merely displaying individual records alone.

30.2.1 EEG DATA COLLECTION AND STORAGE

30.2.1.1 EEG headset

We briefly describe how a simple EEG headset can be built using commonly available materials. Our prototype multifunctional headset includes an EEG sensor, a pulse sensor, a temperature sensor, a microprocessor, and a microprocessor Bluetooth shield. The assembled headset is shown in Fig. 30.2, where the three sensors are mounted on the tips of the three legs on the forehead supports. The microprocessor and the microprocessor Bluetooth shield are mounted on the back, and the earlobe is used as an electrical ground base for the EEG sensor. To test and validate that the headset works properly and that all the sensors are functioning, a test environment was constructed. To approximate a real-world environment, a mobile smartphone application was developed on the Apple iPhone platform. This platform was chosen both for ease of access to development tools and also the wide availability of software development kits from hardware vendors. Both NeuroSky and Red Bear Labs included sample applications that were then transferred to a custom application that uses a simple view to display feedback from the various sensors.

FIGURE 30.2

Prototype of the custom developed headset.

30.2.1.2 Data storage

To provide a reliable storage option that can handle the large amount of incoming data from EEG recording sessions, we store the data in a relational database. The database itself is hosted on a web server. The EEG data is stored in its own separate tables apart from other tables necessary for the basic functionality of the website. This is partly for security reasons, but mostly for clarity and ease of distinction between the EEG data and more general data. When in use, the NeuroSky headset produces a comma separated values (CSV) file that can be quite large for about 3 min of data collection. On average, the resulting file is 8–9 MB in size. These files were manually imported into the database initially, a process that requires that the data file first be transferred to the server and then imported directly into the database. To make this process simpler, we have implemented an interface that allows users to easily and seamlessly upload EEG data files.

After a user logs in to the system, they are given the option to upload a new file. The user can then select the local file they want to upload and provide a unique name for the data sets. At this point, the system will upload the file in the background, while displaying visual feedback to the user in the form of a progress bar (Fig. 30.3). Once the upload has been completed successfully, the file is stored in an upload folder and then immediately parsed into a new table in the database with the name given. Once completed, the EEG data is ready for immediate analysis through the website. If the upload fails for any reason, the user receives an appropriate message and the corrupt data is removed, thereby ensuring consistency.

30.2.1.3 Web visualization of EEG data

Once a repository of EEG data exists, it is necessary that a method exists to review and visualize the stored data. With this purpose in mind, we developed the following web-based visual and statistical review methods. The web server provides a user interface that allows users to view EEG data stored in the database and perform analysis on this data. Fig. 30.4 shows the web interface as it is displaying data that is rendered in waveform mode and statistical mode, respectively. The waveform rendering seen in Fig. 30.4 is generated dynamically within the user's browser. This allows the graph to be zoomed in or moved around, so as to focus on a specific section of the graph.

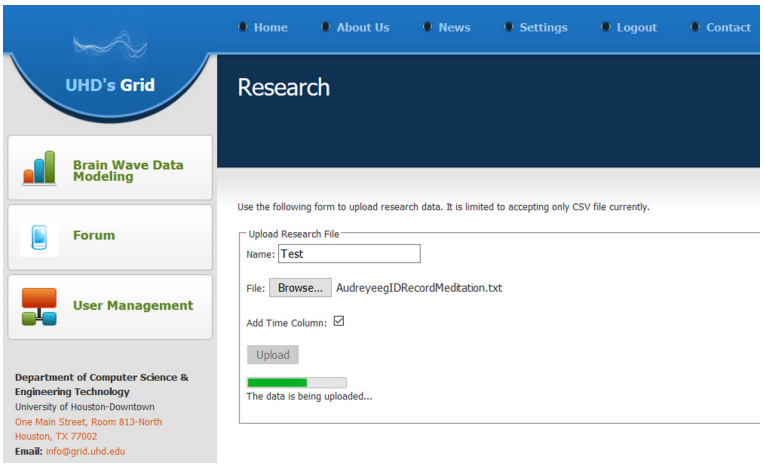

FIGURE 30.3

EEG data upload interface.

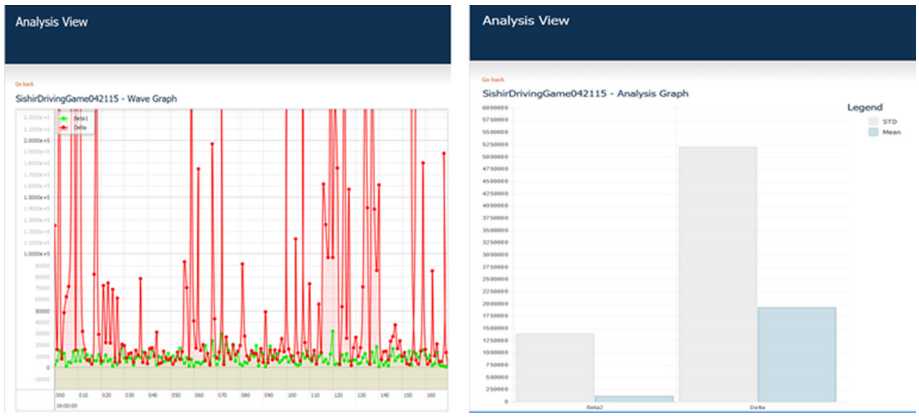

FIGURE 30.4

EEG wave analysis rendered by web interface.

30.2.1.4 Data modeling interface

Once the EEG data has been collected and stored in a standard format, it can be analyzed and modeled. In this section we present first a basic description of the way in which EEG data is measured and quantified. Next we describe the implementation of a variety of machine learning classification algorithms. The inclusion of these analysis methods is central to the overall web-based storage and analysis system, because they allow the user to analyze and test data from various subjects and collection times through a centralized interface. The current version of our toolkit includes the following classification algorithms: K-nearest neighbors (KNN), support vector machine (SVM), boosting, randomized aggregated decision trees (random forest), as well as a naïve Bayesian classifier. To implement this range of

analytic methods, we have used multiple different languages and techniques. For example, our preliminary analysis and modeling of the characteristics of EEG data was done in the statistical language R and Python. Meanwhile, the machine learning algorithms available through the web interface are written in C++, and rely on the open source OpenCV package.

More specifically, to construct a model that can distinguish between two classes each algorithm requires at least two sets of data, one from each class. These data sets are arbitrarily labeled as either 1 or 0, which corresponds to the result that is output by the testing phase. In the case of EEG data classification, the data sets consist of the previously mentioned power spectrum values for the five frequency ranges listed. These values are obtained from a data table stored in the EEG data repository and are assigned to either class 0 or class 1, based on the user's selection. The data is then used by the algorithm in question to train a new model, which can then be tested against other sets of data. The training phase of the classification process is very processor intensive. Because of this, almost all of the actual computation involved in the classification process is deferred to the server, leaving to the client only the task of displaying the results. This means that to construct a model or perform a classification test, the entire data set need not be transferred from the server at all. Instead, the data remains on the server, where the classification is performed, after which the results are communicated to the client.

The classification interface (Fig. 30.5) allows the user to first select a classification algorithm. Then, the user is prompted to select either "Train" or "Test," as desired. Based on this choice, the user is either given the option of selecting multiple tables from the database with which to construct a model,

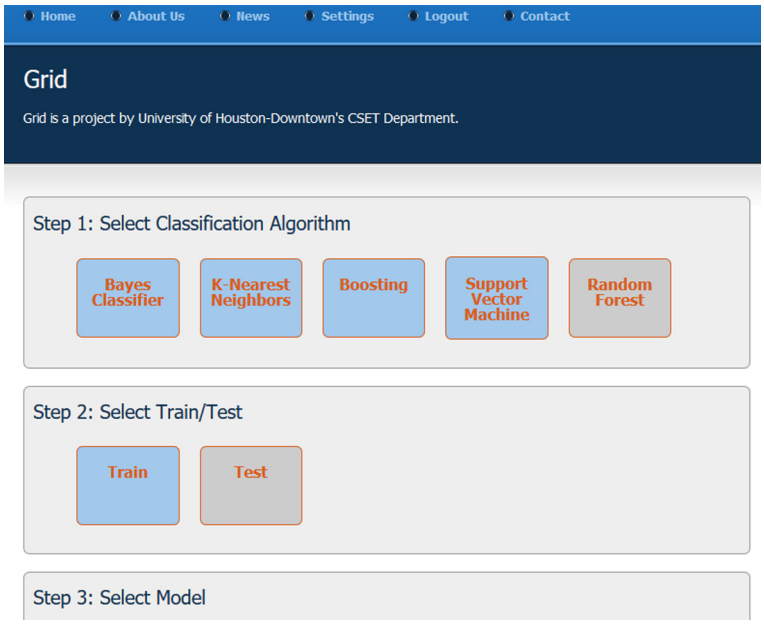

FIGURE 30.5

Algorithm selection pane.

or alternatively the option of selecting a preexisting model against which to test another table from the database. The training process then prompts the user for a name to give the new model, after which the model is trained and stored on the server. A record of this model is added to the database, so that it can be accessed in the future.

Fig. 30.6 describes the algorithm selection pane where the front-end interface is written in HTML, PHP, CSS, and JavaScript. PHP is responsible for the preprocessing and generation of the HTML pages, which use CSS to define the layout and graphical representation of the interface. Meanwhile, JavaScript is also used on the front-end for the purpose of live form handling and submission. Once the form is submitted, the server-side backend takes over. The user's selected options are passed to a PHP processing script that is responsible for executing the desired machine learning algorithm and passing the names of the chosen tables to the algorithms. The classification algorithms themselves are

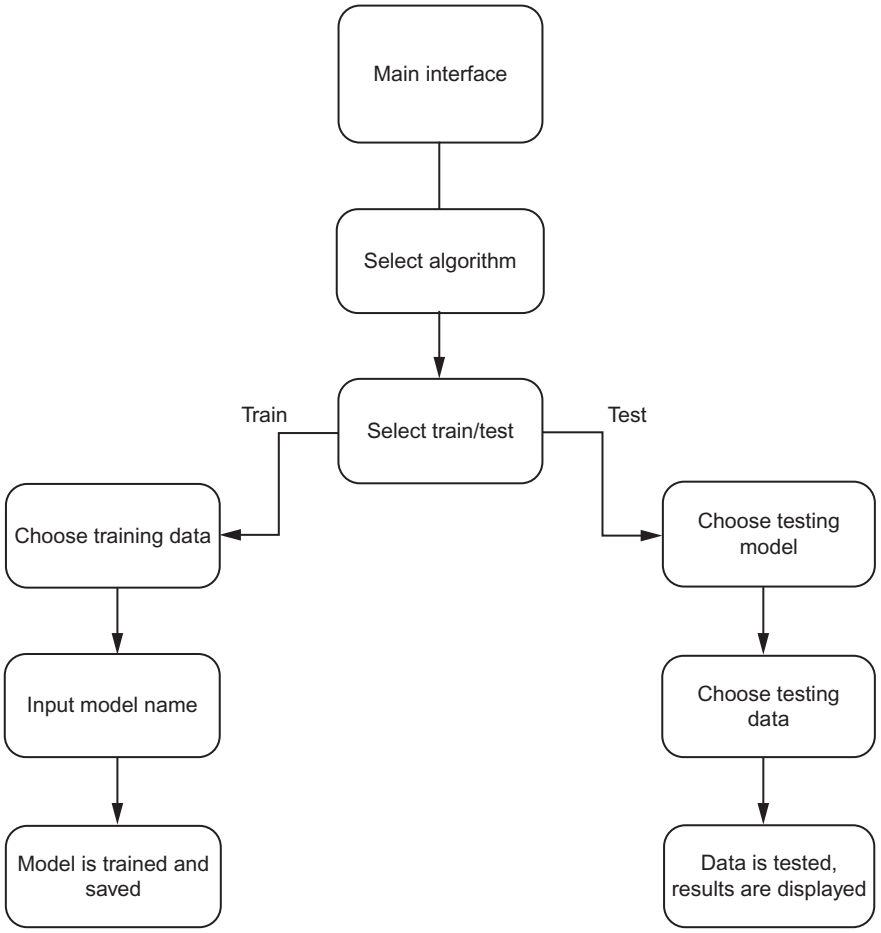

FIGURE 30.6

Algorithm selection pane model.

implemented in C++, and each is split into a separate training and testing executable. After the selected program is run, the results are reported back to the waiting PHP script, which encodes the values into an HTML page and sends it to the user's browser.

30.3 DATA ANALYSIS

The data was collected using the Neurosky Mindwave Mobile headset collected with the EEG ID Android application. For each test, the conditions were tightly controlled, in that each subject sat in the same location, in the same environment, for each sample collection. EEG data was collected for 5 min each in all cases, and so each individual sample consists of 5 min of "Active" data for playing video game, and 5 min of "Idle" data for relaxing with closed eyes. Component frequencies, including five major brain waves: delta (1–3 Hz), theta (4–7 Hz), alpha low (8–9 Hz), alpha high (10–12 Hz), beta low (13–17 Hz), beta high (18–30 Hz), gamma low (31–40 Hz), and gamma mid (41–50 Hz) were extracted from the raw data set using a feature extraction application provided by the Neurosky headset. These frequencies represent specific brain states including deep meditation and high anxiety.

30.3.1 DATA ANALYSIS ON RAW DATA SETS

The testing procedure involves varying not only the number of data sets included in the modeling phase, but also the percentage of each data sets used and the combination of different test subjects' data as presented in the architecture diagram (Fig. 30.7). There are three types of models in our study. The simplest type of model comprises only a single- subject's data, from a single collection time. A more interesting and complex case is when a model is constructed using data from multiple subjects. As the next logical step after single-subject models, we chose to train models that are composed of data from two different subjects (hereafter referred to as "bipartite" models.) Another case is a variation of the two-subject models, wherein the data is taken from the same subjects (homogenous), but the collection is from two separate times. Next, the number of subjects included in each model was then increased to three. We refer to this as a "tripartite" model. This was done to determine whether a good general model could be constructed. Additionally, we also constructed larger models, based on 12, 18, and 24 data sets, respectively. With these models, each of the three subjects provided a third of the data.

For each of these algorithms, an implementation was written in C++ using libraries from the OpenCV machine learning package. The implementations of these algorithms were designed to be run in two discrete phases: training and classification. The training phase takes at least one data file from each class as its input, and outputs a trained model. Similarly, the input for the classification or testing phase is a pre-trained model and at least one data file comprising the test data. The output of the classification phase is a confusion matrix, which shows how many samples from each class were correctly and incorrectly classified. Based on these results, a misclassification rate can be computed, or equivalently percentage values representing the portion of correct classifications. For each set common machine learning algorithms: KNN, SVM, random forest, Bayes, and boosting were chosen with an approach to include representative algorithms based on distance, tree, probability, and hypothesis, respectively.

The analysis of the result obtained on test data sets from same subject as well as different subjects showed that the random forest algorithm provides the highest percentage of correct classifications (Figs. 30.8 and 30.9). The highest accuracy level reached by random forest was 84%, which occurs

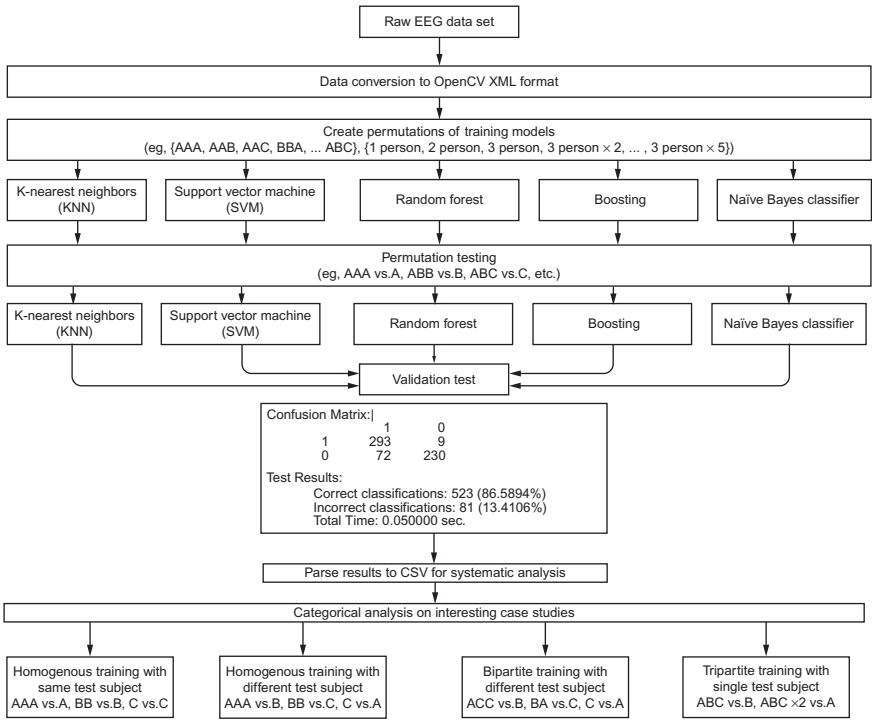

FIGURE 30.7

Architecture of systematic testing and analysis.

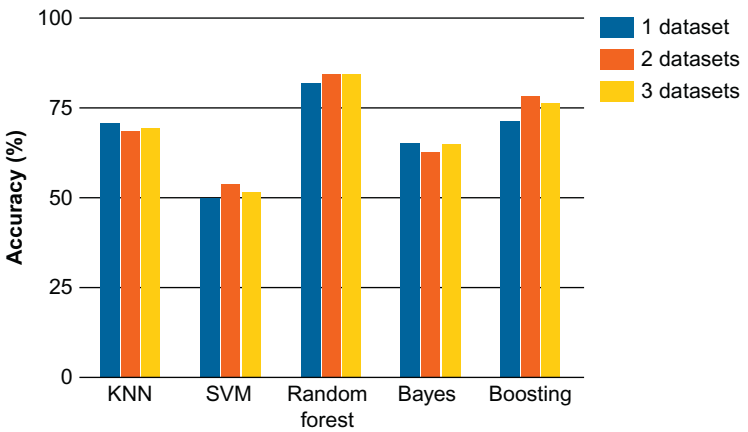

FIGURE 30.8

Test result comparison from same-subject tests.

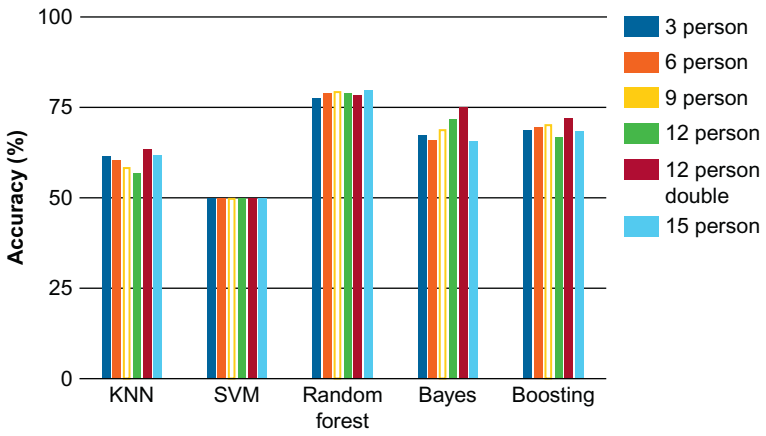

FIGURE 30.9

Test result comparison from tripartite models.

in the cases where a model was trained using three data sets from the same subject and then tested against data sets from the same subject. The results from our tests of the boosting method were the second most accurate, after random forest. Homogeneous training models incorporating data from two unique data sets seem to provide the highest accuracy rates (78%) for boosting, followed closely (76%) by the homogeneous models that were trained with three data sets.

In certain cases, the naïve Bayes classifier also produced relatively accurate results (Figs. 30.8 and 30.9). When using a tripartite training model composed of four data sets from each subject, the accuracy of Bayes reached an average of 75%. When compared to the 68% accuracy of the tripartite models trained with only one data set from each subject, it is tempting to conclude that the addition of more unique training data to a tripartite Bayes model will result in an increased accuracy; however, the accuracy level actually goes down when increasing from 1 to 2 data sets from each subject, and again when increasing from 4 to 5.

The accuracy of KNN ranged from 57% at the lowest, to 70% at the highest (Figs. 30.8 and 30.9). The lowest results were produced in the bipartite tests against a nonincluded subject, and also in the tripartite ($\times 4$) tests. KNN achieved the most accurate results in the single-person, single-data set tests. SVM consistently output results hovering around 50%, a rate that could be achieved by random guessing. Our suspicion is that SVM is not well suited to the classification of our brainwave data, possibly due to our selection of parameters. We tested the SVM algorithm with four different kernel types: linear, polynomial, sigmoid, and radial basis function. The variation of the kernel had no effect on the accuracy level of SVM's classifications, which remained firmly at 50% in all cases.

30.3.2 DATA ANALYSIS ON NORMALIZED MEAN

Another approach of understanding the structure and pattern of the data sets is by applying the machine learning techniques on inherited features of the data sets, such as the central tendency and deviation. Here, we compared the performances of the algorithms on normalized mean data sets alongside the raw

analysis. The data for each major brain wave was normalized to scale the data between 0 and 1, and its mean and standard deviation was used for the analysis. The normalized value of e_i for variable E in the ith row was calculated as

$$\text{Normalized}(e_i) = \frac{e_i - E_{\min}}{E_{\max} - E_{\min}}$$

where $E_{\min}$ is the minimum value for variable E and $E_{\max}$ is the maximum value for variable E.

The box plot (Fig. 30.10) depicts the numerical spread of the normalized data for the all brain waves for active and idle brain states. The combined data sets consisting of all five major brain waves show analogous data variability and the presence of few outliers amongst alpha and gamma waves. The outliers are eliminated for further analysis. The comparison of mean and standard deviation among different brain waves in idle and active states clearly shows that delta waves and beta waves have

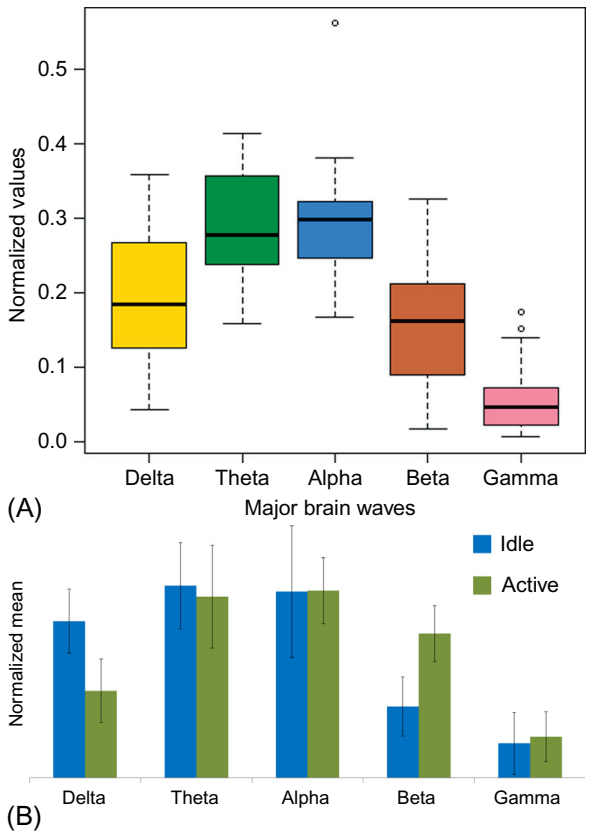

FIGURE 30.10

Visualization of normalized data sets using box plot for brain waves (A), comparison of mean and standard deviation among the brain waves in idle and active brain state (B).

Table 30.2 Correlation Analysis of the Major Brain Waves

	Delta	Theta	Alpha	Beta	Gamma
Delta	1				
Theta	0.53	1			
Alpha	−0.53	−0.50	1		
Beta	−0.75	−0.63	0.08	1	
Gamma	−0.36	−0.51	−0.07	0.35	1

significant differences (Fig. 30.10). Delta waves are clustered at higher values in idle states while beta waves are clustered at lower values, and vice versa in the active brain state. Correlation analysis corroborates that there exists a negative correlation between delta and beta waves during active and idle brain states (Table 30.2). This negative correlation can also be visualized using scatter plot (Fig. 30.11) where idle brain waves tend to cover higher x-axis coordinates while active waves were found more in higher y-axis coordinates.

Delta wave (0–3 Hz) is a dominant wave for idle state and beta wave (12–20 Hz) is a dominant wave of active state. The result also showed an interesting significant difference between these two waves. To further investigate this, we decided to focus on these two representative waves for active and idle state, respectively. Also, the understanding of the pattern of the dominant waves, delta and beta waves, for idle and active states will be an important step in deciphering the complexity of brain waves in different states. The density graphs (Fig. 30.11) further supports the idea that there exists a different pattern of delta and beta waves during active and idle states. To understand these dissimilar features of delta and beta waves, we decided to utilize the statistical machine learning techniques to explore any deep inherent differences.

A classification-based machine learning algorithm was implemented to survey the best algorithm to predict the brain states utilizing only delta and beta waves. The main challenge in this process was the problem of data separation for each brain wave at different brain states. Generally, the brain waves data from different states tend to cluster together, which becomes difficult for classification algorithms to draw a best fitting separation line. Because we were interested in comparing the performances of each algorithm in terms of the correct prediction rate, all 14 samples were used to train the algorithm and cross-validation technique was used to test the error rate of the model. Table 30.3 shows that out of the seven machine learning algorithms used, random forest showed a 4% error rate, while boosting and KNN with $n=4$, showed less than a 20% error. Similarly, SVM, naïve Bayes, neuralnet, and logistic regression showed higher error rates (Fig. 30.12).

The data analysis result demonstrates that the brain waves data performed better with tree-based algorithm random forest, while for other algorithms such as probability-based naïve Bayes and entropy-based neuralnet, the data did not achieve competitive error rate. Boosting and KNN model performed almost equivalently in predicting brain states (Fig. 30.12). KNN correctly predicted 24 out of 28 brain states while for SVM the number of incorrect prediction was slightly lower to 22 out of 28 brain sates (Fig. 30.13). These observations might be different for large sample sizes; however, the analysis does convey a finding that the tree-based model such as random forest learning algorithms are efficient in predicting brain states by analyzing EEG brain waves data for both raw and normalized mean data sets.

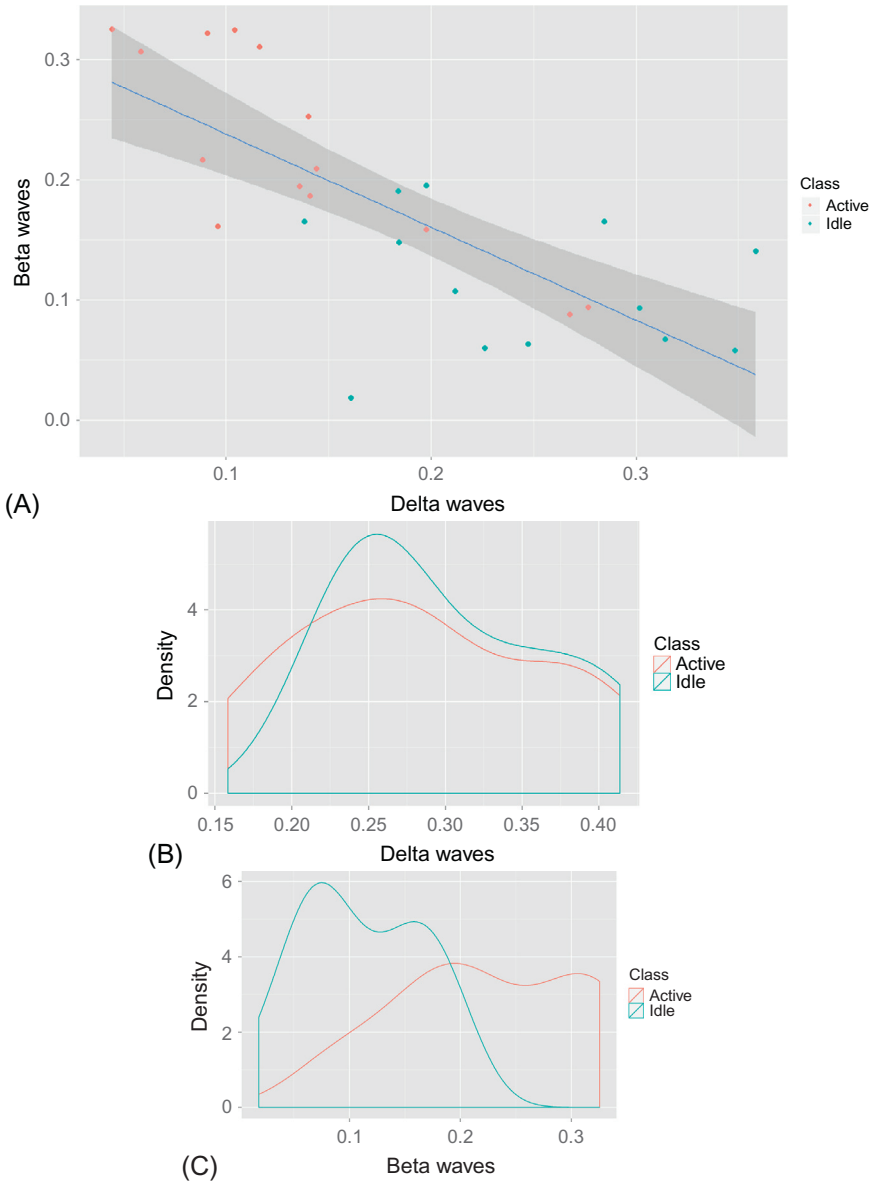

FIGURE 30.11

Scatter plot for major brain waves in active and idle brain states (A). Density graph of delta waves (0.1–3 Hz) (B) and beta waves (13–30 Hz) (C) during active and idle brain states.

Table 30.3 Summary of Classification-Based Prediction Scores

	Model							
	Train	**Logit Reg**	**KNN**	**SVM**	**Naïve Bayes**	**Random Forest**	**Boost**	**Neural Net**
Idle	14	11	13	11	11	14	13	13
Active	14	3	11	11	11	13	11	5
Error rate	0.00	0.50	0.14	0.21	0.21	0.04	0.14	0.36

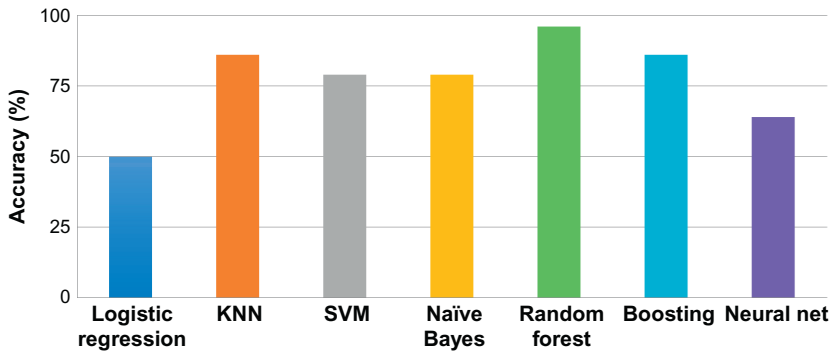

FIGURE 30.12

Comparison of common machine learning algorithms on test data sets.

Fig. 30.14 shows the comparison of all tested machine learning algorithms on normalized mean versus raw data sets. Random forest algorithm showed the lowest misclassification rate. This similarity in the performance of different algorithms with both raw and normalized mean data sets suggests that the pattern of brain waves are conserved from lower raw values to higher level of mean analysis. The overall higher values for normalized data sets might have resulted due to the test on model using the cross-validation technique on trained data sets, while for raw data sets separate train and test data sets were used. The inconsistency shown with the SVM algorithm requires further investigation as the accuracy of SVM depends to a great degree on the type of kernel used by the algorithm, and the values of the parameters passed to the algorithm during the training phase. A combination of a certain kernel and a set of parameters may have resulted in inconsistent misclassification rates.

In addition to the classification-based prediction model, it is also imperative to understand the natural structure and pattern of brain waves data in different states. One of the techniques used in analyzing organization of the data was clustering algorithm. Here, we used both *k*-means algorithm as well as hierarchical algorithm. The analysis showed that the performance of *k*-means and hierarchical clustering were similar (Fig. 30.15). Both algorithms assigned 22 out of 28 data points in the correct states while 6 of them were assigned to the wrong state. For *k*-means analysis, 28 iterations were used to incorporate all data points and 2 clusters were used to separate data points to assign each cluster to active and idle brain states. Similarly, for hierarchical clustering, dendrogram was drawn with two cuts to separate data points into two different states.

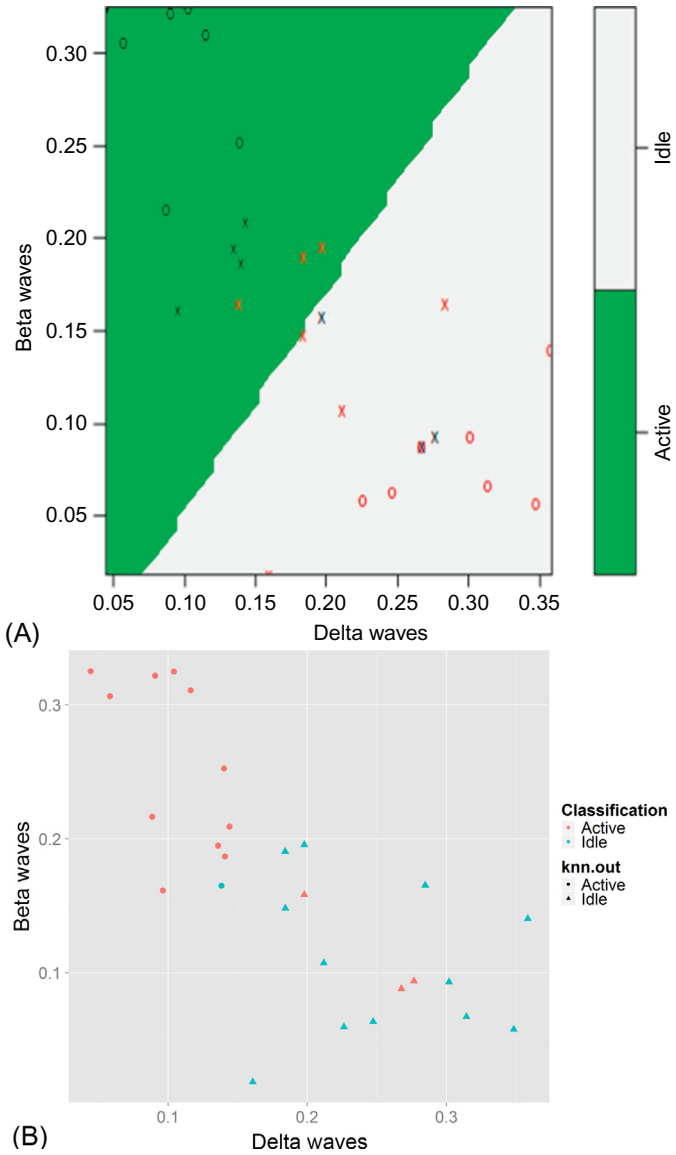

FIGURE 30.13

Prediction of active and idle brain states using SVM (A) and KNN (B) models.

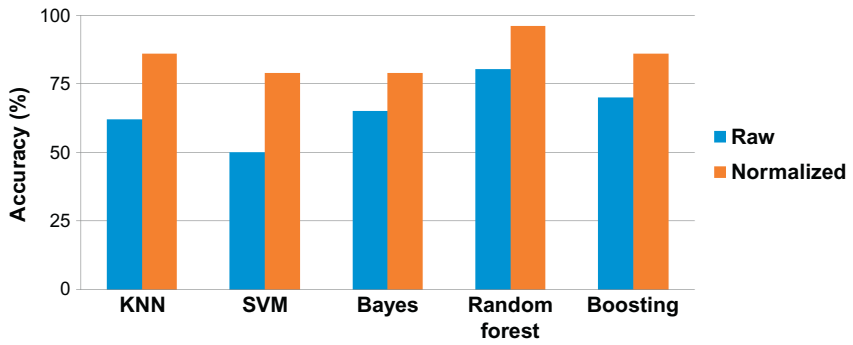

FIGURE 30.14

Comparison of algorithms performance on raw and normalized data sets.

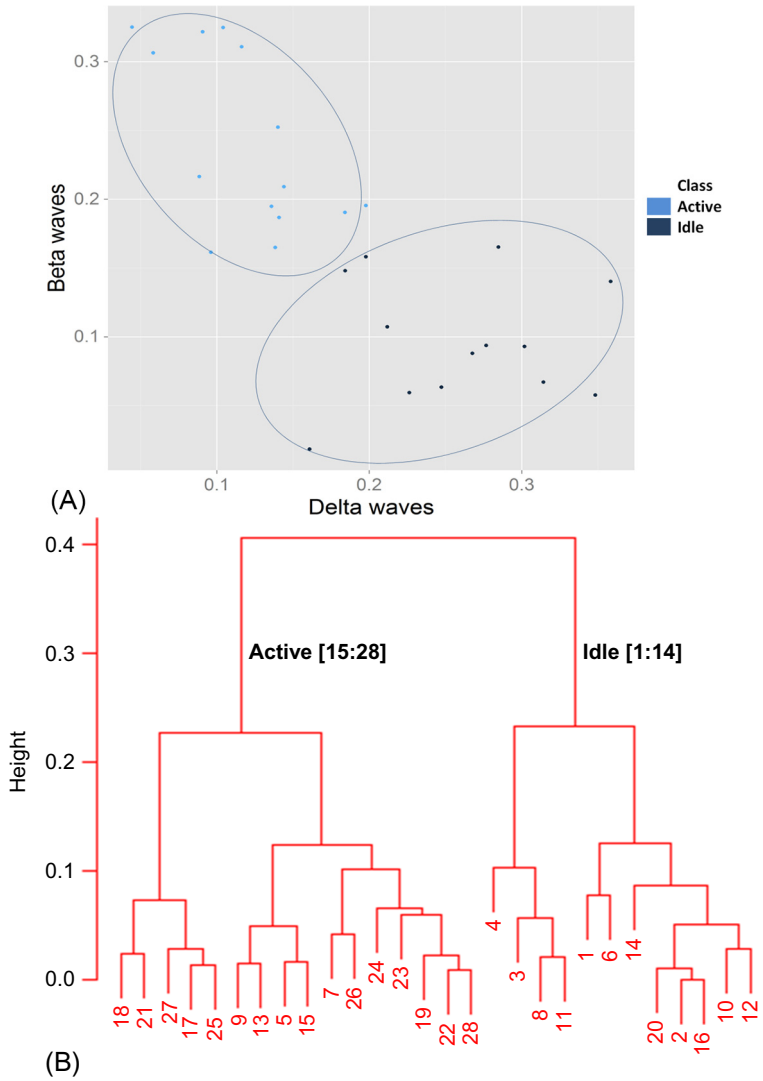

FIGURE 30.15

Cluster analysis: *k*-means (A) and hierarchical (B) for delta and beta brain waves.

30.4 CONCLUSION

In this study we provide a proof of concept for an EEG data analysis system using an EEG headset, mobile application, web server, and web interface with analytics. Additionally, the classification interface provides a way for users to test out various algorithms using different combinations of data for the purpose of constructing training models and testing new data sets on them.

Overall, we have shown that it is possible to capture EEG data from anywhere, while on the move, and be able to immediately see results, as well as delve deeper by performing analysis remotely. Although much more refinement is required in increasing the performance of the data capture and analysis, the system as built provides a good foundation for future improvements and expansions.

Additionally, we have identified an evident pattern of delta and beta waves in idle and active brain states. The finding suggests the importance of analyzing all major brain waves in each state because even though the theta and alpha waves are higher in both states, they did not show any significant difference, and the underlying pattern was observed between low-numbered delta and beta waves. The brainwave modeling experiment often focuses on higher values and single dominant waves; here the result highlights the fact that the inclusion of all major waves in the analysis could delineate the hidden pattern that exists between states. This unique feature of delta and beta waves among the major five brain waves is an interesting finding that could be implemented in recognizing brain states in future brain modeling research projects. Also, we have discovered that the random forest algorithm showed better performance with both raw and normalized mean data sets when compared to common machine learning algorithms used in EEG data analysis. This finding can now be implemented for large-scale data analysis using all major brain waves with increased sample size to conduct a real-time brain wave analysis on data sets utilizing parallel computing architecture.

30.5 FUTURE WORK

The system that we have developed currently shows great promise for future implementation of the additional machine learning algorithms. We are working on incorporating more of those analysis algorithms after developing and testing them first locally before inclusion on the website. We will also develop more visualizations methods for the website as necessary for different algorithms so as to represent results in the best visual appropriate. The classification interface will be expanded to allow the user to alter each of the relevant parameters. We also intend to further develop the mobile application to be more tightly integrated with the website.

The prospective brain waves modeling should design an inclusive system, which incorporates all the major brain waves, and addresses the variables such as specific regions of the brain, inconsistency within samples, limitations of the recording machine, and integrating knowledge form neurobiology in terms of understanding certain brain functions. Additionally, it is imperative for brain wave modeling studies to contemplate the rigorous time-series analysis of brain waves to decipher trends, irregularities, cycles, seasonality, and other variations among waves during different states. Therefore, an improvised advanced machine learning modeling system that includes all major brain waves rather than just dominant representative waves will be implemented on large data sets. The project aims to address the complexity of classification of brain waves data by modeling the major brain waves independently with clinically significant brain regions combined with the feature extraction and time-series analysis. This will achieve an efficient and predictable brain wave modeling system that has potential application in hospitality and clinical industry for self-controlled deep brain relaxation and early diagnosis of various brain abnormalities, respectively.

ACKNOWLEDGMENTS

This work is partially supported by NSF UGI-CSTEM (Grant# 0965952) and NSF Summer REU (Grant# 1262928) and National Computational Science Institute Blue Waters Student Internship program.

REFERENCES

[1] Kropotov J. Quantitative EEG, event-related potentials and neurotherapy. San Diego, CA: Elsevier; 2009.
[2] Larsen E. Classification of EEG signals in a brain-computer interface system [Ph.D. thesis]. Trondheim: Norwegian University of Science and Technology; 2011.
[3] Yang R, Song A, Xu B. Feature extraction of motor imagery EEG based on wavelet transform and higher-order statistics. Int J Wavelets Multiresolution Inf Process 2010;8(3):373–84.
[4] Zhuang T, Zhao H, Tang Z. A study of brainwave entrainment based on EEG brain dynamics. Comput Inf Sci 2009;2(2):80–6.
[5] Yuvaraj R, Murugappan M, Ibrahim N, Sundaraj K, Omar M, Mohamad K, et al. Optimal set of EEG features for emotional state classification and trajectory visualization in Parkinson's disease. Int J Psychophysiol 2014;94(3):482–95.
[6] Direito B, Teixeira C, Ribeiro B, Branco M, Sales F, Dourado A. Modeling epileptic brain states using EEG spectral analysis and topographic mapping. J Neurosci Methods 2012;210(2):220–9.
[7] Lin H. Measurable meditation, In: Proceedings of the international symposium on science 2.0 and expansion of science (S2ES 2010), the 14th world multiconference on systemics, cybernetics and informatics (WMSCI 2010), Orlando, FLorida, June 29–July 2; 2010. p. 56–61.
[8] Loizzo J, Peterson J, Charlson M, Wolf E, Altemus M, Briggs W, et al. The effect of a contemplative self-healing program on quality of life in women with breast and gynecologic cancers. Altern Ther Health Med 2010;16(3):30–7.
[9] Habermann T, Thompson C, LaPlant B, Bauer B, Janney C, Clark M, et al. Complementary and alternative medicine use among long-term lymphoma survivors: a pilot study. Am J Hematol 2009;84(12):795–8.
[10] Lengacher C, Johnson-Mallard V, Post-White J, Moscoso M, Jacobsen P, Klein T, et al. Randomized controlled trial of mindfulness-based stress reduction (MBSR) for survivors of breast cancer. Psychology 2009;18(12):1261–72.
[11] Oh B, Butow P, Mullan B, Clarke S. Medical Qigong for cancer patients: pilot study of impact on quality of life, side effects of treatment and inflammation. Am J Chin Med 2008;36(3):459–72.
[12] Lourenço A, Plácido da Silva H, Carreiras C, Alves A, Fred A. A web-based platform for biosignal visualization and annotation. Multimed Tools Appl 2014;70(1):433–60.
[13] Holmberg N, Wunsche B, Tempero E. A framework for interactive web-based visualization, In: AUIC '06 proceedings of the 7th Australasian user interface conference, vol. 50, JanurarySydney: Australian Computer Society; 2006.
[14] Poliakov A, Albright E, Hinshaw K, Corina D, Ojemann G, Martin R, et al. Server-based approach to web visualization of integrated three-dimensional brain imaging data. J Am Med Inform Assoc 2005;12(2):140–51. doi:http://dx.doi.org/10.1197/jamia.M1671.