Thesis

Detecting Error Related Negativity using EEG Potentials Generated during Simulated Brain Computer Interaction

Submitted by

Prathamesh Verlekar

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2014

Master's Committee:

Advisor: Charles Anderson

Jaime Ruiz
Patricia Davies

Abstract

Detecting Error Related Negativity using EEG Potentials Generated during Simulated Brain Computer Interaction

Error related negativity (ERN) is one of the components of the Event-Related Potential (ERP) observed during stimulus based tasks. In order to improve the performance of a brain computing interface (BCI) system, it is important to capture the ERN, classify the trials as correct or incorrect and feed this information back to the system. The objective of this study was to investigate techniques to detect presence of ERN in trials. In this thesis, features based on averaged ERP recordings were used to classify incorrect from correct actions. One feature selection technique coupled with four classification methods were used and compared in this work. Data were obtained from healthy subjects who performed an interaction experiment and the presence of ERN indicating incorrect responses was studied. Using suitable classifiers trained on data recorded earlier, the average recognition rate of correct and erroneous trials was reported and analyzed. The significance of selecting a subset of features to reduce the data dimensionality and to improve the classification performance was explored and discussed. We obtained success rates as high as 72% using a highly compact feature subset.

Acknowledgements

I would like to express my appreciation and thanks to my advisor Dr. Anderson for being a great mentor and for the several insightful conversations during the development of the ideas in this thesis and for providing constructive feedback on the text. I would also like to thank my committee members Dr. Ruiz and Dr. Davies for reviewing this document and serving on the defense committee. I would also like to thank the members of the BCI group for recording the data and the volunteers who spared their valuable time to provide their recordings. I am grateful to my family and friends for being a constant source of support, encouragement and backing. Finally, special thanks to the developers who created and now maintain tools like LaTeX, Python and Linux.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Advances in computer hardware and cognitive neuroscience have made possible the use of brain waves for communication between humans and computers. Researchers have used these technologies to build brain computer interfaces (BCIs) which are transmission systems that do not depend on the brain's normal output pathways of peripheral nerves and muscles. In these systems, subjects produce signals that can be used to control external devices. Patients incapable of producing any facial or limb movements now have a way to communicate with the outside world [1]. EEG is a medical imaging technique that reads electrical activity generated by brain structures by means of recording from the scalp surface using metal electrodes. But even with the latest techniques, these systems still suffer communication rates on the order of 2-3 symbols per minute [2]. One of the challenges is to accurately capture the complex neurological activity of the brain and to register error detection and response checking. In tasks which involve presentation of a stimulus or target discrimination, errors can be produced either by the subject or by the interface itself. Modern day BCI systems are prone to errors in the recognition of a subject's intent, and those errors can be frequent. Thus, it becomes important to identify the errors so as to improve the performance of these systems. This thesis focuses on identifying these error related potentials and by training classifiers on the recorded data, the accuracy of the BCI is measured.

1.1. Brain Computer Interfaces

In the past century, EEG has gone through enormous developments. Starting from the discovery of electrical currents in the brain by Richard Caton in 1875 to the first recording of brain waves by Hans Berger in 1924 to verification of brain waves in humans by Adrian and

Matthews in 1934, EEG's today can isolate the relative strengths and positions of electrical activity in different regions of the brain based on the state of the subject [3].

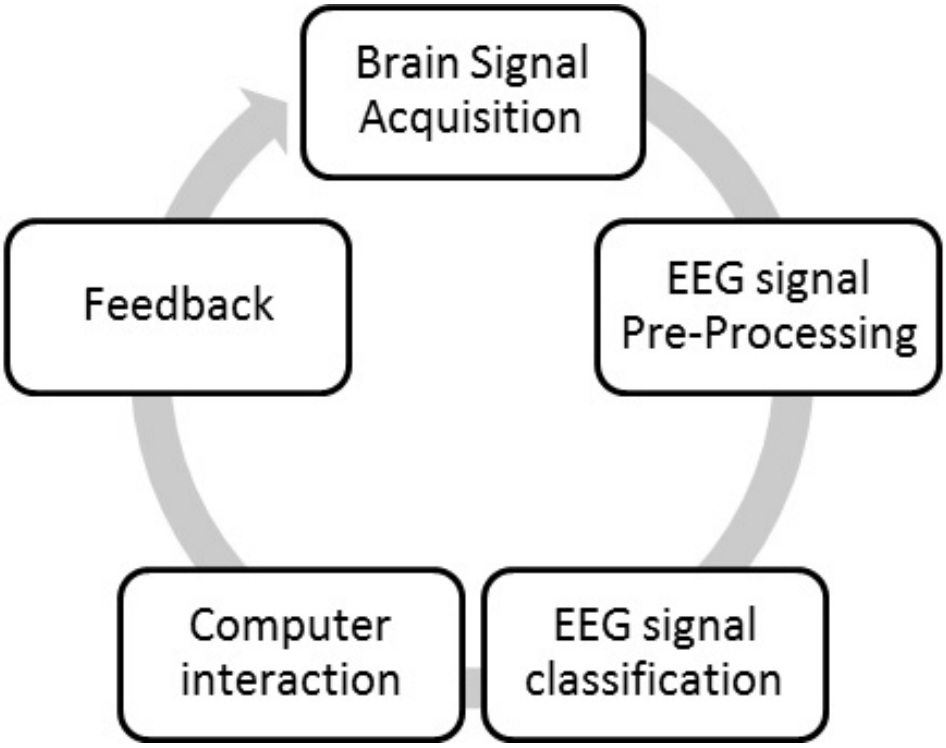

FIGURE 1.1. Common structure of a BCI system. The image is reproduced from [2].

This has led to experts exploring ways to record brain waves and they have come up with two prominent methods of capturing this information: invasive and non-invasive techniques. Since the non-invasive methods such as EEG are preferred due to avoidance of surgery and reduced risk of creating scar tissue around the implant, BCI has become very popular [4]. BCI is a technology which provides a pathway to directly communicate with the brain using an external device and provides the brain with a new control channel [5]. As a part of a non-invasive BCI technique, EEG is used to measure the electrical activity in the brain. The international 10-20 system is a recognized standard used for placement of electrodes on the scalp. The electrodes are made of metal and contact between the scalp surface and the

electrodes is established by means of a conducting gel. Figure 1.2 shows a placement map of 64 electrodes.

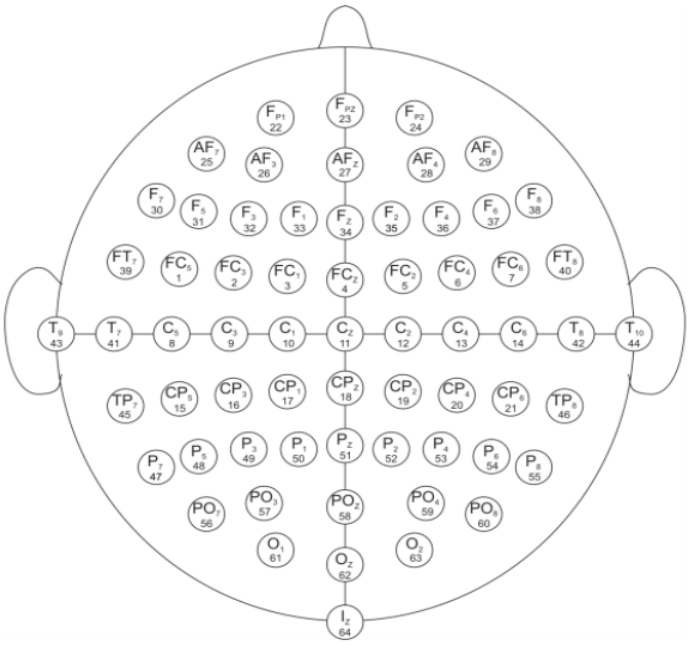

FIGURE 1.2. International 10-20 system for electrodes placement. The image is taken from [6].

Since BCI provides non-muscular communication, it imparts assistance to people with severe motor disabilities such as amyotrophic lateral sclerosis (ALS) which renders the person incapable of performing any physical activities. In addition to applications in the medical domain, BCIs have also been validated in many different areas, including spelling devices, simple computer games, environmental control, navigation in virtual reality and generic cursor control application [7, 8].

1.2. Limitations of State-of-the-Art

Recording and analyzing EEG using BCI techniques is a challenge due to various reasons. EEG is meant to record cerebral activity but it also captures activities from other

unaccounted sources which are termed artifacts. These interference waveforms are electrical potentials not originating in the brain and they often cause serious misclassifications thus reducing the clinical usability of such automated analyzing systems [2]. Isolating and eliminating artifacts in real time EEG recordings is a complex but essential task to the development of practical systems. This includes physiologic artifacts arising from muscle activities like clenching of jaw muscles, eye blinks, eyeball movements, movement of tongue, respiration and heart beats to a certain extent. Extra-physiologic artifacts essentially arise from external sources like equipment or environment. Improper contact of electrode with the scalp, impedance of electrodes exceeding that of ground due to alternating current, interference between adjacent electrodes, movements of other people around the subject, presence of high frequency devices such as TV, cellphones around the recording equipment or photic stimulation cause severe degradation in the quality of the captured EEG [9]. EEG is also associated with poor spatial resolution. The scalp electrodes primarily detect currents originating from superficial layers of the cortex directly underlying the skull. Currents that are not oriented to the scalp are not picked up by the EEG. Thus, neurons buried within the deep brain structures have far less contribution to the EEG signal [10].

Interfering neuropsychological activities of the brain also make recording EEG a difficult proposition. BCI activity depends on two adaptive controllers: the subject who interprets the stimuli and generates brain waves which is provided as an input to the BCI, and the computer itself which processes this information. Different cortical areas of the brain are activated for different tasks and it becomes difficult to recognize the differences between these activities. Additionally, the human brain is a stateful unit, always correlating the current activity to the preceding one. Although the BCI correctly isolates and identifies a particular

kind of activity, its inability to differentiate between the actual task related response and the overall compounded signal still acts as a barrier to obtain optimal EEG recordings.

1.3. Approach

Several studies have discussed boosting the performance of humans already functioning at a normal level using therapeutic BCI technology [5]. In visual stimulus based studies, it is very difficult to reach a level of 100% success even with well trained patients [11]. The human brain has a very distinctive property of behaving as a dual carrier of information pertaining to transmission of control signals to command a device or respond to stimuli and evaluation of cognitive states required for a meaningful interaction within a short span of a millisecond. This includes the state where the brain is aware of erroneous responses which in turn is used to improve the performance of a BCI system. The approach used in this thesis is primarily directed towards identifying the error related potentials generated by the brain and then classifying the incorrect and correct trials using suitable machine learning algorithms. This information can then be used as feedback to train the BCI system to better predict the future instances of error signals. Data are obtained from visual stimulus based experiments specifically designed to capture the erroneous behavior of the brain; it is then pre-processed to eliminate any artifacts and smoothed to remove noisy waveforms. Data are segmented based on the onset of the stimulus to construct a matrix of trials consisting of features represented by individual time points. After obtaining the processed data, various machine learning algorithms are used to fit a model on the training data and using an optimal combination of tuning parameters for these algorithms, classifiers are trained to make an attempt to successfully predict the number of incorrect trails which will help us isolate the occurrences of these error related potentials.

1.4. Overview of Thesis

In Chapter 2, we talk about Error Related Negativity (ERN) as a component associated with recognition of incorrect responses and how it can be used as a degree of correctness in a stimulus based task. It also includes other related work done in analyzing ERN for automatic detection of errors. In Chapter 3, we begin by describing the data used for the thesis along with the protocols used to capture this data. This section also covers the algorithms used for classification of ERN data and their significance. In Chapter 4, we study the outcome of the classification techniques on the data for various subjects. We analyze the results and discuss the importance of capturing the ERN information to improve the accuracy of the BCI system. Chapter 5 summarizes all the results and explores any opportunities for further improvements and considerations.

CHAPTER 2

Background

Interest in cognitive feedback, behavioral monitoring and application of EEG to record certain ERP's for choice reaction time tasks was intensified by the discovery of a component called ERN. The ERN/Ne was originally theorized to represent the activity of a generic monitoring system that detects errors by signaling a mismatch whenever comparisons between the response and the outcome of response selection yield different results [12]. Others suggested ERN to be a response of a conflicting decision. Later, it was discovered that ERN emulates something more than a mental process; it might capture error made by the subject, it might detect a conflicting scenario or some other process involving transmission of a feedback signal in case of an error. This chapter describes in detail the significance of ERN and some state-of-the-art methods used in detection and analysis of ERN. This will lead us to our problem statement for the thesis.

2.1. ERN

A number of groups have recently started exploring a crucial piece of information that deals with awareness of erroneous responses in typical choice reaction tasks. It was observed that when a subject is asked to respond quickly to a visual stimulus, the resulting signal is an error potential (ErrP) primarily due to subject's incorrect motor action [11]. ERN was first observed in accelerated choice reaction time tasks where it was captured as a difference between the correct and error trials. ERN/Ne is a component of the event-related potential (ERP) related to acknowledging erroneous responses in stimulus based visual tasks [13]. It is observed as a negative waveform peaking in the range of 50 to 150 ms after the occurrence of the incorrect response [14]. The ERN occurs on error trials in a wide variety

of speeded-response tasks involving visual, auditory and tactile stimuli, and unimanual, bimanual, foot, oculomotor and vocal responses. It may even be elicited by auditory, visual, and somatosensory error feedback stimuli and by losses in gambling tasks [15].

Studies have found the scalp distribution of the ERN to be maximal at mid-line fronto-central locations, most typically the 10-20 location FCz and has been associated with the anterior cingulate cortex (ACC) as the most probable neural generator site of ERN [16]. Research associates activation of ACC region to tasks involving detection of errors, regulation of emotions and decision making [17]. Because an error is a noticeable indicator signifying disruption of performance, the ERN generally reflects a process involved in evaluating the need for implementing control. Figure 2.1 shows an instance of ERN being observed in recordings of Subject 1 in our study.

ERN could arise from multiple sources which makes it necessary to understand other components related to it in order to study its functional significance. The first one is a positive response locked waveform known as error positivity (Pe) which has a maximum amplitude between 200 and 400 ms after an erroneous response [18]. The most commonly used Pe is the P300 waveform associated with error detection. Another stimulus locked waveform observed is the N200 usually seen in conflict tasks like the Eriksen flanker test [19]. In conflict tasks involving verbal stimuli, a delayed instance of N200 is observed known as N450 [20]. Correct response negativity (CRN) is a waveform mostly observed on correct trails rather than error trails. CRN occurs because the brain labels a response as an error that is not an error according to the subject [21]. And lastly, the feedback related negativity (FRN) is reported to be observed in button press and gambling tasks which involve error-feedback stimuli. FRN occurs approximately 250-300 ms following a feedback stimulus [2]. The presence of these components raises questions such as whether ERN includes feedback

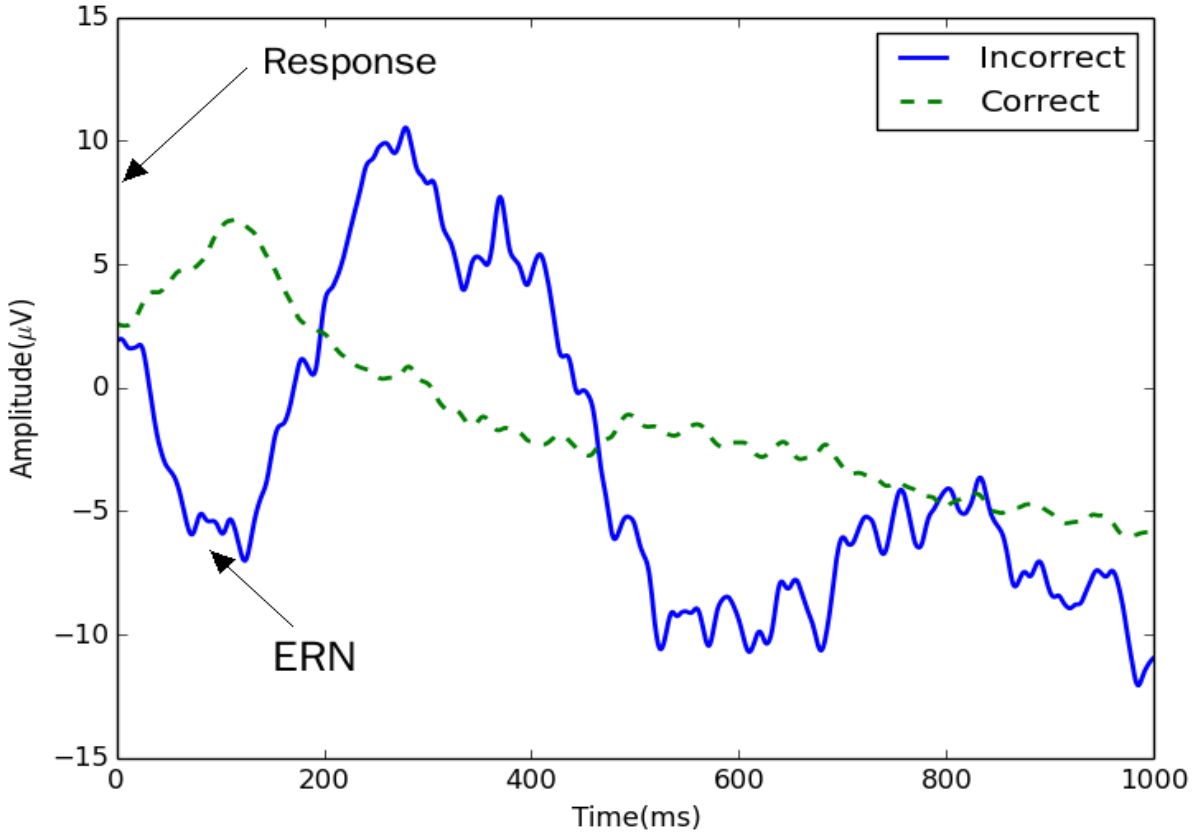

FIGURE 2.1. Average ERP for Subject 1 with correct and incorrect responses. The ERN is marked by a negative deflection around 50-100 ms after the response.

related components, is ERN associated with correct trials and whether ERN is a multiple peak oscillation. A lot of research has been performed over the past few decades to answer these questions which are elaborated in the next section.

2.2. Related work

One of the most reliable signal features of EEG is the P300 evoked potential, which is a positive peak in the signal amplitude at about 300 ms after a stimulus is given to the subject. A number of BCI systems using P300 have been proposed in the literature, including the P300 speller. In the P300 speller system, the subject is shown a matrix of letters, and is told to focus at one of the letters on the grid that he/she desires to choose. The rows and

columns of letters are flashed in a random order. If the letter that the user is looking at flashes, P300 is generated about 300 ms later. We can thus train a classifier to detect this P300 to identify the desired letter [22]. For patients with locked-in syndrome, the P300 evoked in scalp-recorded EEG by external stimuli has proven to be a reliable response for controlling a BCI for some subjects [23]. A notable study performed by Sutton, et al. [24] regarding the significance and neural origins of P300 concluded with a major outcome citing issues related to increase in the number of components identified and to the problem of deciding which components are being dealt with in a particular experiment. This resulted in large scale research on determining the functional significance of ERN and its related components.

Several influential theories have been proposed regarding generation and computation of ERN. These theories combined with state of the art techniques helps capture the patterns required to analyze the error potentials. The *error-detection theory* states that ERN reflects a process that compares the output of the motor system to the best estimate of the correct response at the time of the ERN occurrence [13]. In choice reaction based tasks, an error usually occurs because the subject responds before evaluating the stimulus completely. A comparator computes the difference between the representation of the correct response and a representation of the ongoing response. A discrepancy between these two instances results in a mismatch or error signal. Figure 2.2 describes this approach in detail.

The *conflict monitoring theory* was proposed as a counter to the error detection theory and stated that in a typical choice task, concurrent activation of response signals indicate presence of a conflict which can help track the accuracy without knowing the correct response [26]. The two major claims of the conflict monitoring theory are that the ACC monitors for the occurrence of *response conflict* - concurrent activation of multiple competing responses

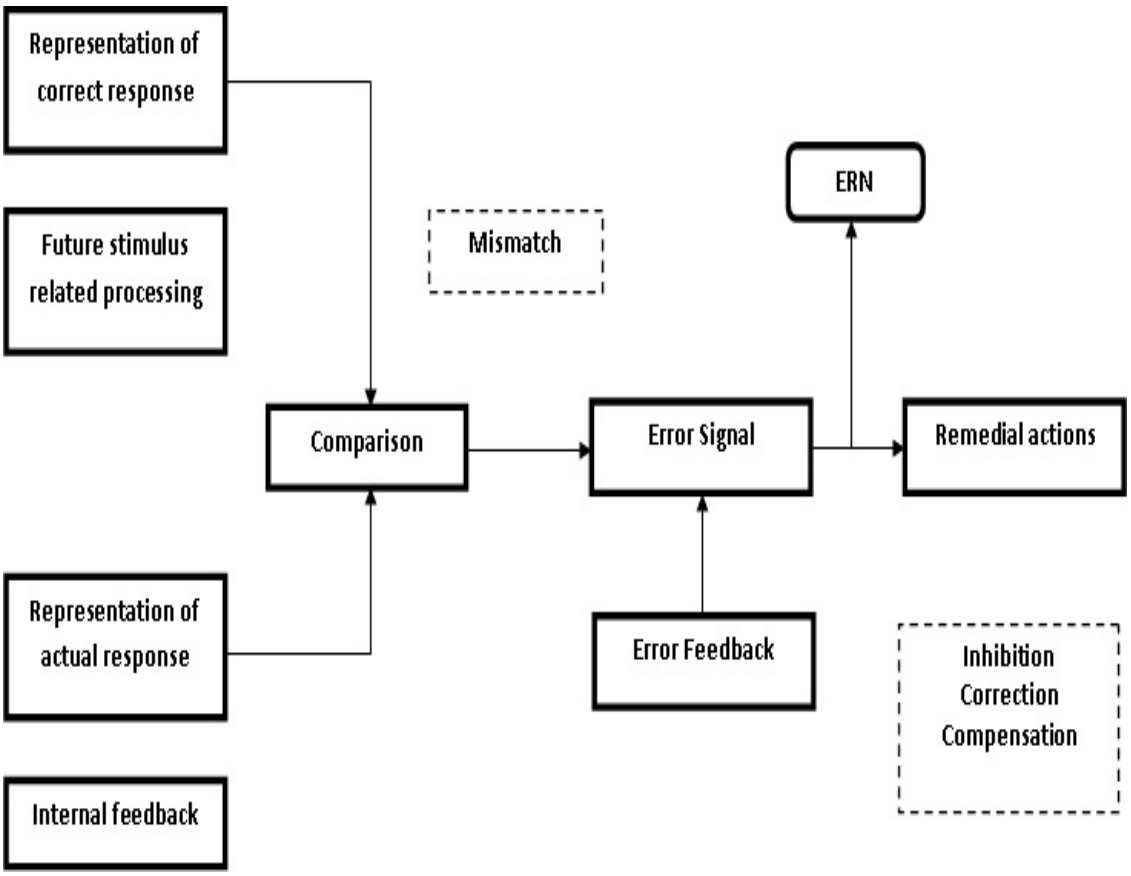

FIGURE 2.2. Schematic diagram illustrating the error-detection theory of ERN. The image is reproduced from [25].

and that it uses this information to signal the need for increased control. Conflict-monitoring model has been widely used with the Eriksen flanker task and the Stroop task to analyze the ERN.

Another very important theory used widely in building artificial intelligence models is the *Reinforcement Learning Theory of the ERN* [27]. According to this theory, when events worse than a specific threshold occur, the basal ganglia produces an error signal based on prior reinforcement signals associated with a response. This error signal is then transmitted to the ACC which is used to improve task performance by altering the manner in which the motor system of the brain is controlled. Hence, in the early learning stage, the brain separates success and failure by means of external feedback. After continuously learning over

a period of time, based on the stimulus and the reward, errors can be detected immediately when a response occurs, without the need to wait for external feedback [28].

Recent studies have shown that in the due course of the experiment, subjects become aware of their mistakes and this aspect manifests itself in the form of an ERN. The *affect/motivation* theory discusses the dynamic link between cognition and emotion and their role in working together to process information and execute a particular action [29]. The ERN reflects an affective response to an error which helps to understand the links between emotion related variables and the ERN amplitude which in turn provides evidence to the fact that affective traits could influence the cognitive processes. These traits have been observed in various patient studies dealing with effects of neurological and psychiatric conditions on the ERN. Studies involving detection of ERN in patients with obsessive-compulsive disorder (OCD) have attributed the exaggerated and repetitive behaviors that characterize OCD to a hyperactive error signal which might manifest itself as an ERN [30]. Studies using questionnaires have reported consistent elevated ERN amplitudes associated with negative affect and anxiety. Finally studies involving monetary incentives suggest that the amplitude of the ERN component increases when monetary incentives are offered for accuracy [12]. But on the contrary, recent studies have found no differences in the ERN for errors associated with low or high penalties, suggesting that the size of a penalty has no impact on the ERN/Ne.

To prove these theories, various paradigms have been used to investigate the presence of ERN. The most classic one is the flanker test which was used to investigate factors affecting selective attention and the extent to which processing of irrelevant information occurs. Eriksen and Eriksen introduced a variant of the flanker test in 1974 which is widely used to capture responses of a subject on a choice reaction task [19]. The test involves identification of a central target character in the presence of surrounding distractors. Usually, arrowheads

or alphabets are used. Arrowhead sequences (←←←←← indicating left or →→→→→ indicating right) or (HHHHH indicating H or SSSSS indicating S) are termed as *congruent* sequences while(←←→←← indicating right or →→←→→ indicating left) or (HHSHH indicating S or SSHSS indicating H) are categorized as *incongruent* sequences. This paradigm examines to what extent irrelevant information is processed in a visual task. It is observed that response times are lower for congruent sequences. During trials involving incongruent sequences, ERN is generated if the brain makes a mistake while identifying the central character correctly. Few studies modify this task to examine age related effects or other neurological indices.

Another task widely used to capture ERN is the Stroop task in which the participants are presented with words and are asked to name the color in which these words are presented [31]. During incongruent trials (i.e., blue printed in red), there is a high degree of response conflict, which signals the need for deliberate control because the relatively automatic behavior (word reading) must be inhibited in favor of a less automatic behavior (naming the color of the ink). In this task, participants exhibit executive control when they override their automatic impulse which results in error potentials during conflict situations.

2.3. Problem statement

After gaining a brief insight about the various theories used for ERN detection, it is evident that capturing the occurrence of ERN is of prime importance to improve the performance of the BCI system. By employing a proper paradigm, tasks were designed to create clashing scenarios in the decision making process of the brain resulting in generation of ERN. Capturing the ERN and analyzing its nature will provide an insight into the error response behavior of the brain and help in classifying the incorrect responses better. The

objective of this study is to identify the presence of ERN in visual stimulus based tasks and to use the underlying patterns in the captured ERN data to classify incorrect trials from the correct ones using the most favorable combination of machine learning algorithms, choice of electrodes linked to predominantly typical ERN generator locations, amount of smoothing applied to the data without compromising its nature, the number of data samples used to train the classifier and the number of features to obtain better classification accuracy.

CHAPTER 3

Methods

To capture the erroneous behavior of the brain accurately, it is important to design appropriate experiments with the objective of intermittently forcing the brain to make a mistake and trigger an error signal in response. It is this error signal which the classification algorithm will make use of to identify incorrect trails from the correct ones. After the data is obtained, it is necessary to pre-process it to get rid of the noisy components, select a set of specific electrodes to work with and convert it into a format suitable for applying machine learning techniques. This chapter explains the protocols followed, the paradigm used to investigate presence of ERN and the various technical specifications enforced to record the data used for the experiments. The effect of smoothing the data using a regularization filter is discussed. Also, the different algorithms employed in this study are explained in detail.

3.1. Experimental setup

The approach used to collect the data targeted towards identification of ERN was the Eriksen flanker test. The visual stimuli was a sequence of five characters containing a combination of the letters H and S. The procedure involved presentation of 480 different trials divided into two sets of 240 trials. The two sets of trials were separated by a time interval of 15 minutes. A batch of 480 trials consisted of 80 trials each of congruent sequences (HHHHH and SSSSS) and 160 trials each of incongruent sequences (SSHSS and HHSHH) respectively. The participants were asked to focus on the central character and press the corresponding key upon presentation of the stimulus. The subjects were 20 to 28 year old able bodied individuals. A sequence of characters persisted on the computer screen for a duration of 250 ms with a fixed value of stimulus onset asynchrony (SOA) initially set at 1200 ms. SOA is

defined as the amount of time between the start of one stimulus and the start of the next stimulus. Based on the error rate, the SOA was dynamically adjusted after a group of 30 trials for each subject. The SOA was reduced by 100 ms if the error rate was less than 15% where the minimum SOA was 800 ms whereas the SOA was increased by 100 ms if error rate was greater than 30% [14].

EEG signals were recorded using the BioSemi system. Thirty-three electrode sites were used, based on the 10/20 international system of electrode placement. The left earlobe acted as a reference while the right earlobe was an active site. Offline, all data were referenced to the average of the left and right earlobes and sampled with a band-pass filter of 0.23 to 100 Hz. The EEG was segmented for each trial beginning at the onset of the stimulus and continuing for 1000 ms. For each subject, correct and error trials were stacked up vertically to form the input data matrix where each segment consisted of time samples from the 1000 ms time window. Each of the time points in this matrix were considered as features for this study. The labels were obtained from the stimulus channel per segment as incorrect or correct based on the type of flashed sequence and they were assigned to the segmented data trials. For the purpose of this thesis, ERN data from three subjects was used. Table 3.1 describes the datasets obtained after segmenting the signals.

TABLE 3.1. Description of the data obtained by segmenting the recordings for 3 subjects. Datasets A and B indicate the two sets of data obtained from a single subject by conducting two sets of trials on the same day

Subject	Dataset	Correct Trials	Incorrect Trials	Total Trials
Subject 1	A	222	16	238
	B	237	03	240
Subject 2	A	226	13	239
	B	224	11	235
Subject 3	A	214	25	239
	B	213	25	238

While the data from thirty three different sites was recorded for each of these subjects, data from the Fz and Cz electrodes were analyzed for this work. The experiments were designed in such a manner that the number of incorrect trials in data set A were less than or equal to those in data set B. There were only 3 out of 240 trials in the data set B for Subject 1 which were marked as incorrect while in the data set B for Subject 2, 11 of the 235 trials were labeled as incorrect. To reduce the effect of noise amplification occurring during data acquisition, data was pre-processed by passing it through a regularization filter [32]. In this method, a quadratic term is used to find a curve by balancing the fit to the original data with the smoothness of the curve. Since a second order derivative gives curvature of a function, constraining this term to be small in value makes the curve smooth. This filter tries to find a balance between the goodness-of-fit term and the factor which controls roughness by varying the regularization parameter λ. The optimal values of λ depend on the data trend, variance, number of points, and the order of the smoothing derivative used. The value of λ was varied to obtain the best smoothing and this data was used for further analysis. A comparison of the original and the smoothed signal of Subject 1 for correct and incorrect trials using the electrode Cz is shown in Figure 3.1.

From the Figure 3.1, it can be seen that the incorrect response has more variation compared to the correct response. This can be attributed to the fact that the trials labeled incorrect were very few in number in comparison to the trials marked as correct. As a result, due to fewer trials labeled incorrect were averaged, it's response showed more fluctuation than the average response of the trials marked as correct. A similar graph showing the comparison of the original and smoothed signals for Subject 2 using Cz electrode is shown in Figure 3.2.

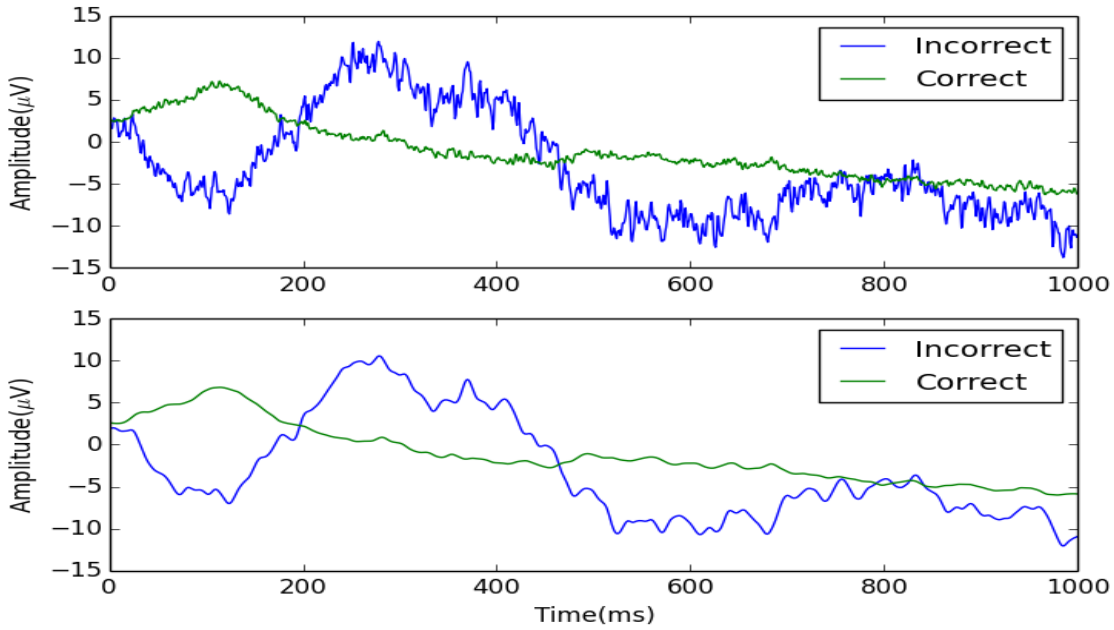

FIGURE 3.1. The average correct and incorrect responses for channel Cz before and after smoothing for Subject 1.

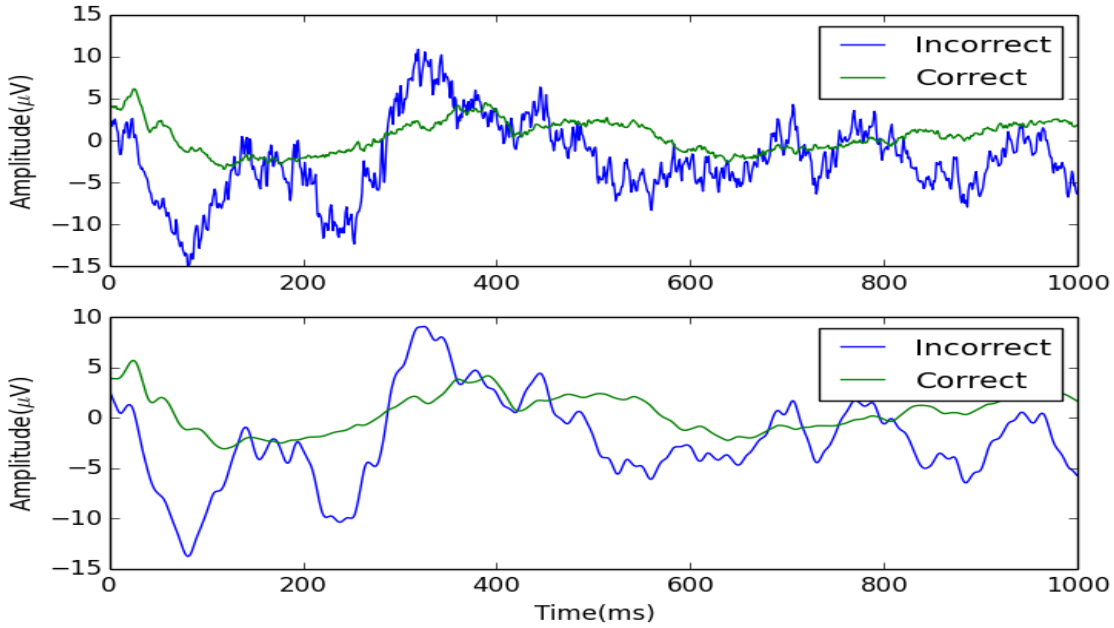

FIGURE 3.2. The average correct and incorrect responses for channel Cz before and after smoothing for Subject 2.

3.2. Algorithms

We have used four algorithms for the classification of ERN data: Linear discriminant analysis, quadratic discriminant analysis, neural networks and support vector machines. We have also used a ranking based feature selection technique called recursive feature elimination to work with a reduced feature based data set. These techniques are described as follows.

- LDA and QDA:

  Linear discriminant analysis (LDA) and Quadratic discriminant analysis (QDA) belong to a method of classification called generative modeling where we try to estimate the within class density of the feature vector given the class label. Combined with the prior probability of classes, the posterior probability of the class labels can be obtained by the Bayes formula. LDA tries to find a linear combination of features to separate classes of trials while QDA is a more generic version of LDA in which classes are separated by a quadratic function. The only major difference between LDA and QDA is that LDA assumes the Gaussian distributions for different classes share the same co-variance structure whereas this constraint is not present in QDA.

  If the feature vector is X and the class label is Y, then according to Bayes rule, the optimal decision on Y is to choose a class with maximum posterior probability given X provided we have the joint distribution of X and Y. If the prior probability for class k is given as π_k, $\sum_{k=1}^{K} \pi_k = 1$, then the posterior probability is

$$P(C = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \tag{3.2.1}$$

For LDA and QDA, we assume that the density for X provided every class k follows a multivariate Gaussian distribution given by

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-1/2(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)} \tag{3.2.2}$$

where p is the dimension, Σ_k is the covariance matrix and μ_k is the mean of the Gaussian distribution. For LDA, we assume for different k that the covariance matrix is identical due to which the classifier becomes linear. For optimal classification using the generative modeling approach, we maximize the product of the prior and the within class density to obtain the final classifier given by

$$\hat{C}(x) = arg \max_k \left[x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \right] \tag{3.2.3}$$

If we assume that the covariance matrix can be different for each class, we can estimate the covariance matrix Σ_k separately for each class $k = 1, 2, \ldots, K$. This changes the discriminant function to a quadratic one for QDA as shown below

$$f_k(x) = -\frac{1}{2} \log |\pi_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log(\pi_k) \tag{3.2.4}$$

Since QDA allows more flexibility for the covariance matrix, it tends to fit the data better than LDA, but then it has more parameters to estimate.

- Neural Networks:

  Classifying observations using prior information is the most popular application of neural networks. The input, hidden and output neurons are connected together by a feed-forward structure with signals flowing from inputs, forwards through any hidden units, eventually reaching the output units. Classification process usually involves three stages: propagating the signal to activate the artificial neurons, training the network using a specified technique (e.g., back propagation gradient) and then testing the network for classification accuracy [33].

  In a typical neural network, the input layer comprises n neurons of input signal $X = (X_1 \dots X_n)$, with the number of neurons of the hidden layer being chosen empirically. The output layer comprises K neurons for the K classes with the output being $Y_k \in [0, 1] \; \forall \; Y_k, k = 1, \dots, K$. Each connection between two neurons is associated with a weight factor which is modified by successive iterations during the training of the network according to input and output data. For a classification task, the neural network learns a discriminant function which separates the classes. If the classes can be separated by a straight line, the classes are said to be linearly separable which can be learned by neural networks without any hidden units. If a non-linear function is required to ensure class separation, multiple hidden units are used. The number of hidden units can be varied to adjust the non-linearity of the model. If there are no hidden units, the model becomes a simple linear logistic regression model over input data.

  To train a neural network, following process is followed:

  (1) The input data is fed to the network and propagated until it reaches the output layer which produces a predicted output.

(2) The difference between the actual output and the predicted output(error value) is evaluated.

(3) The neural network then uses a learning algorithm for adjusting the weights e.g., backpropogation.

(4) The forward process is repeated until the error between predicted and actual outputs is minimized.

To classify a new pattern after training, we use the *maximum likelihood* discriminant. We try to select weight w using maximum likelihood on a training set. Thus, we try to identify which selection of w was most likely to generate the labels in the training set. We employ negative log-likelihood as the objective function to minimize. This negative log-likelihood objective function is defined as

$$LL(w) = -\sum_{i=1}^{N}\sum_{k=1}^{K} T_{ik} \log f_k(x_i)$$

where w is the weights of the network, T_{ik} is the k^{th} component of indicator variable for training sample x_i and $f_k(x_i)$ is the output of the network defined by

$$f_k(x_i) = \frac{e^{w_k^o.Z}}{\sum_j e^{w_j^o.Z}}$$

where w^o are the weights of the nodes in the output layer and Z are the outputs of the hidden layer corresponding to the input x_i.

The final classifier output is given as

$$C(x) = \arg\max_k f_k(x)$$

The activity of neurons in the hidden layer is determined by an activation function represented by the hyperbolic-tangent function. For testing, the input data is fed into the network and the desired values are compared to the network's output values. The performance of the trained network is judged on the accuracy of the classification results [34]. We have used a gradient descent method called Scaled Conjugate Gradient (SCG) to minimize the objective function [35].

- Support Vector Machines:

  Support Vector Machines (SVMs) are a popular machine learning method for classification, regression and other learning tasks. SVM algorithm developed by Vapnik is based on statistical learning theory [36]. In classification, we try to find an optimal hyperplane that separates two classes by minimizing the norm of the weight vector w, which defines the separating hyperplane which is equivalent to maximizing the margin between two classes. Therefore, the objective is to choose a hyperplane with small norm while simultaneously minimizing the sum of the distances from the data points to the hyperplane. We obtain a quadratic programming problem where the number of variables is equal to the number of observations. The objective here is to use support vector machines for classification applied to capture correct and incorrect trials in a stimulus based task.

  In the standard two-class classification problems, we are given a set of training data $(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)$ where the input $x_i \in R^n$. We wish to find a classification rule from the training data, so that when given a new input $x \in R^n$, we can assign a class y from $\{-1, 1\}$ to it. The standard SVM solves the following problem:

$$min \ \ \frac{1}{2} \mid\mid w \mid\mid^2 + C \sum_{i=1}^{l} \xi_i$$

$$\begin{aligned} Subject\ to \qquad & y_i \left(w^T \phi \left(x_i \right) + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \ \ C{>}0, \ \ i = 1, 2, \ldots, l \end{aligned}$$

where C determines the trade-off between the flatness of the $f(x)$ and the amount up to which deviations larger than ϵ are tolerated. This is called ϵ -insensitive loss function $\mid \xi \mid_\epsilon$ and is described by:

$$\mid \xi \mid_\epsilon = \begin{cases} 0 & \text{if } \mid \xi \mid \leq \epsilon \\ \mid \xi \mid - \epsilon & \text{if } \mid \xi \mid > \epsilon \end{cases}$$

CHAPTER 4

Results

This chapter presents the results of the experiments conducted on the data obtained by the procedure listed in Chapter 3 to test the performance of various classification algorithms. We begin by investigating the effects of smoothing on the sampled data. We apply the above mentioned classification techniques to the datasets and report the results and discuss their significance. This chapter also describes the workflow utilized in this study to design the experiments and provides a comparative analysis for all the methods used. Finally we offer some insight on the effects of using a lower dimensional data set by means of performing ranking based feature selection.

4.1. Experimental design

For the purpose of this thesis, data recorded from three subjects on the same day was utilized. For each subject, the recording experiment was split into two series of trials separated by a rest period of fifteen minutes thus providing us with two sets of data: A and B. The various factors influencing the outcome of the experiments were the number of subjects, the number of samples used for training and testing purposes, effect of smoothing the data, the types of classification algorithms used, the combination of best values of the tuning parameters within these algorithms and the effect of feature selection on the classification accuracy. The analysis was divided into three major steps:

(1) For initial analysis, the data belonging to the first series of trials (data set A) was used. This would help us understand whether a data set with a relatively low number of samples could be used to successfully identify incorrect trials. This data was randomly partitioned into training (80%) and testing (20%) sets using a stratified

approach, i.e., maintaining the proportion of samples belonging to incorrect and correct trials in each set. This data was then passed through a regularization filter to reduce the effect of noisy components. The parameter for the smoothing function λ was tuned to obtain a compromise between optimal smoothing and preservation of signal. The best value of λ was chosen by visual inspection of the smoothed curve [32]. LDA and QDA were applied on the original and the transformed data and the accuracy in the form of balanced success rate (BSR) was reported [37].

(2) The next phase of analysis involved dividing the data series into two parts: the one obtained before the rest period (data set A) as training set and the one recorded after the rest period as testing set (data set B). For some algorithms, the training data was further split in a 80:20 ratio using the stratified approach into training and validation sets ,respectively, for the purpose of obtaining optimally tuned parameters. This approach gave more samples for the classifier to train on a certain number of samples to validate the model against and provided sufficient number of testing samples as well. The evaluation process on the validation set was conducted to choose the best values of tuning parameters for the algorithms listed in Chapter 3 which were in turn used to perform the task of identifying the incorrect trials from the correct ones. LDA, QDA, Neural networks and SVMs were applied to this data and the classification accuracy was reported.

(3) As a way of trying to estimate the impact of specific time samples in the detection of ERN, Recursive Feature Elimination (RFE) [38] was used to select the top 10, 100 and 500 features. RFE is an embedded feature selection technique since it includes a

classification process to eliminate features. Linear SVM is used to produce a decision boundary which separates the 2 classes. During this learning phase, each feature is assigned a weight which are in turn used for feature ranking. Features with values of weights as zero or close to zero are discarded [39]. The process is repeated on the reduced set until the desired number of features is eventually obtained. The data was modified to incorporate these top ranked features and all the above mentioned classification algorithms were applied on this reduced data set. The BSR is reported and the contribution of high scoring features is discussed.

In the first two analyses, the segment size used was the entire time window of 1000 ms. This would result in a dimensionality problem since the number of samples is less than the number of features. Hence, we used the third approach to down size the data based on selecting a subset of features using RFE [38]. The results and their related discussion are presented in the next few sections.

4.2. LDA and QDA

The first set of experiments involved applying LDA and QDA on data set A for the three subjects. The purpose of using data set A for initial analysis was two fold: to test the performance of these algorithms against an unbalanced training data with relatively few samples to train the classifier and to verify the effect of smoothing the data on the classification accuracy. Since the sample size of the data is less than the number of features for data set A, the covariance matrices turn out to be singular during the formulation of QDA. Hence, the success rates for QDA could not be evaluated for the initial analysis. In LDA, averaging of the sample covariance matrices of the two classes resulted in the average covariance matrix being non-singular. Thus, the issue of singularity of the covariance matrix

does not arise when we apply LDA. We investigated the effects of selecting a smaller feature set and provided a comprehensive comparison of the performance of all the algorithms later in this chapter. All the values of BSR were averaged over 10 runs and then reported. The balanced success rates (BSR) on the test data are reported in Table 4.1.

TABLE 4.1. Balanced success rates for LDA using data set A

Subject	Method	Electrode	LDA
Subject 1	Before smoothing	Cz	0.432
		Fz	**0.507**
	After smoothing	Cz	0.573
		Fz	**0.575**
Subject 2	Before smoothing	Cz	0.419
		Fz	**0.584**
	After smoothing	Cz	0.537
		Fz	**0.598**
Subject 3	Before smoothing	Cz	0.471
		Fz	**0.536**
	After smoothing	Cz	0.546
		Fz	**0.547**

From Table 4.1, we obtain two pieces of information: 1.) smoothing the data improves the accuracy irrespective of the electrode considered for analysis and 2.) data set obtained using the Fz electrode resulted in better success rates compared to the data set constructed using Cz electrode. We performed one tailed t-test to determine any statistical significance between the BSRs for each subject before and after smoothing. For Subject 1, using Cz ($t=2.59$, $p<0.05$) suggested the result is statistically significant compared to using Fz ($t=0.61$, $p>0.05$). For Subject 2, using Cz gave us $t=3.11$, $p<0.05$ while for Fz, we noted $t=1.25$, $p>0.05$. Comparing Cz for Subject 3 gave $t=1.42$, $p>0.05$ while that for Fz gave $t=0.19$, $p>0.05$. For subject 2 using Cz, the difference in means before and after smoothing was significant but we did not notice any significant statistical difference for Subject 3 using either Fz or Cz. Using a random classifier would yield an accuracy of 0.5. The best accuracy we obtained was 0.63 for both Subjects 1 and 2. Since smoothing the data gave us improved

BSR, with the difference in the means being statistically significant for Subjects 2 and 3, further analysis was performed on the smoothed data. The optimal value of the smoothing parameter obtained was $\lambda = 10^{-5}$ which was used throughout the experiments.

The next task was to investigate the effect of increasing the size of the training data on the classification accuracy. The data set A was entirely used as the training set and the samples from data set B were used for testing purpose. Since fixed partitions of the training and testing data sets were used for this analysis, we could not perform any tests of statistical significance in this case. QDA did not work on this data due to the issue of singularity of covariance matrices which we address later in this chapter by reducing the number of features using feature elimination technique. LDA was applied to this smoothed data and the observed values of BSR were reported in Table 4.2. The highest accuracy we obtained was 0.67 for Subject 2.

TABLE 4.2. Balanced success rates for LDA using data sets A and B

Subject	Electrode	LDA
Subject 1	Cz	0.590
	Fz	**0.631**
Subject 2	Cz	0.575
	Fz	**0.666**
Subject 3	Cz	0.572
	Fz	**0.613**

The results obtained in Table 4.2 reiterate the fact that using the data recorded from the Fz site gave better accuracies compared to using data from the Cz site. In addition, compared to results acquired from the previous task listed in Table 4.1, using a larger data set for training improves the classification accuracy using either of the electrodes. The general consensus from this initial analysis was that smoothing along with a larger training set helps improve the classification accuracy. Also, an interesting observation was the fact that LDA worked well with data obtained from the Fz site compared to the Cz site. The algorithms

discussed in the next couple of sections involve extracting a validation set out of the training data which is used to obtain the best values of certain algorithm specific tuning parameters which were used to classify the testing samples as correct or incorrect.

4.3. Neural networks

This section presents the results of applying neural networks to our ERN classification problem. We utilized the most commonly used neural network classifier architecture of nonlinear logistic regression network with input, hidden and output units connected by a feed-forward structure followed by a conversion of the outputs to form a multi-normal distribution. Data set A was divided into training and validation partitions and the model was tested against data set B. The back propagation technique was used and the neural network was trained with the scaled conjugate gradient algorithm [35] to maximize the training data likelihood and the training was stopped when the error gradient decreased under 10^{-12}. The validation set was used to determine the optimum number of hidden units. The number of hidden units were varied from 1 through 20 and the best value was determined by validating the network on the validation set and plotting the log of the data likelihood for validation and test sets. Figure 4.1 shows the log likelihood plots for selecting the best number of hidden units for each of the three subjects using Cz electrode.

The optimum number of hidden units is determined by the trial which gives highest value of likelihood on the training set. Table 4.3 gives the number of hidden units which were chosen for each subject which were obtained from the validation process.

Table 4.3. Number of hidden units obtained from validation process

	Subject 1	Subject 2	Subject 3
Cz	5	6	5
Fz	12	6	3

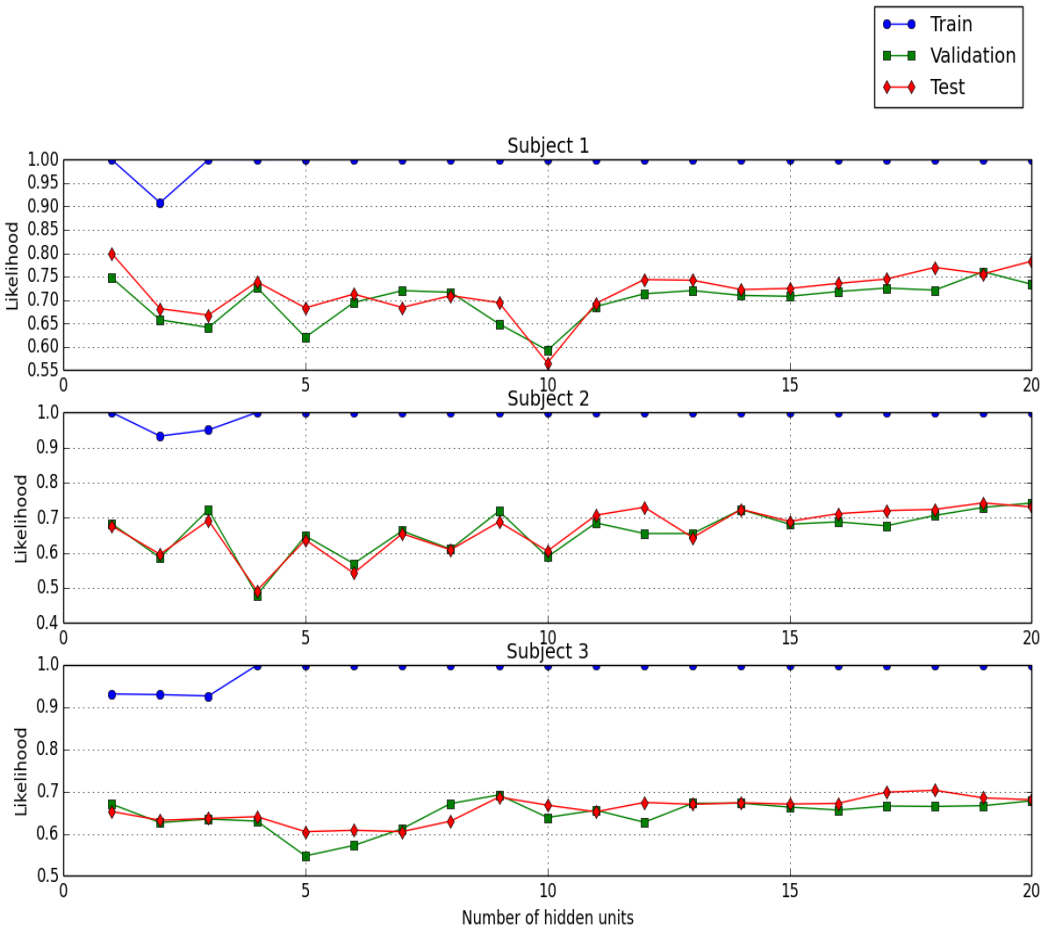

FIGURE 4.1. Likelihood values of train, validation and test sets for three subjects using Cz electrode for varying number of hidden units.

The network was fed with the training data and the best number of hidden units obtained from the validation task were used to train the network again. The neural net was then tested on data set B and the likelihood values, computed from the log likelihood results, for the train, validation and testing data set along with the BSR values obtained are reported in Table 4.4.

TABLE 4.4. Likelihood and BSR for neural network trained using tuned parameters from validation set

Electrodes	Data sets	Subject 1		Subject 2		Subject 3	
		Likelihood	BSR	Likelihood	BSR	Likelihood	BSR
Cz	Train	1.0	**0.72**	1.0	**0.67**	1.0	**0.67**
	Valid	0.53		0.64		0.62	
	Test	0.57		0.64		0.66	
Fz	Train	1.0	0.61	1.0	0.66	1.0	0.56
	Valid	0.73		0.70		0.59	
	Test	0.68		0.70		0.55	

In contrast to the results obtained using LDA from Table 4.2, we noted better success rates using Cz electrode as the data source compared to Fz. We also achieved equal or better values of BSR using the Fz electrode indicating the effectiveness of NN over LDA. Another observation is that the neural net implementation drastically improved the accuracies over LDA irrespective of the choice of electrodes or subjects. The maximum success rate obtained was 72% for Subject 1 using the Cz electrode compared to 66.6% for Subject 2 using the Fz electrode for LDA. The only 2 cases where we observe a lower BSR for NN compared to LDA are for Subject 1 using the Fz electrode (61% compared to 63.1%) and for Subject 3 using the Fz electrode (59% compared to 61.3%). So, we can conclude from the experiment that the NN performs better than LDA for all three subjects and that the classification accuracy is better when the signals from the Cz electrode are used instead to Fz. The final algorithm we tested was support vector machines where the same data was used and the results are compared to LDA and NN.

4.4. Support vector machines

The final algorithm applied to this ERN data was Support Vector Machines (SVM). The *scikit-learn* package was used to train the data, estimate the best tuning parameters and then use support vector classifier to predict the output on test data [40]. The same

data sets employed for evaluating the performance of neural networks were used for SVM analysis as well. The validation set was held out from the training data which was used to evaluate different settings for the estimators such as C, gamma and the degree of the polynomial. Since the number of samples in the training set which are labeled incorrect were relatively few in number, a stratified K fold mechanism combined with an exhaustive grid search approach was used along with 5-fold cross validation to determine the best values of the hyperparameters which were then provided as inputs for final evaluation on the test set. We used coarse grid values to avoid overfitting. The kernels used for testing were *linear, polynomial(degree 3) and Radial basis function (RBF)*. The hyperparameters varied according to the kernel type were: C varied from 10^{-2} to 10^{9}, *degree* varied as 2 or 3 and *gamma* bearing values from 10^{-5} to 10^{4}. C and *gamma* were incremented in multiples of 10. Table 4.5 shows the best values obtained by the cross validation process.

TABLE 4.5. The best values of hyperparameters selected after 5-fold cross validation

Subject	SVM kernel → Electrode ↓	Linear	Polynomial		RBF	
		C	C	Degree	C	Gamma
Subject 1	Cz	10	10	2	10	0.1
	Fz	0.01	0.1	2	0.01	0.01
Subject 2	Cz	1	10	2	1	0.01
	Fz	10	1	2	10	0.01
Subject 3	Cz	100	100	3	1000	1e-5
	Fz	1000	1000	3	100	1e-5

The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly while the *gamma* parameter is the inverse of the width of the RBF kernel which roughly defines the area of influence of a support vector on the classification accuracy.

From Table 4.5, it can be noted that we obtained higher C values for Subject 3 compared to the other 2 subjects from the validation process. This indicates the SVM was looking to classify the training examples correctly for Subject 3 while it tried to make the classifier flat for Subjects 1 and 2. Also, the gamma values are relatively smaller for Subject 3 compared to other subjects indicating the individual trials might not be close to each other for the accuracy to get affected by a single trial. The significance of these parameters will be discussed after we obtain the accuracies on the test data. These optimized hyperparameters were then used to predict the test labels. The BSR obtained were reported in Table 4.6.

TABLE 4.6. Balanced success rates of SVMs using linear, polynomial and RBF kernel

Subject	SVM kernel→ Electrode ↓	Linear	Polynomial	RBF
Subject 1	Cz	0.62	**0.65**	0.50
	Fz	0.66	0.50	0.50
Subject 2	Cz	**0.59**	0.54	0.58
	Fz	0.52	0.53	0.54
Subject 3	Cz	**0.67**	0.65	0.63
	Fz	0.66	0.60	0.57

Initial observation of the results obtained by SVM gave mixed results. Considering the results for each subject, we obtained better success rates for data from Cz electrode compared to that from Fz electrode which was consistent with the findings from NN study, but we could not achieve the highest BSR reached by NN in any of the experiments with SVM. Subject 3 using the Cz electrode gave us the highest classification accuracy of 67% for SVM. The results were improved compared to LDA but were lower than those obtained using NN. Some of the kernels were not able to predict a single incorrect trial but could predict all the trials which were correct. These were indicated by the entries in Table 4.6 with BSR of 0.5. Since the number of features are greater than the number of samples, the grid search approach does not necessarily give the optimum result. Hence it was important to choose a

subset of features before giving the data to SVM [41]. The next approach solves this problem by performing feature selection and evaluating the results.

4.5. Feature selection

The last set of experiments conducted on ERN data was feature ranking and selection of a subset of features. We used the popular feature selection algorithm Recursive Feature Elimination (RFE) to give us the top 10, 100 and 500 ranked features. We used the RFE package of the $scikit - learn$ module with a step size of 1 and varied the desired number of features as 10, 100 and 500 which returned a ranked list of indices of the top features. Data set A was used as training set while data set B was used for testing purposes. The data was then altered to consider the selected top ranked features only, thus eliminating the problem where the number of dimensions were exceedingly larger than the number of samples. This allowed us to apply all the previous techniques on this reduced feature set based data, including QDA which resulted in singular covariance matrices during initial analysis. The optimized values of the tuning parameters obtained in Table 4.5 are not the same here since the dimensions of the data sets changed. So, we used the same approach as in the previous tasks to evaluate the best values of the tuning parameters. For SVMs, we tested the performance using a linear kernel on the reduced data. The values of the tuning parameters used for this experiment were derived and listed in Table 4.7.

The results for all the algorithms using these tuned parameters incorporating the top 10 ranked features are consolidated in Table 4.8. The highlighted value indicates the best accuracy obtained for the algorithm.

We observed improvements on a lot of results after applying feature selection. Firstly, as QDA could not be applied during the previous analysis, these are the only set of results

TABLE 4.7. Best values of parameters for NN and SVMs using linear kernel using feature selection for top 10 features

Subject	Electrode	NN	Linear SVM
		Hidden units	C
Subject 1	Cz	5	1.0
	Fz	5	0.1
Subject 2	Cz	6	0.01
	Fz	5	1.0
Subject 3	Cz	3	1.0
	Fz	3	10.0

TABLE 4.8. BSR for LDA, QDA, NN and SVM using top 10 ranked features by RFE. The figures in parentheses indicate the number of incorrect trials correctly predicted

Dataset		LDA	QDA	NN	SVM
Subject 1	Cz	**0.64 (1/3)**	0.50 (0/3)	0.62 (1/3)	0.63 (1/3)
	Fz	0.55 (1/3)	0.50 (0/3)	0.60 (1/3)	0.58 (1/3)
Subject 2	Cz	0.62 (7/11)	0.55 (4/11)	0.62 (4/11)	**0.67 (4/11)**
	Fz	0.57 (6/11)	**0.59 (6/11)**	0.57 (4/11)	0.58 (4/11)
Subject 3	Cz	0.61 (16/25)	0.49 (6/25)	**0.66 (17/25)**	0.64 (15/25)
	Fz	0.56 (14/25)	0.53 (14/25)	0.62 (10/25)	0.61 (11/25)

for this method that we could account for. Secondly since our initial problem statement stressed the importance of capturing the presence of ERN in trials, we also noted the number of incorrect trials successfully predicted by each of the algorithms along with the overall balanced success rate. This metric helped us visualize how accurate the classifier was towards predicting the number of samples which elicited the presence of ERN. For LDA, we achieved highest BSR of 64% for Subject 1 using Cz electrode which was contrasting to the findings in Table 4.2 where Fz gave better result. But in this case, only 1 out of the 3 incorrect samples were predicted successfully (33%). For Subjects 2 and 3, success rates using Cz electrode were higher with better prediction rate of the incorrect labeled class (63.6 % and 64% respectively). For QDA, using Fz gave higher accuracies than using Cz for Subjects 2 and 3. We obtained the overall best success rate of 59% using QDA for Subject 2 using Cz

along with a 55% accuracy of successfully predicting incorrect labels. Results for any given subjects were not encouraging using QDA.

For NN, we trained the network with data containing the top 10 features and used the best values of hidden units obtained by the validation process. The BSR followed a similar trend with Cz data giving better results than Fz for all the subjects. For Subject 3, we achieved the highest BSR of 66% with 17/25 incorrect trials identified successfully and it required the least number of hidden units as well. We inspected the feature set selected by RFE and obtained the following indices: [324, 224, 332, 350, 341, 231, 208, 313, 124, 202] indicating the features related to the presence of ERN were indeed considered for Subject 3. BSR for Subjects 1 and 3 did not improve compared to the previous analysis.

A similar approach was used for SVMs as in the previous analysis in which we performed stratified 5 fold cross validation with grid search mechanism to get the best values of the hyperparameter C and then used these values to predict the test labels. Again, using the Cz electrode gave us better prediction accuracies compared to Fz. Using SVMs for Subject 1 gave us a success rate of 63%. For Subject 2, the BSR improved to 67% which is the maximum we achieved using 10 features for any classification method. For the third subject, SVMs performed nearly as good as NN with an accuracy of 64% compared to 66% for NN. Taking into account the results of Table 4.8, it could be conjectured that using Cz gives better classification results.

We also analyzed the features selected by RFE to study the effect of data constructed by features closer to the occurrence of ERN on the classification accuracy. We constructed plots of the average incorrect and correct responses with the top 10 selected features marked on it for each of the 3 subjects as shown in Figures 4.2, 4.3 and 4.4.

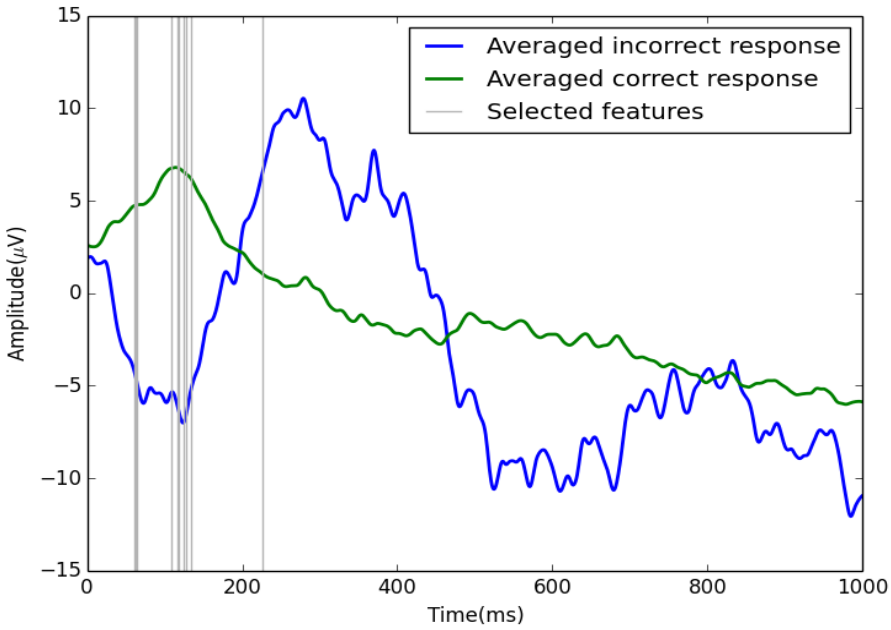

FIGURE 4.2. Top 10 selected features indicated on an averaged response for Subject 1 using Cz electrode.

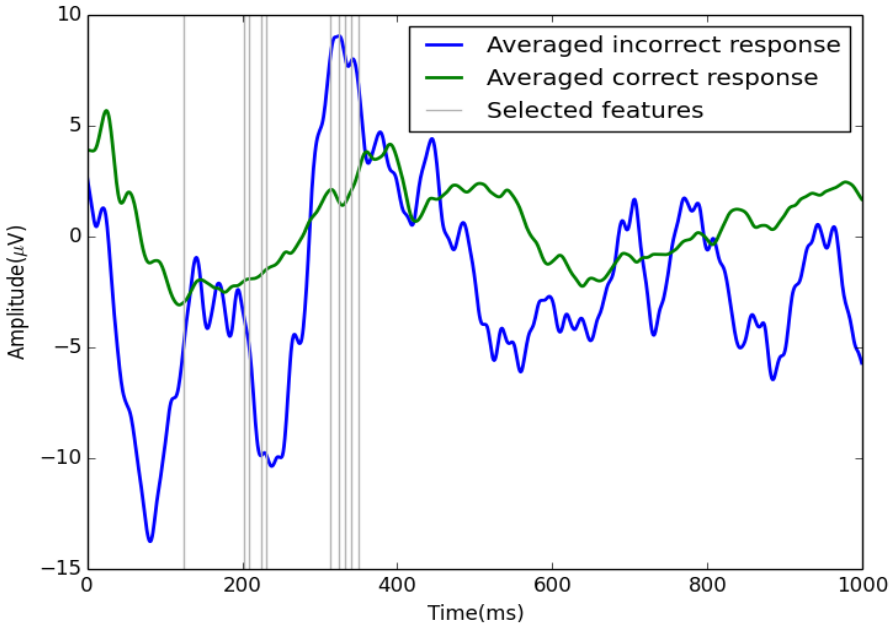

FIGURE 4.3. Top 10 selected features indicated on an averaged response for Subject 2 using Cz electrode.

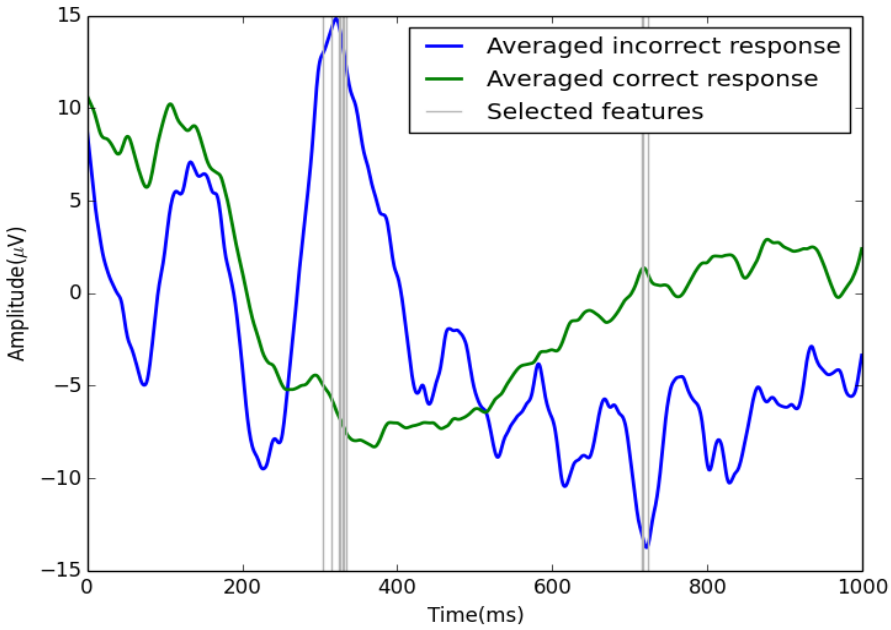

FIGURE 4.4. Top 10 selected features indicated on an averaged response for Subject 3 using Cz electrode.

Figure 4.2 indicates that RFE picked up features closer to the occurrence of ERN, i.e., in the range of 50-150 ms for Subject 1. For Subject 2, 1 of the top 10 features was selected by RFE which was close to the ERN occurrence. Four of them were placed in the range of 200-250 ms while the remaining features were picked from the location related to the positivity following the ERN. For Subject 3, none of the features were selected from the time range of the occurrence of ERN. 7 of the 10 features were picked from 250-350 ms which was the positivity following the ERN while 3 features were picked from a later time range as shown in Figure 4.4.

The same process was repeated using 100 and 500 features. Feature selection was performed using RFE and the classifiers were applied on the ERN data. We could not obtain results for QDA due to a runtime error in computation of the determinant of the covariance matrix for 100 and 500 features. The results obtained using the top 100 and top 500 features were reported in Tables 4.9 and 4.10, respectively.

TABLE 4.9. BSR for LDA, NN and SVM using top 100 ranked features by RFE

Dataset		LDA	NN	SVM
Subject 1	Cz	0.45 (1/3)	0.64 (1/3)	0.65 (1/3)
	Fz	0.60 (2/3)	0.62 (1/3)	0.61 (1/3)
Subject 2	Cz	0.60 (8/11)	0.60 (3/11)	0.52 (2/11)
	Fz	0.58 (7/11)	0.49 (2/11)	0.50 (0/11)
Subject 3	Cz	**0.62 (17/25)**	**0.66 (10/25)**	**0.72 (17/25)**
	Fz	0.55 (14/25)	0.60 (8/25)	0.53 (7/25)

TABLE 4.10. BSR for LDA, NN and SVM using top 500 ranked features by RFE

Dataset		LDA	NN	SVM
Subject 1	Cz	0.44 (1/3)	**0.61 (1/3)**	0.64 (1/3)
	Fz	0.41 (1/3)	0.53 (1/3)	0.60 (1/3)
Subject 2	Cz	**0.61 (7/11)**	0.54 (3/11)	0.59 (3/11)
	Fz	0.56 (6/11)	0.49 (1/11)	0.48 (0/11)
Subject 3	Cz	0.57 (14/25)	0.59 (7/25)	**0.64 (11/25)**
	Fz	0.57 (14/25)	0.52 (4/25)	0.60 (10/25)

As we increased the number of selected features, we did not observe a drastic improvement in performance. For Subject 1, LDA degraded in performance while NN and SVM gave a slight improvement in success rates indicating a compact feature set gives comparable results than using more number of features. Also, increasing the size of the feature set did not improve the accuracy of identifying the incorrect samples. A similar trend was observed for Subject 2 where using 100 features resulted in BSR less than or equal to those obtained while using top 10 features for all the algorithms. On using 500 features, accuracy improved to 61% which was the highest we achieved using LDA. For the third subject, we obtained slightly improved values for LDA, NN and SVM using 100 features and Cz electrode. For this feature set, SVM gave the highest accuracy of 72% using Cz with a 68% classification accuracy for successfully predicting samples with incorrect labels. The success rate did not improve on increasing the selected features to 500 but Cz gave higher BSR compared to Fz for all the algorithms using 500 features.

To analyze the features selected and their relevance in detecting ERN, we plotted the top 100 features on an averaged data response for Subject 3 using Fz electrode as shown in Figure 4.5. On manual inspection of the indices of the selected features, it was observed that almost 60% of the features were placed in the time range of 200 - 400 ms. The trend of feature selection is similar to Figure 4.4 where RFE picked up the some features near the region of the positivity after the ERN and the remaining ones from the time range of 650-750 ms.

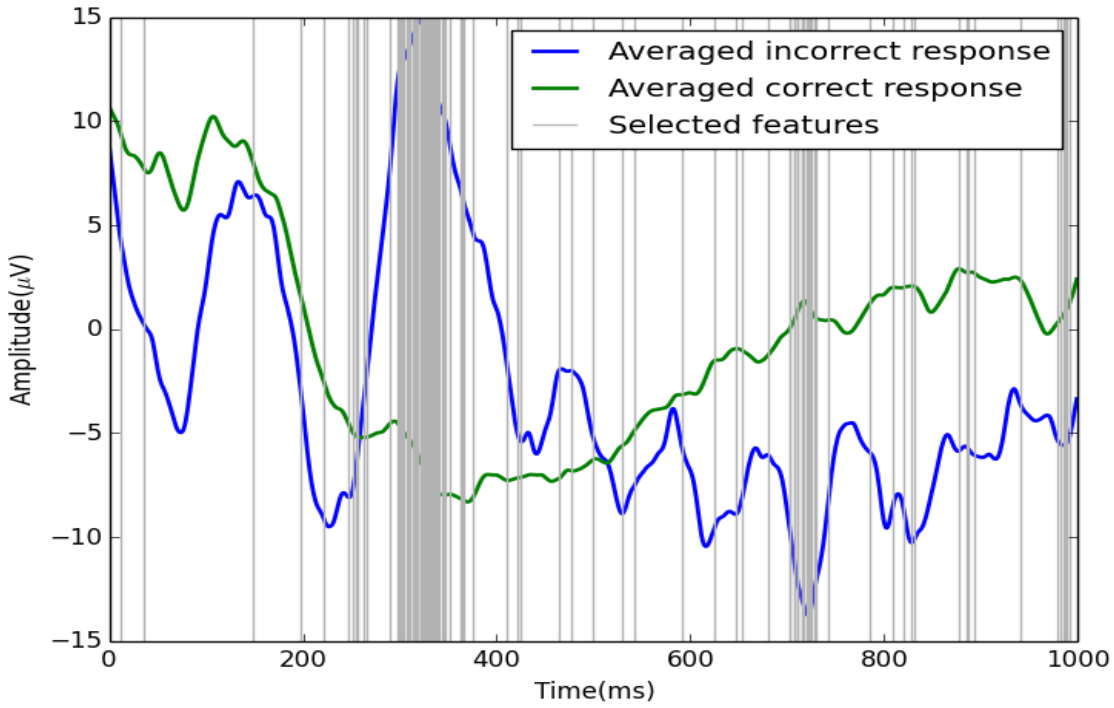

FIGURE 4.5. Top 100 selected features indicated on an averaged response for Subject 3 using Cz electrode.

CHAPTER 5

Conclusions

This chapter compiles the results obtained throughout this experiment and discusses the best possible combination of data collection electrode and classification algorithm suitable for the subjects considered. Importance of feature selection and the need to provide statistical evidence is highlighted. We also discuss the possibilities of future application of this work and mention some more approaches to explore which were beyond the scope of this work.

5.1. Summary

The goal of this thesis was to identify different classification algorithms to be applied on ERN data, to explore the effectiveness of feature selection and to use the optimal features to further improve the classification accuracy. We managed to achieve success rates greater than 60% for the three subjects by varying a combination of factors such as choice of participants, selection of electrode, type of classification algorithm used and evaluation of hyperparameters. Another notable outcome of the experiments was that feature selection not only helped to reduced the dimension of the data but also gave us good success rates.

Our initial analysis from Table 4.1 supported by the evidence provided by the t-test suggested the need to smooth the data. After performing the smoothing operation, we obtained improved results as mentioned in Table 4.2 indicating that smoothing helps alleviate noise and improve the accuracy. Without using any feature selection, we achieved the highest success rates of 72% for Subject 1 using NN, 67% for Subject 2 using NN and 67% for Subject 3 using NN as well as SVMs. Thus, we can conclude that neural nets tuned to work with an optimized number of hidden units gave us the best results. Also, we can conclude that using the data from Cz gave better results than using Fz without any kind of feature selection.

When we extracted the top 10 ranked features by RFE and used the reduced dimension data, an increase in classification accuracy was observed. For Subject 1, it was LDA which gave us a success rate of 64%. For Subject 2, SVM was the best candidate with 67% BSR but it could not detect more than half of the samples which were labeled incorrect. LDA and QDA could identify 6 out of 11 incorrect samples but their overall success rate was low. Subject 3 gave the best success rate of 66% with a NN implementation having 3 hidden units. SVM came close to matching the accuracy of neural net at 64%. A simple inspection of the feature set given by RFE revealed a few features which were very close to the point of occurrence of ERN. But in some cases, RFE returned a feature set which when used, resulted in very poor success rates too. Thus, the choice of feature selection algorithm is important as well. In all of the above experiments, we again observed superior accuracies using the Cz electrode. Thus, we can successfully conclude from the experiments that the use of the Cz electrode gives better success rates for classification of ERN data. Also, to conduct tests of statistical significance, data sets A and B should be combined and multiple repetitions using different combinations of the training and testing set should be considered.

Finally, we selected more features and sought to test these algorithms for their performance. In almost all the cases, except for Subject 1 for LDA, use of Cz gave us superior BSR compared to using Fz, implying that using Cz electrode gives better results when the task is to detect the presence of ERN. Also, the performance degraded severely in some cases (e.g., LDA for Subject 1 for 500 features) while it improved, too (e.g., SVM for Subject 3 using Cz for 100 features). This can be attributed to the fact that features closer to the occurrence of ERN were picked up by the feature selection algorithm.

5.2. Future work

This work is essentially an attempt to identify the occurrence of ERN in stimulus based tasks and has a lot of scope for improvement in the future. Firstly, more subjects can be used to train the algorithms and a generalized trend among the subjects can be studied. This work used data recorded from able bodied subjects ranging from 20 to 28 year old. This range can be expanded to consider a higher age group or people with disabilities.

Different smoothing techniques or filters could be explored and their impact on ERN detection could be studied. Also, down sampling the data or selecting a time window around 200 ms which is likely to contain ERN information could be used instead of the entire time window.

We could not apply tests of statistical significance since the same partition of data sets A and B were used for training and testing purposes respectively. These data sets can be combined and then random partitioning can be performed to generate new training and testing sets which can be tested for statistical significance.

The two electrodes, Cz and Fz, were considered individually for all the analyses. The effects of considering both these electrodes together can be explored. We used four state of the art classification techniques in this study which can be expanded to include other algorithms such as K-nearest neighbors classifier, decision trees and random forests. Also the scope of feature selection in this work was limited to RFE which can be expanded to include others such as L1-based feature selection and forward feature selection.

Finally, these experiments were conducted in an offline setting and it is difficult to predict the performance in a real time setting. One of the potential applications is to detect errors caused by a BCI interface. This classification work can serve as an initial stage to identify ERN and this information can be fed back to the system to improve its performance.

Bibliography

[1] K. L. Kilgore, P. H. Peckham, M. W. Keith, G. B. Thorpe, K. S. Wuolle, A. M. Bryden, and R. L. Hart, "An implanted upper-extremity neuroprosthesis: Follow-up of five patients," *The Journal of Bone & Joint Surgery*, vol. 79, no. 4, pp. 533–41, 1997.

[2] J. B. Ochoa, *EEG Signal Classification for Brain Computer Interface Applications*. PhD thesis, Ecole Polytechnique Federale De Lausanne, 2002.

[3] M. Teplan, "Fundamentals of eeg measurement," *Measurement Science Review*, vol. 2, no. 2, pp. 1–11, 2002.

[4] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767 – 791, 2002.

[5] R. L. Wilson, "Ethical issues of brain computer interfaces," *International Association for Computing and Philosophy*, 2013.

[6] B. Blankertz, *Wadsworth BCI Dataset*, 2004. `https://www.bbci.de/competition/iii/desc_II.pdf`.

[7] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive berlin braincomputer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539 – 550, 2007.

[8] B. Graimann, B. Allison, and G. Pfurtscheller, "Brain computer interfaces: A gentle introduction," *The Frontiers Collection*, pp. 1–27, 2010.

[9] S. R. Benbadis, *Introduction to sleep Electroencephalography*. John Wiley and Sons, Inc., 2006.

[10] R. Roche and P. Dockree, "Introduction to eeg methods and concepts: What is it? why use it? how to do it. advantages? limitations?," in *Proceedings of the Sixth European Science Foundation ERNI-HSF meeting on 'Combining Brain Imaging Techniques'*, pp. 18–25, 2011.

[11] P. Ferrez and J. del R.Millan, "Error-related eeg potentials generated during simulated brain computer interaction," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 3, pp. 923–929, 2008.

[12] F. P.W. and del R.Millan J., "The error-related negativity as a state and trait measure: Motivation, personality, and erps in response to errors," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 3, pp. 923–929, 2008.

[13] M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke, "Effects of crossmodal divided attention on late {ERP} components. ii. error processing in choice reaction tasks," *Electroencephalography and Clinical Neurophysiology*, vol. 78, no. 6, pp. 447 – 455, 1991.

[14] P. L. Davies, S. J. Segalowitz, and W. J. Gavin, "Development of response-monitoring erps in 7- to 25-year-olds," *Developmental Neuropsychology*, vol. 25, no. 3, pp. 355 – 376, 2004.

[15] E. S. Kappenman and S. J. Luck, *Oxford Handbook of Event-Related Potential Components*. New York: Oxford University Press, 2012.

[16] W. J. Gehring, B. Goss, M. G. H. Coles, D. E. Meyer, and E. Donchin, "A neural system for error detection and compensation," *Psychological Science*, vol. 4, no. 6, pp. 385–390, 1993.

[17] N. Sander, K. R. Ridderinkhof, J. Blom, G. P. Band, and A. Kok, "Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task," *Psychophysiology*, vol. 38, no. 5, pp. 752–760, 2001.

[18] M. Ullsperger and M. Falkenstein, "Erp correlates of erroneous performance," in *Errors, Conflicts, and the Brain: Current Opinions on Performance Monitoring*, MPI special issue in human cognitive and brain sciences, pp. 5–14, Max Planck Institute for Human Cognitive and Brain Sciences, 2004.

[19] B. A. Eriksen and C. W. Eriksen, "Effects of noise letters upon the identification of a target letter in a nonsearch task," *Perception & Psychophysics*, vol. 16, no. 1, pp. 143–149, 1974.

[20] M. Liotti, M. G. Woldorff, R. P. III, and H. S. Mayberg, "An {ERP} study of the temporal course of the stroop color-word interference effect," *Neuropsychologia*, vol. 38, no. 5, pp. 701 – 711, 2000.

[21] F. Vidal, T. Hasbroucq, J. Grapperon, and M. Bonnet, "Is the 'error negativity' specific to errors?," *Biological Psychology*, vol. 51, no. 2-3, pp. 109–128, 2000.

[22] P. Jaeyoung, K. Kee-Eung, and J. Sungho, "A pomdp approach to p300-based brain-computer interfaces," in *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, (New York, NY, USA), pp. 1–10, ACM, 2010.

[23] E. Donchin, K. M. Spencer, and R. Wijesinghe, "The mental prosthesis: Assessing the speed of a p300-based brain-computer interface," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, pp. 174–179, 2000.

[24] S. Samuel and R. Daniel, "The late positive complex," *Annals of the New York Academy of Sciences*, vol. 425, no. 1, pp. 1–23, 1984.

[25] M. G. Coles, M. K. Scheffers, and C. B. Holroyd, "Why is there an ern/ne on correct trials? response representations, stimulus-related components, and the theory of error-processing," *Biological Psychology*, vol. 56, no. 3, pp. 173 – 189, 2001.

[26] M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen, "Conflict monitoring and cognitive control," *Psychological Review*, vol. 108, pp. 624–652, 2001.

[27] C. B. Holroyd and M. G. H. Coles, "The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity," *Psychological Review*, vol. 109, pp. 679–709, 2002.

[28] C. B. Holroyd, N. Yeung, J. D. Cohen, M. G. H. Coles, C. B. Holroyd, N. Yeung, and D. O. Psychology, "A mechanism for error detection in speeded response time tasks," *Journal of Experimental Psychology: General*, vol. 134, pp. 163–191, 2005.

[29] P. Luu, D. M. Tucker, D. Derryberry, M. Reed, and C. Poulsen, "Electrophysiological responses to errors and feedback in the process of action regulation," *Psychological Science*, vol. 14, pp. 47–53, 2003.

[30] W. J. Gehring, J. Himle, and L. G. Nisenson, "Action-monitoring dysfunction in obsessive-compulsive disorder," *Psychological Science*, vol. 11, no. 1, pp. 1–6, 2000.

[31] A. Reddy, J. Shobha, M. Naidu, R. Pilli, and U. Pingali, "A computerized stroop test for the evaluation of psychotropic drugs in healthy participants," *Indian Journal of Psychological Medicine*, vol. 35, no. 2, pp. 180–189, 2013.

[32] J. J. Stickel, "Data smoothing and numerical differentiation by a regularization method," *Computers and Chemical Engineering*, vol. 34, no. 4, pp. 467 – 475, 2010.

[33] A. Coolen, "A beginner's guide to the mathematics of neural networks," in *Concepts for Neural Networks* (L. Landau and J. Taylor, eds.), Perspectives in Neural Computing, pp. 13–70, Springer London, 1998.

[34] D. Reby, S. Lek, I. Dimopoulos, J. Joachim, J. Lauga, and S. Aulagnier, "Artificial neural networks as a classification method in the behavioural sciences," *Behavioural Processes*, vol. 40, no. 1, pp. 35 – 43, 1997.

[35] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[36] V. Vapnik, S. E. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation and signal processing," in *Advances in Neural Information Processing Systems 9* (M. Mozer, M. Jordan, and T. Petsche, eds.), pp. 281–287, MIT Press, 1997.

[37] A. Ben-Hur, "Pyml a python machine learning library focused on kernel methods," 2007. `http://pyml.sourceforge.net/`.

[38] P. Evangelista, M. Embrechts, and B. Szymanski, "Taming the curse of dimensionality in kernels and novelty detection," in *Applied Soft Computing Technologies: The Challenge of Complexity* (A. Abraham, B. de Baets, M. Kppen, and B. Nickolay, eds.), vol. 34 of *Advances in Soft Computing*, pp. 425–438, Springer Berlin Heidelberg, 2006.

[39] I. Rezarta, *Feature generation and analysis applied to sequence classification for splice-site prediction.* PhD thesis, University of Maryland, 2007.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," tech. rep., Department of Computer Science, National Taiwan University, 2003.