Development of Multiple Class Human-Computer Interaction System using Machine Learning Algorithm for Eyeball Movement

by

Mubtasim Rafid Chowdhury

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Biomedical Engineering

Khulna University of Engineering & Technology
Khulna 9203, Bangladesh
August, 2019

Declaration

This is to certify that the thesis work entitled " Development of Multiple Class Human-Computer Interaction System using Machine Learning Algorithm for Eyeball Movement" has been carried out by Mubtasim Rafid Chowdhury in the Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh. The above thesis work or any part of this work has not been submitted anywhere for the award of any degree or diploma.

Signature of Supervisor

Rafid

Signature of Candidate

Approval

This is to certify that the thesis work submitted by Mutasim Rafid Chowdhury entitled "Development of Multiple Class Human-Computer Interaction System using Machine Learning Algorithm for Eyeball Movement" has been approved by the board of examiners for the partial fulfillment of the requirements for the degree of M.Sc. Engineering in the Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh in August 2019.

BOARD OF EXAMINERS

1. Chairman
(Supervisor)

Dr. Md. Nurunnabi Mollah
Professor, Department of Electrical & Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

2. Member

Head
Department of Biomedical Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

3. Member

Dr. A.B. M. Aowlad Hossain
Professor, Department of Electronics & Communication Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

4. Member

Nasrin

Dr. Nasrin Akhter
Assistant Professor, Department of Biomedical Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

5. Member
(External)

Prof. Dr. Mohammad Shorif Uddin
Professor, Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

Acknowledgement

I express gratitude to the almighty ALLAH that the project has been accomplished perfectly and successfully. I am very much delighted to present this research work on developing multiple class Human-computer interaction system in Structural Engineering.

I would like to express profound respect, deepest gratitude and hardest thanks to my thesis supervisor ***Prof. Dr. Md. Nurunnabi Mollah,*** Department of Electrical and Elecronic Engineering, Khulna University of Engineering & Technology. The door to ***Prof. Nurunnabi*** office was always open whenever I ran into a trouble, spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank my committee members, ***Prof. Dr. A.B.M. Aowlad Hossain*** and ***Dr. Nasrin Akhter*** for their crucial suggestion on my research projects, and my external examiner ***Prof. Dr. Mohammad Shorif Uddin*** for his valuable time to read and giving opinion on my thesis.

I would like to thank my respective teachers in the Department of Biomedical Engineering, Khulna University of Engineering & Technology for their valuable supports. I would also like to thank my seniors, friends, roommate and everybody who were involved directly or indirectly to run the thesis work successfully.

Finally, I must express my very profound gratitude to my parents, elder brother and younger sister for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

August, 2019 Mubtasim Rafid Chowdhury

Abstract

Modern technologies in the field of Biomedical Engineering are flourishing astonishingly in recent times. Human-computer interaction (HCI) is one of the newest additions in this field. It is the study of the way people interact with computers and how the computers are or are not developed for interacting with human successfully. Electrooculography (EOG) machine can be used as HCI device. It is a technique for measuring the corneo-retinal standing potential which is present between the front and the back of the human eye. Pairs of electrodes are generally attached either above and below the eye or to the left and right of the eye to detect the eye movement. A potential difference occurs between the electrodes. Considering that the resting potential is constant, the recorded potential is a measure of the eye's position. EOG device can pick up these resting potentials while moving the eyeball in different directions. Data classification is important for HCI systems. It is to identify a new observation which belongs to a set of categories. Classification is done based on a training dataset having observations whose category membership is familiar. Various algorithms can be used to classify bio-signal data like ECG, EEG and EOG. These algorithms are called machine learning algorithm. It is a set of mathematical approaches to teaching computers to train based on large amount of data without step by step human instruction. In this research, EOG data are classified using machine learning algorithm to develop a multiple class HCI system.

EOG data for different directional eyeball movement is acquired with the help of Biopac MP3X Acquisition unit. By placing the disposable surface electrodes on the right position of the skull and connecting all the leads and wires to the proper channel, the setup is ready to pick up the EOG data. Subjects are instructed to follow the LED sequence in the navigational setup. The data of 7 subjects aged between 22 to 48 years are taken for this experiment. The data is then saved using Biopac Student Lab Software and then preprocessed to prepare an EOG dataset for classification. With Weka 3.9.2, the classification procedure is done on the prepared dataset. Six classification algorithms i.e. naïve bayes, support vector machine, logistic regression, k-nearest neighbor, random forest and bagging are applied on the dataset. Comparison is shown among the algorithms based on different parameters. In the EOG dataset, features are also added which can be correlated with the classes. This correlation method is performed in IBM SPSS Statistics 25 software to find the most significant features related to the class.

From the classification result, the accuracy of the different classifiers are obtained. The accuracy of Naïve Bayes is 30.7692%, SVM is 30.7692%, Logistic regression is 53.8462%, KNN is 7.6923%, Random forest is 84.61% and Bagging is 92.31% respectively. From the comparison among the classifiers based on different parameters, bagging has the highest and KNN has the lowest accuracy among them. The proposed method is then compared with other researches where it is seen that other methods applied only two or three algorithms but in the proposed method six machine learning algorithms are used. It is observed that bagging is more suiTable algorithm for EOG data than other algorithms used in the mentioned related works. As for the correlation, only the chi-square test is performed as Fisher's exact test can be performed for 2x2 matrix whereas there are 9 classes in the dataset. From the chi-square test result, it is seen that the mean of channel 1 (horizontal channel) and channel 2 (vertical channel) used to acquire EOG data for eyeball movement are the most significant features. These two featuress are directly related to the classes.

Contents

List of Tables

List of Figures

Figure No **Description** **Page**

List of Abbreviations

HCI	Human-Computer Interaction
EEG	Electroencephalogram
ECG	Electrocardiogram
EMG	Electromyogram
EOG	Electrooculogram
GSR	Galvanic Skin Response
MEG	Magnetoencephalogram
REM	Rapid Eye Movement
ML	Machine Learning
SVM	Support Vector Machine
RF	Random Forest
KNN	K-Nearest Neighbor
LR	Logistic Regression
ANN	Artificial Neural Network
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
RAE	Relative Absolute Error
TPR	True Positive Rate
FPR	False Positive Rate
ROC	Receiver Operating Characteristic
MCC	Mathews Correlation Coefficient
GPS	Global Positioning System
PDA	Personal Digital Assistant
RFID	Radio Frequency Identification
MLPNN	Multilayer Perceptron Neural Network
ANOVA	Analysis of Variation
FFNN	Feed Forward Neural Network
TDNN	Time Delay Neural Network
DT	Decision Tree
DTCWT	Dual Tree Complex Wavelet Transform
ORF	Open Reading Frame
BSL	Biopac Student Lab
WEKA	Waikato Environment for Knowledge Analysis
SPSS	Statistical Package for the Social Sciences
LED	Light Emitting Diode
ASCII	American Standard Code for Information Interchange

List of Symbols

k	Kappa Statistic
E_i	Relative Absolute Error
$\lambda(x)$	Likelihood Ratio
Θ	Parameter Space
Θ	Complement of Θ
χ^2	Chi-square Statistic
σ	Standard Deviation

CHAPTER I

INTRODUCTION

1.1 Introduction to Human-Computer Interaction (HCI) System

Human-computer interaction system is the design and use of computer technology which interfaces between people and computer [1]. It is at the center point of various field of researches such as computer sciences, design, behavioral sciences, media studies etc. This term was first used in 1975 but made popular by Stuart K. Card, Allen Newell, and Thomas P. Moran in their seminal 1983 book, '**The Psychology of Human–Computer Interaction'.** The interface between human and computers is very important to develop the interaction system. The emerging multi-modal and Graphical user interfaces (GUI) create an opportunity for humans to interact with computers in a way which cannot be attained by other interface paradigms [2]. Human-computer interaction is more likely related to the design, evaluation and implementation of interactive computing for human use. User satisfaction is very much needed for this system. If the system is poorly designed, it can cause unexpected problems. In Figure 1.1, a general block diagram of HCI system is shown.

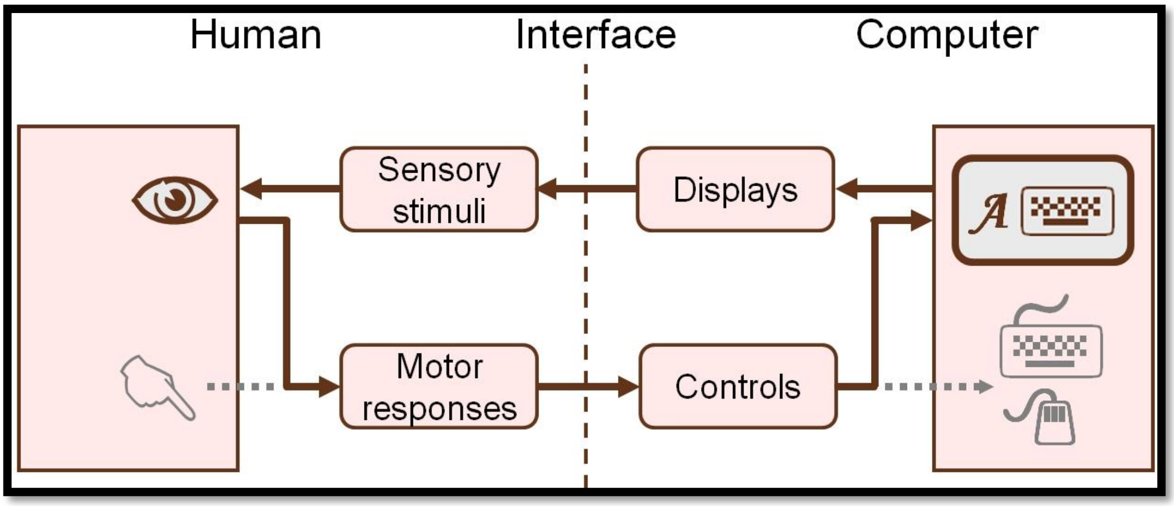


Figure 1.1: General Block Diagram of HCI System

1.1.1 Needs for HCI system:

Human-computer interaction deals with computational artifacts, systems and infrastructures and how humans make use of them. There are various needs of HCI systems in our modern life to make it easier and smoother [3]. These needs are given below:

- To increase productivity
- To decrease user training time and cost
- To decrease user errors
- To increase accuracy of data input and data interpretation
- To decrease need for ongoing technical support

1.1.2 Types of Human-computer interface:

There are various types of human-computer interface as human has many ways to interact with the computer [4]. Types of HCI systems are given below:

- Visual Based Interface
- Audio Based Interface
- Graphical User Interface
- Menu Driven Interface
- Voice User Interface
- Command Line Interface
- Touch Sensitive Interface

1.2 Control Signals for HCI system

Human-computer interface can be controlled using various methods. These methods are depended on different types of signals. Mainly, there are two types of control signals to operate the HCI devices [5]. These signals are called biosignal which can be continually measured and observed in living beings. They are:

- Electrical bio-signal
- Non electrical bio-signal

1.2.1 Electrical Bio-signal

Electrical biosignals are referred to the change in electric current produced by the sum of an electrical potential difference across a tissue, organ or cell system [6]. Best well-known electrical biosignals are:

- Electroencephalogram (EEG)
- Electrocardiogram (ECG)
- Electromyogram (EMG)
- Mechanomyogram (MMG)
- Electrooculogram (EOG)
- Galvanic skin response (GSR)
- Magnetoencephalogram (MEG)

EOG, ECG, EEG and EMG are calculated with a differential amplifier which measures the difference between two electrodes attached on the skin [7]. The galvanic skin response measures the electrical resistance and the MEG measures the magnetic field created by the electrical currents of the brain [8,9]. These signals can be used to operate the HCI systems.

1.2.2 Non-electrical Bio-signal

Non-electrical biosignals are the signals which are generated by the movements of different organs or tissues in living beings without producing any electrical current. For HCI system, various non-electrical biosignals are used. Mechanical signals (e.g. the mechanomyogram or MMG), acoustic signals (e.g. phonetic and non-phonetic utterances, breathing), chemical signals (e.g. pH, oxygenation) and optical signals (e.g. movements) are several non-electrical biosignals which can be used to control HCI systems [5].

1.3 Different Non-Electrical Bio-signal Based HCI System

Many of the human-computer interaction systems are based on non-electrical biosignal. Few of the systems are described as below:

1.3.1 Voice User Interface

Voice interaction is the ability to interact with a device through voice recognition and processing [10]. In modern days, voice interfaces are getting developed rapidly which is brilliant offering to our daily lives. These interfaces are used in various areas such as TVs, smartphones, wheelchairs, directory assistance services (DAS) and a wide range of products [11].

1.3.2 Hand Movement Based HCI System

Most of the human-computer interaction systems are hand controlled. Hands are the most usable organ of the human body. People use them to operate various HCI systems such as mouse, keyboard, cellphones, joystick, joystick controlled wheelchair etc [12].

1.3.3 Touch Sensitive HCI System

Touch sensitive device is an electronic visual display that the user can operate through simple or multi-touch gestures by touching the display with one or more fingers. There are various touch sensitive HCI devices like smartphones, smartwatches, smart televisions, Tablet pcs and so on. Touch screen devices have simpler user interfaces. For people who are new or uncomforTable with normal desktops, touch screens are easy to use [13]. But screens have to be big enough otherwise there is a chance of mistyping. If a touch screen were to crush the whole screen would be unresponsive, that is one of the biggest problems of touch screen devices [14].

1.4 Different Electrical Bio-signal Based HCI Systems

Electrical bio-signal based human-computer interaction systems are new addition to the modern technology and research. Few of the such systems are discussed below:

1.4.1 Electromyography (EMG) Based System

EMG is a device which detects the electrical currents that are generated in a muscle when it is contracting [15]. A muscle fibre contracts when it gets an action potential. EMG based HCI has variety of applications including clinical applications, interactive computer gaming, electric powered wheelchair etc [16]. These systems have some advantages as EMG signals are easy to acquire and have high magnitude compared to other bio-signals. But EMG signals are susceptible to noise such as equipment noise, interaction of different tissues, motion artifacts and electromagnetic radiation etc.

1.4.2 Electrooculography (EOG) Based System

EOG is used to acquire the corneo-retinal standing potential that presents between the front and the back of the human eye. This technique can be used to develop assistive HCI devices. Various HCI devices can be controlled using EOG such as wheelchair, mouse pointer, interactive computer games, smartphones etc [17] [18] [19]. These kind of EOG controlled devices are very useful for the people who are severely paralyzed.

1.4.3 Electroencephalography (EEG) Based System

EEG is the recording of electrical activity of the brain, where the signal originates from post-synaptic potentials and transfers through the skull to the scalp [20]. Brain computer interface is one of the latest HCI systems [21]. It collects EEG data from brain and converts it into device control commands using signal processing techniques. The measured EEG signals are amplified, filtered and digitized in a computer where feature extraction and classification are performed. This is one of the most important technologies for the people who suffer from neuromuscular disorders, since BCI provides them the means of communication, control and rehabilitation tools to restore their lost abilities. EEG techniques are non-invasive but signal processing and pattern recognition is a big challenge.

1.5 Electrooculogram (EOG)

Electrooculogram (EOG) is the electrical signal that corresponds to the potential difference between the retina and the cornea of the eye. This potential can be considered as a steady electrical dipole with a negative pole at the fundus and a positive pole at the cornea. EOG is a steady state potential in which the steady dipole may be used to measure the eye position by placing surface electrodes to the left and right of the eye. When the person looks straight ahead, the steady dipole is symmetrically placed between the two electrodes. In this case, the output is measured to be zero. When the person looks towards left, the positive cornea becomes closer to the left electrode, which becomes more positive. The EOG output for horizontal angle of gaze is found to be approximately ±30° of arc. Electrodes may also be placed above and below the eye to record vertical eye movements. The EOG value varies from 50 to 3500 µV with a frequency range of about 0.15 to 30Hz [22].

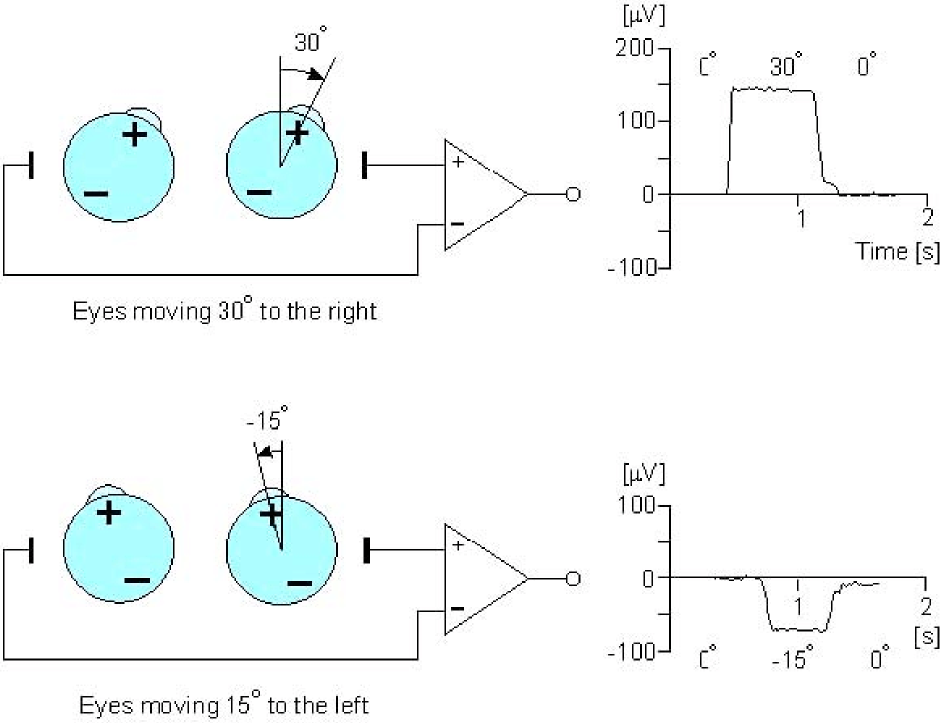


Figure 1.2: Electrooculogram

1.5.1 Human Eye

The normal eye is an approximately spherical organ about 24 mm in diameter. Located at the back of the eye, the retina, is the sensory portion of the eye. The light transmitting parts of the eye are the cornea, anterior chamber, lens and vitreous chamber, named in the order in which these structures are traversed by light. A transparent fluid, the aqueous humor, is found in the anterior chamber. The vitreous chamber is filled by a transparent gel, the vitreous body. The aqueous humor provides a nutrient transport medium, but it is also of further optical significance. It is normally maintained at a pressure(20-25mmHg) that is adequate to inflate the eye against its resistive outer coats (the sclera and choroid) [23].

Eye is one of the most complex sense organs present in the body. The eye can be considered as the master sense organ among all the others. The separation of the two eyes in animals (which is about 6 cm in human beings) helps in forming a distinct image by each eye which is processed by the brain through superimposition of the two images thereby giving a perception of depth and 3-dimensional virtue to the surrounding objects. The curvature and architecture of the eye also helps in identifying the distance of the light source [24].

1.5.2 Types of the Eye Movements

The ocular muscles mentioned in the above paragraph work individually or in synchrony to provide the overall motion of the eyeballs thereby positioning the eye in the direction of vision. These movements, which have been deeply studied in neurophysiology and psychology, can be broadly classified into four categories: saccades, smooth pursuit movements, vergence movements and vestibulo-ocular movements.

Saccades comprise of the short, quick burst of movements that occur in the eye that are associated with the normal functioning such as reading, gazing etc. The Rapid Eye Movement (REM) stage of the sleep which is associated with lucid dreaming also consists of saccades movement of the eye. These movements are mostly reflexive in nature but can also be voluntary [25].

Smooth pursuit movements are voluntary movements of the eye which are much slower in nature. This movement is associated with the tracking of an object which usually occurs after an initial saccade which is used to locate the object [26].

Vergence movements are those which help in proper focusing, pupil dilation, convergence and divergence of the eye based upon the distance of the object. Unlike the other muscles which work uniformly for both the eyes, vergence movements function independently for each such that the perception of depth and distance can be well achieved by the superimposition of the two independent images formed by each eye in the brain [25].

Vestibulo-ocular movements are stabilizing movements of the eye with respect to the movement of the head. These movements prevent in keeping the object under the circle of vision even while the head is moving [25].

1.5.3 Advantages of EOG Signal

EOG signal has some advantages over other signals [27]. These advantages are described as below:

- EOG based recording techniques are simple and cheaper than other methods and can be recorded with minimal discomfort.
- EOG readings can be measured even when eye is closed, for example during sleep.
- Fully or partially abled persons have a dominant vision which can be used as a residual influential tool in developing their rudimentary works through human-computer interfacing.

1.5.4 Disadvantages of EOG Signal

There are few drawbacks of EOG signal [28]. They are given below:

- The corneo-retinal potential is not fixed but has been found to vary diurnally, and to be affected by light, fatigue, and other qualities. Consequently, there is a need for frequent calibration and recalibration.
- Additional difficulties arise owing to muscle artifacts and the basic nonlinearity of the method.
- EOG signals are subject specific. Signal amplitude and duration vary from person to person.
- EOG signal amplitude is of microvolt range and highly susceptible to noise.
- EOG signals are very much sensitive and therefore fluctuate with head movements.

1.6 Machine Learning

Machine Learning (ML) is the study of statistical models and algorithms that computers use to efficiently execute a specific task without using direct instructions but depending on patterns and inference [29]. In artificial intelligence, machine learning is one of the fundamental things. In order to make predictions for doing a specific task, machine learning algorithms create a mathematical model from sample data. This is known as ‘training data’.

There are many machine learning algorithms available to make the HCI devices work effectively [30]. Some example of these algorithms are given below:

1) Support Vector Machine (SVM), 2) Random Forest (RF), 3) K-Nearest Neighbor (KNN), 4) Logistic Regression (LR), 5) Bagging, 6) Naïve Bayes, 7) Artificial Neural Network (ANN) etc.

1.7 Correlation

Data correlation is a widely used term which define the relationship and association among quantities. It predicts one to another quantity by describing association between random variables [31]. So, it can be used for data analysis to know how one variable is depended on another variable.

1.8 Scope of the Study

People are inventing new technologies every day. These modern technologies make life lot easier for the people. HCI devices are one of the newest additions in this technological advancement. Most of the HCI devices are hand controlled. Nowadays, hands-free devices are becoming very popular [32]. Voice controlled devices are already being used [33]. But eye controlled devices are still under research. If these devices are implemented in day-to-day life, they will be very helpful for the able-bodied as well as paralyzed people. Mainly there are two types of paralysis. One is partial paralysis and other is full-body paralysis. The people who are suffering from full-body paralysis is referred to quadriplegia. For quadriplegic people, eye controlled HCI devices can be very helpful. These devices can also be useful for partially paralyzed people as well as able-bodied people. This research can help to develop an intelligent system which can detect eyeball movement and control other devices according the eyeball directions.

1.9 Motivation

Nowadays computer operated devices are very popular among people. Day by day people are becoming more dependent on these kind of devices. To make these devices intelligent and useful, computer needs to understand and analyze the signal which has been picked from human. Machine learning concept is introduced to help the computer understood those signals. With this new concept, the HCI devices can be handled with ease and accurately. These devices are not only useful for able bodied human beings but also for paralyzed people. Hands free HCI devices can be very helpful for those paralyzed persons. Even if a person is full body paralyzed (quadriplegia), he/she can move his/her eyeball. Eyeball contolled devices can be very effective for paralyzed as well as normal people. Based on these concepts, eyeball movements are classified in this research to make an intelligent multiple class HCI system.

1.10 Objective of the Study

In this research, EOG data have taken from the subjects when they are looking at different directions. The main objectives of the study are given below:

- To make a detection system for eyeball movement.
- To develop a multiple class HCI system.
- To classify the dataset taken from the subjects using different algorithms.
- To find the correlation between the dataset and the selected features.
- To show the comparison among different algorithms for better accuracy and performance.

1.11 Thesis Outline

The followings are the outline of the thesis:

- In **Chapter I,** the introduction of the thesis is reported. This chapter explains about Human-Computer Interaction Systems, different types of HCI, EOG based HCI and machine learning. Finally, the objectives of the thesis are briefly discussed.
- In **Chapter II,** background knowledge including some important terminologies of the thesis are described.
- In **Chapter III,** a comphrensive literature survey of EOG data classification is presented. The investigations of the previous researchers are reported here.

- In **Chapter IV,** methodology of the experiment is discussed. It includes experimental setup, classification algorithms and procedure, correlation test and methods.

- In **Chapter V,** findings of this study (Results & Discussions) are presented. This chapter shows the result of different classification algorithms as well as performance test among the classifiers. It also includes the correlation test result on EOG dataset.

- In **Chapter VI,** summary of the findings of the study and some recommendations for future research have been presented.

CHAPTER II

BACKGROUND

2.1 Introduction

Background knowledge is very important before starting a research. Without proper backgrund knowledge, a successful research is not possible. Gathering information related to the research work is said to be the background. The gathered information should be analyzed to get a rough idea about the research. Here some important terms related to the classification and correlation methods for developing the multiple class HCI system are described.

2.2 Some Important Terminologies

Some important terminologies related to this research are given below:

2.2.1 Machine Learning

Machine learning is one of the most important factors in artificial intelligence. It focuses on gathering information from their experience and making predictions based on that. There are mainly four types of machine learning [34]. They are listed as follows:

1. **Supervised Learning:** This learning algorithm creates a mathematical model of a set of data that has both the inputs and the desired outputs. Classification and regression are included in supervised learning algorithm.

2. **Unsupervised Learning:** This learning algorithm finds out similarities in the data and take action based on the absence or presence of such similarities in each new piece of data. This unsupervised learning can be used to estimate density in statistics and in some other fields also.

3. **Semi-supervised Learning:** In semi-supervised learning, the combination is used to create the desired result. In real world the available data are a combination of labelled and unlabelled data.

4. **Reinforced Learning:** In reinforced learning, the machine learns from the past experience and captures the best knowledge to give accurate decisions based on the received feedback. The machine is trained by trial and error method.

2.2.2 Attributes

In machine learning, each object is described by a number of variables that is related to its properties. These variables are called attributes. There are two types of attributes: Categorical and Continuous Attributes. Categorical attributes are corresponding to nominal, binary and ordinal variables. And continuous attributes are corresponding to integer, interval-scaled and ratio-scaled variables. There is another attribute called 'ignore attribute', corresponding to the variables which do not have significance for the application [35].

2.2.3 Instances

When a set of variable values is corresponding to each of the objects, it is called a record or instances. The complete set of data available for a task is said to be a dataset. A dataset illustrated as a Table, with each row indicating an instance. Each column has the value of the variables for each of the instances [36].

2.2.4 Data Cleaning

In machine learning, preparing the dataset is a very important step. For preparing the dataset, data cleaning is a must. Data cleaning is a procedure to remove noise and inconsistent data from the dataset. It is time-consuming and labor-intensive procedure but it is absolute necessary for successful data mining [36].

2.2.5 Cross-validation

In machine learning, a model cannot be fit on the training data and it is uncertain that the model will work. For this, it has to assured that the model gets the correct patterns from the data and it is not getting too much noise. Cross-validation can be used for this reason. It is a technique in which the model is trained using the subset of the dataset and is assessed using the complementary subset of the dataset [37]. There are three steps involved in cross-validation which are given below:

- Some portion of the sample dataset has been reserved.

- Train the model using the rest dataset.
- The model should be tested using the reserve portion of the dataset.

2.2.6 Percentage Split

Percentage split in machine learning is a re-sampling method where the mentioned percentage will be set for training the model and the rest of the data will be set for testing that model. Against the trained data, the algorithms is trained and the accuracy is measured on the test data set.

2.2.7 Kappa Statistic

Cohen’s Kappa statistic measures inter-rater agreement for categorical items. It is more like a robust measure than simple percentage agreement measurement, as k considers the possibility of agreement occurring by chance. Some researchers have proposed that it is conceptually easier to calculate the disagreement between items [38].

2.2.8 Mean Absolute Error (MAE)

In a set of predictions, mean absolute error calculates the average magnitude of the errors without considering their directions. It’s the average over the test sample of the absolute differences between prediction and actual observation. In that all individual differences have equal weight [39]. By considerin scatter plot of n points where point j has coordinates of y_j and $\widehat{y_J}$, it is written that

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \widehat{y_J}| \tag{1}$$

2.2.9 Root Mean Squared Error (RMSE)

RMSE is a measure of the differences between sample/population values by a model and the values observed. It represents the square root of the second sample moment of the differences between predicted values and observed values [39].

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \widehat{y_J})^2} \tag{2}$$

2.2.10 Relative Absolute error (RAE)

The relative absolute error is relative to a simple predictor, which is just the average of the actual values. Instead of the total squared error, this error is just the total absolute error. It takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor [40].

$$E_i = \frac{\sum_{j=1}^{n} |P_{ij} - T_j|}{\sum_{j=1}^{n} |T_j - \bar{T}|} \tag{3}$$

Here,

$E_i = Relative\ absolute\ Error$

P_{ij} = the value predicted by the individual program *i* for sample case *j* (out of *n* sample cases)

T_j = target value for sample case *j*

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

2.2.11 True Positive Rate (TPR) or Recall

True positive rate measures the proportion of actual positives that are accurately identified [41].

$$TPR = \frac{TP}{P} = \frac{TP}{(TP+FN)} = 1 - FNR \tag{4}$$

Here,

TP = True Positive ; P = Condition Positive ; FN = False Negative ; FNR = False Negative Rate

2.2.12 False Positive Rate (FPR)

False positive rate is an error in data mining in which a test result improperly shows the presence of a condition that is not actually present [41].

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = 1 - TNR \tag{5}$$

Here, FP = False Positive ; N = Condition Negative ; TN = True Negative ; TNR = True Negative Rate

2.2.13 Precision

Precision is the fraction of relevant instances among achieved instances. It can can be calculated at a given cut-off rank and also considered the retrieved documents [41]. Precision can be written as below:

$$Precision = \frac{|\,(relevant\ document) \cap (retrieved\ document)|}{|(retrieved\ document)|} \tag{6}$$

2.2.14 F-measure

F-measure is the harmonic mean of precision and recall. It combines both precision and recall. It is approximately the average of the two [41]. Traditional F-measure is given by:

$$F = \frac{2\,.precision\,.recall}{(\,precision + recall)} \tag{7}$$

2.2.15 Receiver Operating Characteristic (ROC) Curve

Receiver operating characteristic(ROC) curve is a graphical plot that describes the diagnostic ability of a binary classifier system as its discrimination threshold is checked. The ROC curve is showed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [42]. ROC curve is a comparison of TPR and FPR as the criterion changes.

2.2.16 Correlation

In almost every business there are some correlation with each individual products and correlation helps to understand relationship between them to build a better business statistically.

2.2.16.1 Types of Correlation

Correlation can be categorized in three types [43]. They are:

- Positive Correlation: Positive correlation means that the variables are related with each other and move to the same direction. So, the values of the variables increase proportionally.

- Negative Correlation: If as a result of increased value of one variable, the value of another variable decrease then it is known as negative correlation. So, the direction is not same.

- Neutral Correlation: Correlation can also be zero or neutral that means that the variables are not related to each other.

2.2.17 Mathews Correlation Coefficient (MCC)

MCC is a measure of the quality two-class classifications. It is actually a correlation coefficient between the observed and predicted binary classifications. It can be calculated from the confusion matrix [44]. The formula is given below:

$$MCC = \frac{\{(TP\ x\ TN) - (FP\ x\ FN)\}}{\surd\{(TP+FP)\ (TP+FN)\ (TN+FP)\ (TN+FN)\}} \tag{8}$$

2.2.18 Likelihood Ratio

Likelihood ratio is a measurement of comparing the goodness of fit of two statistical models- a null model against an alternative model [45]. It is denoted by λ. The formula is given below:

$$\lambda(x) = \frac{\sup\{\ L\ (\theta \mid x): \theta \in \Theta_0\ \}}{\sup\{\ L\ (\theta \mid x\ ): \theta \in \Theta\}} \tag{9}$$

Where, Θ = parameter space and θ = complement of Θ

2.2.19 Uncertainty Coefficient

Uncertainty coefficeint is a measure of nomianl association. It is based on the information entropy [46]. The uncertainty coefficient is given as:

$$U(X\ |Y) = \frac{H(X) - H(X \mid Y)}{H(X)} \tag{10}$$

Where,

H(X) = Entropy of a single distribution

H(X | Y) = Conditional Entropy

2.2.20 Contingency Coefficient

Contingency coefficient is a coefficient of association which gives the information about two variables or data sets if they are independent or dependent of each other [47]. The equation of contingency coefficient is given below:

$$C = \sqrt{\frac{\chi 2}{N + \chi 2}} \tag{11}$$

2.2.21 Information Gain

Information gain is a measure of the decrease in disorder which can be achieved by partitioning the original data set. It is measured by how much of a term can be utilized for classification of information, as to calculate the importance of rhetorical items for the classification. The formula of the information gain is given as follows:

$$G(D,t) = \sum_{i=1}^{m} P(C_i) \log P(C_i) + P(t) \sum_{i=1}^{m} P(C_i|t) \log P(C_i|t) + P(\bar{t}) \sum_{i=1}^{m} P(C_i|\bar{t}) \log P(C_i|t) \tag{12}$$

Where, C is a set of document collection, in which there is not the feature t [86].

2.2.22 Mean / Arithmetic Mean / Average

Mean is basically the central value of a discrete set of numbers. Mean can be calculated by dividing the sum of values with the number of values. For x_1, x_2,….., x_n samples denoted by $\bar{x}$, the mean can be calculated as follows:

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{13}$$

Where, n = number of values

2.2.23 Variance

In a data set, variance is a calculation of the spread between numbers. It gives information about how far each number in the set from the mean. Variance is one of the important parameters in asset allocation. It is denoted by σ^2. The formula for variance is as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{14}$$

Where,

x_i = the ith data point

$\bar{x}$ = the mean of all data points

n = the number of data points

2.2.24 Standard Deviation

For data set, random variable or probability distribution, the standard deviation is the square root of its variance. The formula for calculating standard deviation is given as below:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad (15)$$

2.2.25 Skewness

Skewness is a measure of symmetry, or more specifically, the lack of symmetry. A distribution or data set, is symmetric if it looks the same to the left and right of the center point [87].

For univariate data X_1, X_2, ..., X_N, the formula for skewness is:

$$g_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^3/N}{s^3} \qquad (16)$$

It can be computed as the adjusted Fisher-Pearson coefficient of skewness:

$$G_1 = \frac{\sqrt{N(N-1)}}{N-2}\frac{\sum_{i=1}^{N}(X_i - \bar{X})^3/N}{s^3} \qquad (17)$$

2.2.26 Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. High kurtosis data sets usually have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers [87].

For univariate data X_1, X_2, ..., X_N, the formula for kurtosis is:

$$Kurtosis = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^4/N}{s^4} \qquad (18)$$

Some sources use the following formula of kurtosis which is referred to as "excess kurtosis":

$$Kurtosis = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^4/N}{s^4} - 3 \tag{19}$$

2.2.27 P-Value

P-value is a measure of the strength of the evidence against the null hypothesis the strength of the evidence that is provided by our sample against the null hypothesis. The P-value is probability of getting the observed value of the test statistic or a value of with even greater evidence against the null hypothesis if the null hypothesis is actually true the probability of getting the observed value of the test statistics or a value with even greater evidence against the null hypothesis if the null hypothesis is actually true [48].

2.2.28 Bar Chart

A bar chart is a chart or graph that displays the values of different categories of data with different lengths as rectangular bars. This bar can be showed vertically or horizontally. A bar graph compares different discrete categories. The specific categories are shown in one axis of the chart and a calculated value is displayed in another axis. It is scaled in such a way that all the data can fit on the graph. Bar chart also displays a graphical view of categorical data. These categories are qualitative. In a column bar chart, the categories are displayed in horizontal axis and the height of the bar shows the values of each category.

2.2.29 Pie Chart

Pie chart is circular statistical graph. To describe numerical proportion, it is divided into slices. The arc length of each slice is proportional to the quantity it displays. In the business world and the mass media, pie charts have been used frequently.There are several types of pie chart such as 3D pie chart, doughnut pie chart, exploded pie chart, polar area diagram, multilevel pie chart, spie chart, square chart.

2.2.30 Histogram

A histogram is an accurate display of statistical information about the numerical data. Karl Pearson is the guy who first used this. Histogram uses the rectangular shape structures in successive numerical intervals of equal size in order to represent the frequency of data items. It is actually an approximation of probability distribution of a continuous variable. To produce a histogram, first the entire range of values is divided into series of intervals and then count the number of values which are fall into each interval. The main difference between the bar chart and histogram is that bar chart deals with two variables whereas histogram deals with one variable.

CHAPTER III

LITERATURE REVIEW

3.1 Introduction

Literature review is the most important work before an experimental study to know the limitation and scope of the research. In this chapter, a detailed literature review on developing a multiple class HCI system for eyeball movement is presented. Firstly,a brief introduction of different available classification techniques is given. Then a detailed review of previous researches on the classification of EOG data with different techniques is presented.

3.2 Overview of Human-Computer Interaction

Dr. Alex Roney Mathew, et al (2011) has given an overview about Human-Computer Interaction (HCI) system in their paper [49]. They show that functionality and usability are very important for HCI. Based on these, there are different types of HCI systems.

Existing HCIs are desisgned based on some important factors such as:

- Physical
- Cognitive
- Affective

Some of the existing HCIs are keyboads, joysticks, tape devices, microphones etc.

Recent HCI devices are combined with existing devices in animation and networking. There are various recent devices like Global Positioning Systems (GPS), thermal vision, Personal Digital Assistant (PDA), Radio Frequency Identification (RFID) etc.

In this overview, it is shown that there are two types of design for HCI system. They are:

- Unimodal
- Multimodal

There are many unimodal interfaces such as face recognition, body movement tracking, eye movement tracking, speech recognition etc.

Multimodal interfaces are very useful for disabled persons. These interfaces can be used in surgeries through visual sensors, automated arms and robots and data processing.

As from this paper, it is understood that advancement in HCI will make life lot easier and will help the people in many ways.

3.3 Classification Using Machine Learning Techniques

Shweta H. Jambukia, et al (2015) has shown a survey on ECG classification according to the arrhythmia types.Different issues in ECG classification, preprocessing techniques, various ANN based classifiers for ECG data are highlighted here [50]. Main issues of ECG classification are:

- Lack of standardization
- Variability of the features
- Individuality of the ECG pattern
- Non existence of optimal classification rules
- Patients may have different waveforms
- Beat variations

According to these researchers, ECG data can be classified into two ways i.e. ECG beat classification and ECG signal classification. For preprocessing and feature extraction most widely used techniques are wavelet and Pan-Tompkins algorithms. It is seen from the survey that neural network based algorithms are best suited for ECG classification.

From this survey, it is seen that most researchers have focused on sensitivity, specificity and accuracy to find the performance of the classifiers.The researchers have shown in the survey that neural networks are best suited for ECG signal classification and as for ECG beat classification, Multilayer Perceptron Neural Network (MLPNN) gives good accuracy.

3.4 Classification of EOG Signals Using Artificial Neural Network and Support Vector Machine

Lim Jia Qi, et al (2018) has shown feature extraction method for the EOG data in their research [51]. Statistical features are calculated using their respective formulae, autoregressive coefficient is derived from Burg method and Power Spectral Density is estimated using Yule-Walker method.

Next eye movements are classified using Artificial Neural Network (ANN) and Support Vector Machine (SVM). Multilayer feed forward ANN is chosen as supervised learning which is implemented with Levenberg-Marquardt algorithm. ANN classification is performed in MATLAB version 2013a.

As for SVM, radial basis function is chosen for classification. It is done in same MATLAB version 2013a as ANN.

After the classification, a comparison between ANN and SVM is shown to decide which is the best suited algorithm for EOG data. In this experiment, researchers have proved that the combination of statistical feature extraction and SVM are better than ANN with accuracy of 69.75%.

3.5 Feature Extraction of EOG Signal

In the last two decades, EOG signal plays a vital role is controlling Human-Computer Interface. S. Aungsakul,et al (2012) has extracted fourteen features from EOG signal before classification [52].

Maximum peak and valley amplitude values, the maximum peak and valley position values, the area under curve value, the number of threshold crossing value, EOG variance are extracted from vertical and horizontal signals of eye movements.

A wireless system (Mobi6-6b, TMS International BV, Netherlands) is used to acquire EOG data from two channels: vertical and horizontal. Data is taken from five volunteers. For each directional movement, 15 datasets were obtained.

After that, features were obtained from EOG data using different formulae and based on some definitions. To remove background noise and avoid involuntary eye movements, theonset threshold is set at 50µV in this research.

Analysis-of-variation (ANOVA) test is performed to find the better feature among all the features for EOG data.

3.6 EOG Signal Detection and Verification

Lawrence Y. Deng,et al (2009) has designed a multi-purpose eye-movement tracking system [53]. Pragmatic relaible low-cost electrodes are chosen for EOG signal recordings. For signal amplification, INA 128 instrumentation amplifier is used. To balance the signal, an adder was added with the instrumentation amplifier.

For this experiment, the researchers have designed a band-pass filter to isolate the disturbance from grid electricity. For isolating the signals of frequency higher than 23.417Hz, one first-class low pass filter and then another second class low pass filter are used.

After analog to digital coversion, the signals are sent to SPCE061A microprocessor through RS-232 interface. By oscillograph, the fuzzy set values for EOG analog signal are obtained.

The system is checked for controlling TV, DIY eye-sight level detection and eye-controlling game after implementing the eye-movement tracking interface. With few adjustment, the researchers has achieved more than 90% precision of the system.

3.7 Sensing and Processing of EOG signal

C. Kavitha et al(2015) has chosen EOG signal for controlling Human Machine Interface (HMI) system as wheelchair motion due to some advantages over other bio-potential signal [54]. These advantages are:

- EOG signal recordings are easy to acquire and cost effective.
- It can be recorded with minimal discomfort.
- EOG can be measured even when the eye is closed.
- For partially or completely disabled people, EOG can be used for controlling HMI to give them mobility.

To pick up horizontal and vertical EOG signals, Ag/AgCl made electrodes are made. For EOG signal processing intrumentation amplifier, band pass filter, low pass filter, DC bias removal and operational amplifier are used.

Java programming in Processing tool synchronizing with Arduino tool are used to classify the signal. For transmitting signal to the wheelchair module, Zigbee is used as transmitter and receiver. To give the direction for this module, EOG data is processed by Arduino. Wheelchair is controlled wirelessly.

In this resarch, scientist has developed a new “Navigation Point Algorithm” to control the wheelchair. With the help of this algorithm, the wheelchair is directed from the start point to the goal point. An obstacle detection system is also installed to avoid collision.

In this study, the researchers have showed that EOG data can be successfully used to control Human Machine Interface.

3.8 Use of Neural Network in EOG Signal

Dr. S. Ramkumar et al (2016) has classified eye movement data using neural network. Data was taken from 10 subjects aged between 21 and 44 years [55]. Here, Feed Forward Neural Network (FFNN) and Time Delay Neural Network (TDNN) are used for classification.

In this study, eight basic eye movements are taken into account and other four additional eye movements such as rapid movement, lateral movement, open and stare are also considered.

EOG signals are obtained by a two channel AD instrumentation bio-signal amplifier. Cup shaped electrodes are placed below and above both eye to pick up the signal. For each eye movement, 10 trials are taken. A notch filter is used while acquiring the EOG data to remove power line artifacts.To remove some other artifacts from the original, a band-pass filter is used.

For feature extraction from each band, algorithm based on the Parseval theorem is proposed. As for classification, FFNN is trained using Levenberg back propagation algorithm to classify EOG data. TDNN with 6,7,8 hidden neuron architecture is also used.

In this study, out of all the samples 75% are used for training and 100% are used to test the network. From this experiment, it is observed that the mean performance of FFNN varied from 80.72% to 91.48%. TDNN is comparatively better for EOG classification. The mean performance of TDNN varied from 85.11% to 94.18%.

3.9 Comparison of K-Nearest Neighbour, Support Vector Machine and Decision Tree for EOG Based HCI

Babita et al (2017) has shown a comparison of KNN, SVM and Decision Tree (DT) classifiers for EOG signal [56]. First, EOG signal is acquired by placing g.LADYbird electrodes in left and right corner of the eye with the help of g.USB amplifier from g.tec machine. Data has been taken from 12 subjects aged 24 to 26 years.

16 features are extracted from the time domain using dual tree complex wavelet transform (DTCWT). There are three steps for wavelet analysis. They are:

- Decomposition
- Thresholding
- Reconstruction

In DTCWT, threshold are used to remove the noise from EOG signal.

After feature extraction, classification is done to implement the EOG based HCI using SVM, KNN and DT.

Performance comparison for the classifier is done based on the confusion matrix, receiver operating characteristics (ROC) and performace indices i.e. sensitivity, specificity, precision, accuracy and F1 score.

The researchers has shown in this study that KNN is the best algorithm for horizontal EOG signal. It has almost 100% accuracy.

3.10 Use of Fisher’s Exact Test

Statistical dependency analysis has a problem to identify the most significant dependency rules. Usually, the significance is measured either by Fisher’s Exact Test of the chi-square measure (χ^2 – measure).

Wilhelmiina Hamalainen has introduced an efficient algorithm to search the top-K globally optimal dependency rules by using Fisher’s Exact Test as a measure function [57]. This algorithm searches for both positive and negative dependency rules. It is based on the common branch-and – bound strategy.

Fisher’s Exact Test and the corresponding goodness measure are used to calculate the statistical significance of dependency rules. The problem of redundancey is also analyzed.

Search algorithm is used to find out the non-redundant dependency rules. Finding the top-K rules is also more efficient as the strictier conditions which are used to prune the search space. The whole search space can be displayed by enumeration tree.

The main idea of the branch-and-bound search is to calculate the best goodness measure,P_F values. The proposed Kingfisher algorithm is well scalable and can handle dense and high-dimensional datasets effectively.

3.11 Performance of Chi-square Test in biological Sequence:

Leila Pirhaji et al (2008) has applied Pearson’s Chi-square test to identify the signals appeared in the whole genome of Escherichia coli [58].

First, DNA sequence is translated in all six frames. Then Open Reading Frame (ORF) is found. The longest sequence in the genome starting with the start codon and ended up with a stop codon gives the information about ORF.

Pearson’s Chi-square test was performed to compare the frequency of nucleotide in the whole genome and the frequency of the one in a window. Chi-square satistic is calculated by

$$\chi 2 \ = \sum_{i=1}^{n} \frac{(O-E)\text{^}2}{E} \tag{20}$$

After Chi-square test, linguistic complexity is applied. Then, the evaluation of complexity in a text, CWF and the evaluation of entropy are calculated.

To estimate the measure around a specific position of an ORF, the measure for each point in the window is computed around the point and the average of measure values of these points is considered.

At the end, Pearson’s Chi-square test is compared with the measure acquired from linguistic complexity. It is seen that the lingusitic complexity is much lower than the chi-square test.

Pearson’s Chi-square test is also used to locate which parts of ORF had significant effect on discrimnating genes from pseudogenes. The resul of this experiment descibes that there is a region near the start codon with high-value of chi-square statistic.

CHAPTER IV

METHODOLOGY

4.1 Introduction

Methodology is an important part of any successful work. The details of theoretical and experimental work procedure of this study are presented in the methodology chapter including theoretical analysis, necessary equipments, experimental setup, electrode placement, data acquisition, data classification, classification algorithm and correlation methods.

4.2 Theoretical Analysis

Theoretical analysis is one of the major parts to perform a successful and productive research. This is the first step for a research. So this has to be done thoroughly. Studying about the chosen research topic is actually called theoretical analysis. Gathering information related to the topic is very important in this analysis. There are several steps involved in theoretical analysis. They are:

- Background Discussion
- Research paper analysis related to the chosen research topic
- Learning necessary methods to handle the hardware
- Learning necessary software

4.2.1 Background Discussion

For this experiment, some important terminologies like machine learning, attributes, instances, data cleaning, cross-validation, kappa statistic, mean absolute error, relatve absolute error, true positive rate, false positive rate, precision, F-measure, ROC curve, likelihood ratio, uncertainty coefficient, contingency coefficient etc. are discussed in the background section.

4.2.2 Research Paper Analysis

Gathering information from the previous research papers plays a vital role in completing the experiment. Previous researches give a proper guideline how to proceed with the research. Analyzing the procedure, advantages and disadvantages of the method used in previous works will help to overcome the problems and find an innovatiove solution for that.

In this experiment, several research papers have been studied related to EOG data classification using machine learning algortihm. This analysis is described in the literature review section.

4.2.3 Learning Hardware Information

Sometimes hardware has been used to perform an experiment. To learn about the hardware is a must. Without knowing the proper way to handle the hardware, it is nearly impossible to complete the research.

In this experiment, Biopaca MP3X Acquisition unit has been used. A thorough study has been done on the hardware, the power connection, the lead connection and the safety measures to acquire the EOG data from a subject.

4.2.4 Learning Different Software

Software is a set of instructions or programmes to operate a computer or a machine. Several software have been used in performing a research. So learning those software is very important.

Firstly, BSL (Biopac Student Lab) Analysis 4.1 software is learned to acquire the EOG data. To perform the classification task, Weka 3.9.2 (Waikato Environment for Knowledge Analysis) software has been learned. For showing the correlation between the features and the class, IBM SPSS (Statistical Package for the Social Sciences) Statistics 25 software has been learned.

4.3 Experimental Procedure for Developing Multiple Class HCI System

There are several steps involved in the experimental procedure for developing multiple class HCI system including necessary equipment, experimental setup, electrode placement, data acquisition, data classification and correlation method. These steps are describes as follows:

4.4 Necessary Equipment

Equipments are the most important element for any type of project. Good quality equipments are always required for a successful experimental investigation. Different equipments required for this thesis work are listed below.

1. BIOPAC MP3X Acquisition Unit
2. Desktop with Biopac Student Lab Software
3. Arduino Board

4.5 Important Accessories

There are some important accessories which are needed to perform the research properly. They are:

1. Electrolyte Gel
2. Disposable Electrodes (6 pieces per subject)
3. Electrode Lead Cables
4. Connection Cables
5. Wall transformer
6. Power Cable
7. LED (Light Emitting Diode)
8. Connecting Wires
9. Battery- 9V

4.6 Experimental Protocol

Some protocols are followed before performing the experiment. These protocols are given as follows:

1. At first, the experimental procedures are explained to the subjects along with experiment details and risk factors.

2. The subject information such as age, gender, height (cm), weight (kg), blood group, eyesight problem, color blind, sleep duration (hrs), left/right handed, blood pressure, heart disease are taken as features.

3. Then the subjects are told to follow the sequence in navigational setup which is shown Figure 4.1.

4.7 Eyeball Movement Control Using Navigational Setup

4.7.1 Arduino Based Navigational Setup

A navigational setup is prepared for this experiment to obtain eyeball movement data in all directions. Nine LEDs are placed on a corksheet putting one in the center, four at each corner, two at the top and down middle position and other two at left and right side middle to prepere the setup. The corksheet is 70cm long and 96cm wide. All these LEDs are navigated using the Arduino system. Atmega328 is used in that sytem to control the LEDs.

4.7.2 Time Frame for Eye Navigation

Time duration for shifting the LED glow from center point to other LED positions is 1200ms. This navigation setup is kept 50cm away from the eye of the subject to acquire proper EOG data. Six disposable electrodes are used for data acquisition. These electrodes are connected to the Biopac MP36 Acquisition Unit through electrode lead cables.

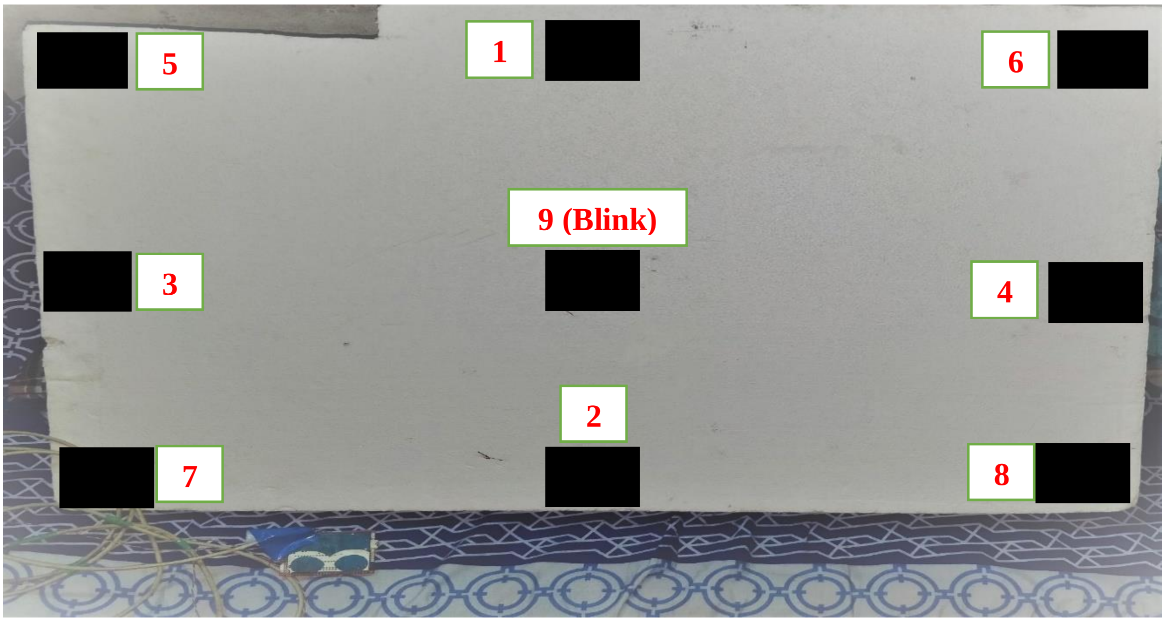


Figure 4.1: Navigational Setup

Two electrodes are used for horizontal channel, two are for vertical channel and other two are used as reference electrodes. Acquisition unit is connected to the desktop by USB connection cables to obtain and analyze the data. In Figure 4.1 the navigational setup system along with Arduino board is shown.

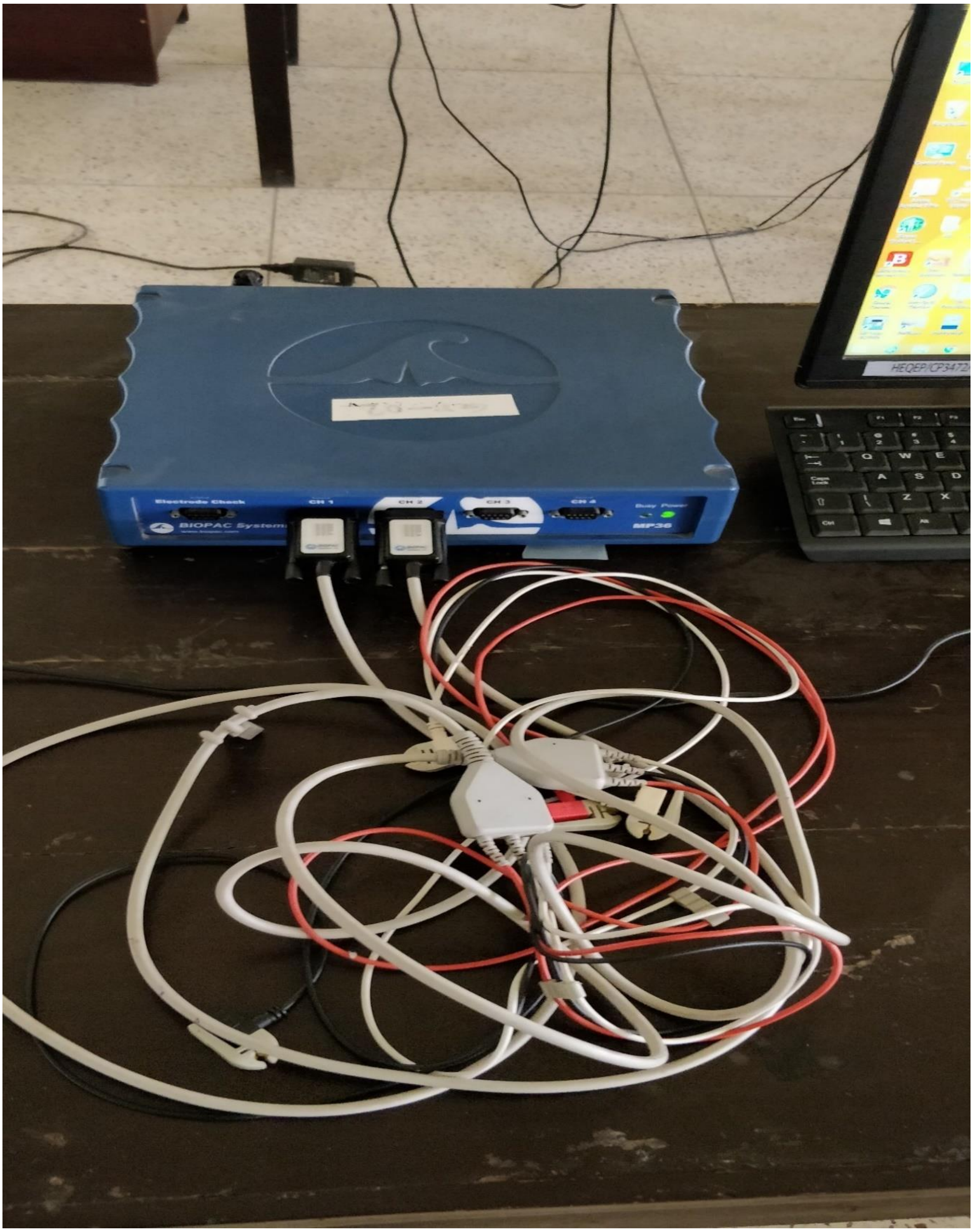


Figure 4.2: Biopac Data Acquisition Unit

In Figure 4.2 Biopac MP36 data acquisition unit and the electrode leads connect to that unit are shown.

4.8 Electrode Placement

Electrodes are one of the main equipments to perform the experiment. Electrode placement is very important to acquire the EOG data. Good experimental result is dependent on the proper placement of the electrodes. There are two channels in Biopac MP3X Acquisition unit for EOG: horizontal and vertical channel. For the horizontal (left-right) eye movement, two electrodes are placed on the temporal bone (squamous) of the both side of the skull. For vertical (up-down) eye movement, one electrode is placed on the supraorbital or superciliary ridge and another is placed on the infra-orbital foramen of the right side skull. Two electrodes are used as reference electrodes. One is placed on glabella of the skull and other is placed on the supraorbital foramen of the left side skull [59].

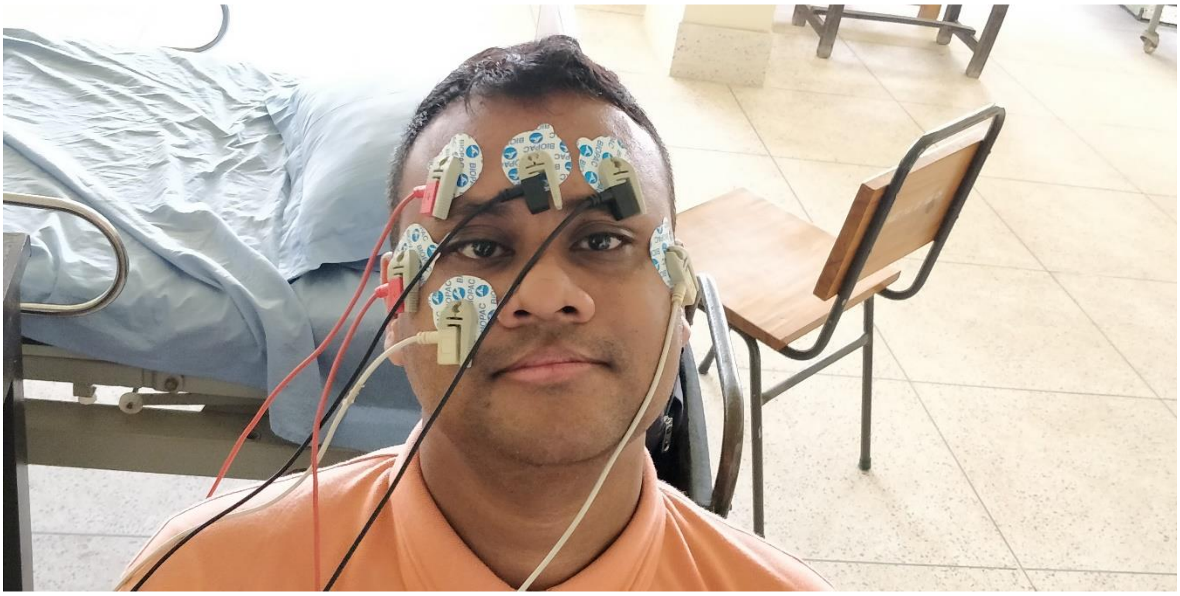


Figure 4.3: Electrode Placement

In Figure 4.3 the electrode placement for this experiment is shown.

4.9 EOG Data Acquisition

7 subjects are chosen for this research aged between 22 to 48 years. Among the subjects, four are male and three are female. All eight basic eye movements as well as blink of the eye are considered to perform this experiment. For each eye movement, 15 trials are taken from all the subjects. Time duration for performing one set of trial (15 trials) is 20 sec.

4.9.1 Software for Data Acquisition

After placing the electrodes on the subject, those electrodes are connected to the Biopac MP3X Acquisition Unit's both channel (horizontal and vertical) through electrode lead cable. Acquisition Unit is connected to the desktop by USB connection cables. EOG data is acquired by Biopac Student Lab software (BSL). Guided lessons and BSL pro options for advanced analysis are included in this software [60]. With the help of this software, the data can be saved easily and reviewed later for further analysis.

4.9.2 Procedure

All the 7 subjects are told to perform the same set of tasks. They are instructed to move their eyeball as the direction of the glowing LED in the navigational setup. In the setup, LEDs are programmed to glow in a pattern. First, it will glow from center to up position. Then it will glow from center to down position. After that the subject is told to look from center to left and then right position as the LED glows in that pattern. When all four major directions are completed, the LEDs are programmed to glow from center to all the corner positions consecutively and the subjects are informed to follow the LED glowing pattern. After all the corner positions, each subject is given task to perform the blink activity for 20 sec with 15 trials. For each eye movement, EOG data is saved in the Biopac Student Lab software as .csv and .matlab format to perform the classification and correlation techniques.

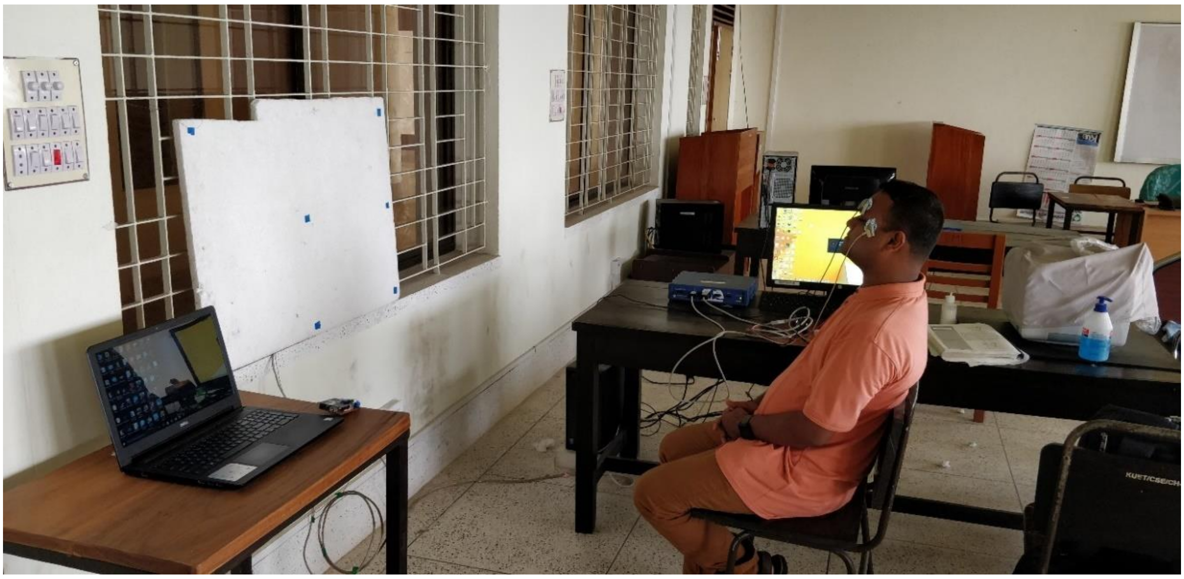


Figure 4.4: Data Acquisition

The whole setup for data acquisition is shown in Figure 4.4.

4.10 Feature Extraction: Finding Mean, Standard Deviation, Kurtosis, Skewness and Variance

In MATLAB, mean, standard deviation, kurtosis, skewness and variance can be extracted from the EOG signal. These are some important features which are important for the EOG dataset. To find those parameters the matlab library functions are used. First all the signal of 7 subjects are laoded into MATLAB. With the help of mean( ), std( ), kurtosis( ), skewness() and var( ) library functions, those parameters of both EOG channels are calculated.

4.11 EOG Data Classification

To develop a multiple class HCI system, it is very important to classify the data accurately using machine learning algorithm. Classification is very much needed to make the system more intelligent and reliable. In this experiment, the EOG data is taken from 7 subjects. There are total 9 types of eye movements which need to be classified. So, total 63 data are collected from the subjects. There are several steps involved in classification procedure of the data. The following flowchart in Figure 4.5 shows the classification steps:

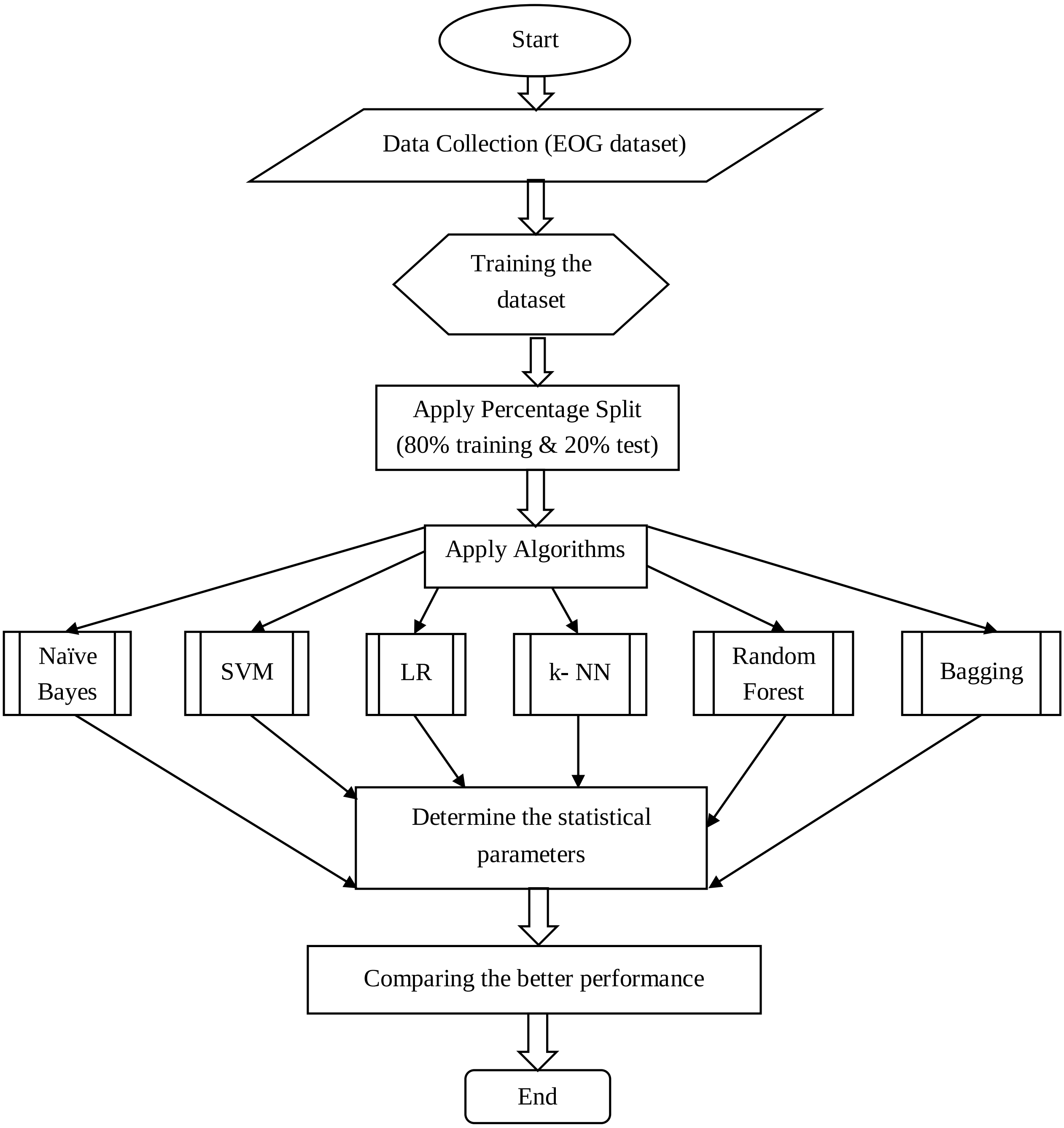

Start
Data Collection (EOG dataset)
Training the dataset
Apply Percentage Split (80% training & 20% test)
Apply Algorithms
Naïve Bayes
SVM
LR
k- NN
Random Forest
Bagging
Determine the statistical parameters
Comparing the better performance
End

Figure 4.5: Flowchart of Classification Procedure

Description of classification steps are given below:

4.11.1 Software Used in Data Classification

For classification purpose, Weka 3.9.2 software has been used. Weka stands for Waikato Environment for Knowledge Analysis. It is a machine learning software which is coded in Java. University of Waikato, New Zealand has developed this software. For data analysis and predictive modeling. Weka has various visualization tools and algorithms along graphical user interfaces [61] . This software has several advantages which are given below:

- Under the General Public License, Weka has free availability.
- It has good portability as it is completely implemented in the Java programming language and can work on any latest computing platform.
- It has a comprehensive collection of data preprocesing and modeling techniques.
- It's graphical user interfaces make it very easy to use.

Weka can perform several standard data mining tasks more precisely i.e.data preprocessing, clustering, classification, regression, visualization and feature selection. Weka's techniques are utilized on the prediction that the data is one flat file or relation type data [36].

4.11.2 Classification Algorithms

Each classification algorithm has a different way to classify the data. Some algorithms have There are some similarities among the classifiers. The classification algorithms used in this experiment are discussed below:

4.11.2.1 Naïve Bayes Classifier

Naïve Bayes classifiers are simple probabilistic classifiers. In this classifier, Bayes' theorem is applied with strong independenece assumptions between the features. It is lightly scalable and a number of parameters linear in the number of variables in a learning problem is required for this [62] [85].

4.11.2.1.1 Mathematical Model of Naïve Bayes

Naïve Bayes is a conditional probability model. Suppose a problem needs to be classified which is represented by a vector $x = (x_1,\ldots\ldots, x_n)$ denoting some n features (independent variables). It indicates this instance probabilities as $p(C_k \mid x_1,\ldots., x_n)$ for each of k possible outcomes or classes C_k.

Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k \mid x) = \frac{p(C_k)\, p\,(x \mid C_k\,)}{p\,(x)} \tag{21}$$

Using Bayesian probability, the above equation can be written as

$$Posterior = \frac{Prior \times Likelihood}{Evidence} \tag{22}$$

Using the chain rule, the conditional probability can be showed as

$$p\ (C_k\,, x_1, \ldots., x_n) =$$
$$p\ (\,x_1 | x_2, \ldots, x_n, C_k)\ p\ (x_2 |\ x_3, \ldots., x_n, C_k) \ldots.\ldots p\ (x_{n-1} |\ x_n, C_k)\ p\ (x_n \mid C_k)\ p\ (C_k) \tag{23}$$

For conditional independence assumptions having feature x$_j$ for j≠I, it can be written,

$$p\ (x_i \mid x_{i+1}, \ldots, x_n, C_k) \approx p\ (x_i |\ C_k) \tag{24}$$

The joint model can be defined as

$$p\ (\,C_k \mid x_1, \ldots., x_n) \propto p\ (C_k, x_1, \ldots\ldots, x_n)\ \approx p\ (C_k)\ p\ (x_1 | C_k)\ p\ (\,x_2 | C_k) \ldots\ =$$
$$p\ (C_k) \prod_{i=1}^{n} p\ (\,x_i \mid C_k) \tag{25}$$

The conditional distribution over the class variable C under the above independence assumptions

$$p\ (C_k \mid x_1, \ldots., x_n)\ = \frac{1}{z}\ p\ (C_k)\ \prod_{i=1}^{n}\ p\ (\,x_i \mid C_k\,) \tag{26}$$

where the evidence $z = p\ (x) = \sum_k p\ (\,C_k)\ p\ (\,x \mid C_k)$

Bayes classifier combines this model with a decision rule [63]. It is the function that assigns a class label $\hat{y} =\ C_k$ for same k as follows:

$$\hat{y} = \underset{k \in \{1,\ldots,,K\}}{argmax}\ p\ (\,C_k) \prod_{i=1}^{n} p\ (\,x_i \mid C_k) \tag{27}$$

4.11.2.1.2 Implementation of Naïve Bayes in Weka

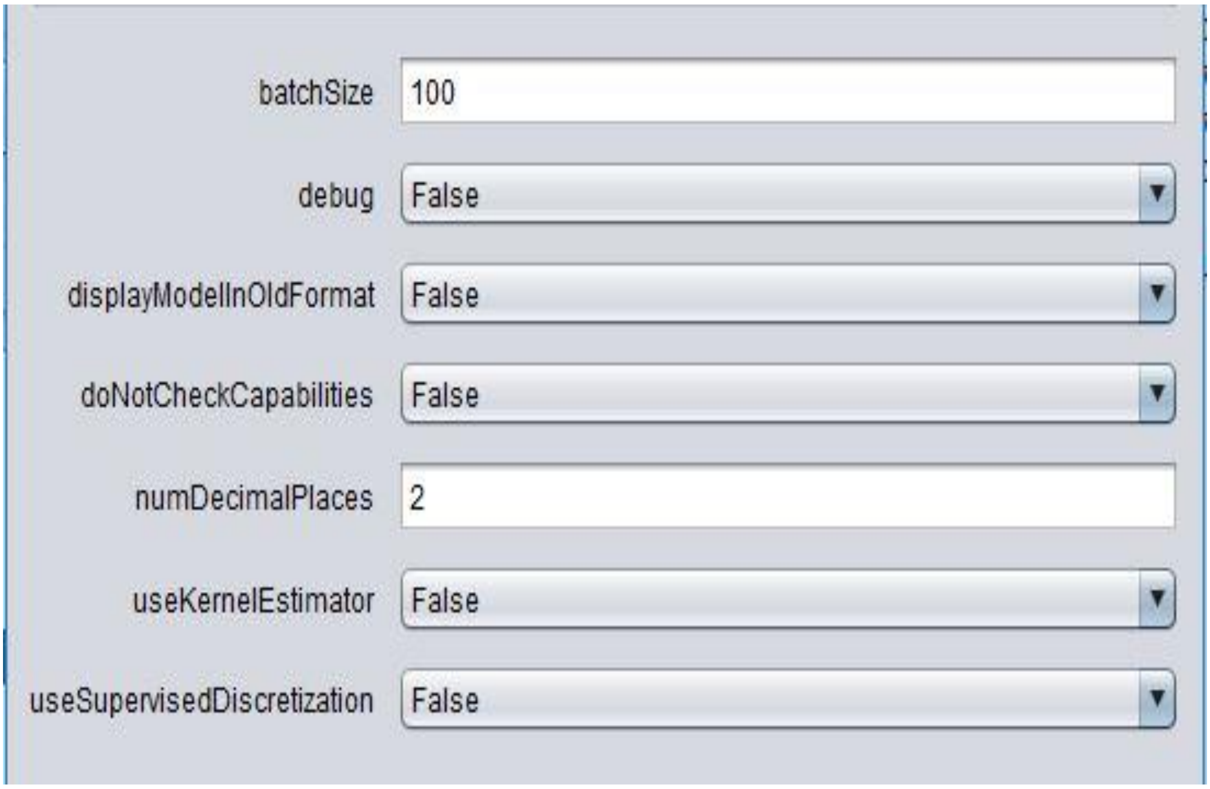
In Weka 3.9.2, first ‘Explorer’ application is selected to insert and classify the data. After inserting the data using ‘Preprocessing’ tab, ‘Classify’ is clicked to perform classification process on the dataset.

In ‘Classify’ window ‘Choose’ option is clicked to select the classifier. In ‘bayes’ category, the ‘NaiveBayes’ classifier is chosen. After this, ‘Percentage Split’ is set to 80% from the ‘Test Option’ which is shown just below the ‘Choose’ menu.

Then by clicking the ‘Start’, the naïve bayes classifier is used to classify the EOG dataset.

4.11.2.1.3 Properties of Naïve Bayes

Properties of the naïve bayes classifier are given below:



batchSize	100
debug	False
displayModelInOldFormat	False
doNotCheckCapabilities	False
numDecimalPlaces	2
useKernelEstimator	False
useSupervisedDiscretization	False

Figure 4.6: Properties of Naïve Bayes

The properties of naïve bayes classifier is shown in Figure 4.6 where it is seen that the batchsize is 100 and the debug mode is kept false.

4.11.2.1.4 Advantages of Naïve Bayes Algorithm

There are various advantages of Naïve Bayes Algorithm [64]. They are given below:

- Easy to use, very simple and fast.

- Need less training data.

- For both binary and multi-class classification problems, it can be used.
- It can make probabilistic predictions.
- It is not sensitive to irrelevant features.

4.11.2.1.5 Disadvantages of Naïve Bayes Algorithm

There are few disadvantages of naïve bayes which are given below [64]:

- Class conditional independence in this classifier because of that there is loss of accuracy.
- Dependencies among variables cannot be modelled by naïve bayes classifier.

4.11.2.2 Support Vector Machine (SVM)

SVM is a classifier which finds a hyperplane in an N-dimensional space (N-th number of featues) that distintly classifies the data points. Hyperplanes are decision boundaries help classify the data points. The dimension of the hyperplane depends on the number of features. Support vectors are data points that are closer to the hyperplane and effect the position and orientation of the hyperplane [65].

4.11.2.2.1 Mathematical Model of Support Vector Machine

A training dataset of n points of the form is given as $(\overrightarrow{x_1}, y_1), \ldots\ldots, (\overrightarrow{x_n}, y_n)$ where the y_i are either 1 or -1.

Any hyperplane can be written as the set of points $\vec{x}$ statisfying

$$\vec{w}.\vec{x} - b = 0 \tag{28}$$

where $\vec{w}$ is the normal vector to the hyperplane

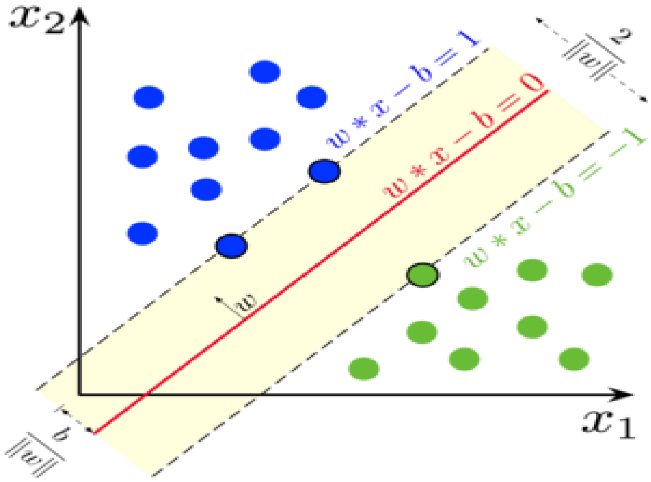


Figure 4.7: Maximum-margin hyperplane and margins for an SVM [60]

In Figure 4.7 maximum-margin hyperplane and margins for an SVM trained with samples from two classes are shown. Samples placed on the margin are called the support vectors [66].

SVM classifier can be written by minimizing an expression of the form

$$[\frac{1}{n}\sum_{i=1}^{n} \max(\,0, 1 - y_i\,(\,w.x_i - \,b))] + \lambda\|w\|^2 \qquad (29)$$

4.11.2.2.2 Implementation of SVM in Weka

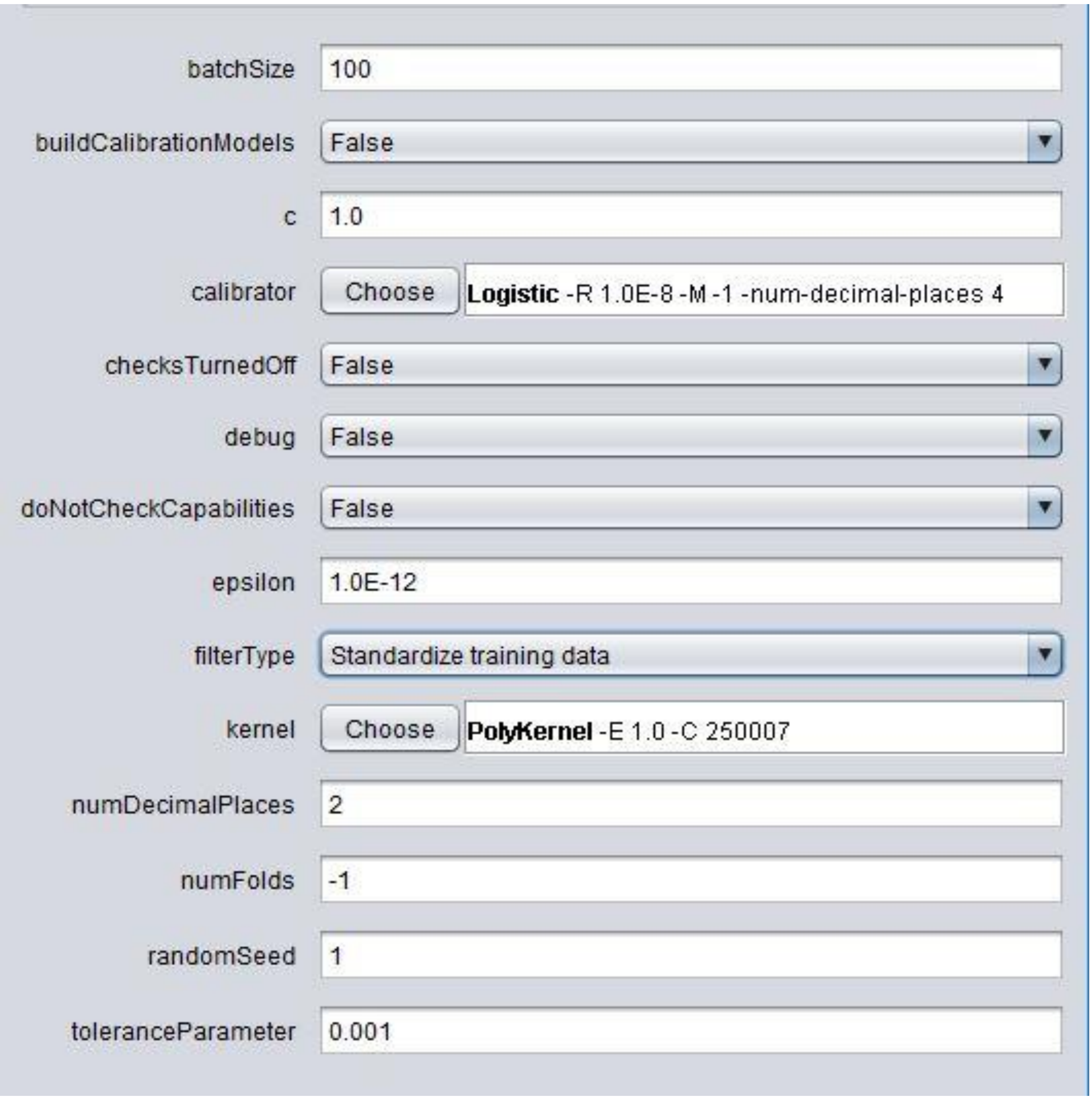
In Weka 3.9.2, first ‘Explorer’ application is selected to insert and classify the data. After inserting the data using ‘Preprocessing’ tab, ‘Classify’ is clicked to perform classification process on the dataset.

In ‘Classify’ window ‘Choose’ option is clicked to select the classifier. In ‘functions’ category, the ‘SMO’ classifier is chosen which is the ‘Support Vector Machine’ classifier. After this, ‘Percentage Split’ is set to 80% from the ‘Test Option’ which is shown just below the ‘Choose’ menu.

Then by clicking the ‘Start’, the SVM classifier is used to classify the EOG dataset.

4.11.2.2.3 Properties of SVM Classifier

Properties of the SVM are given below:



batchSize	100
buildCalibrationModels	False
c	1.0
calibrator	Choose **Logistic** -R 1.0E-8 -M -1 -num-decimal-places 4
checksTurnedOff	False
debug	False
doNotCheckCapabilities	False
epsilon	1.0E-12
filterType	Standardize training data
kernel	Choose **PolyKernel** -E 1.0 -C 250007
numDecimalPlaces	2
numFolds	-1
randomSeed	1
toleranceParameter	0.001

Figure 4.8: Properties of SVM

In Figure 4.8 it is seen that the batchsize of SVM classifier is 100. Calibrator chosen for SVM is logistic and the kernel used in that classififer is polykernel. Filter type is chosen as standardize training data.

4.11.2.2.4 Advantages of SVM

Various advantages of SVM are described below [66]:

- SVM can perform even if there is not much information on the data.
- For unstructured and semi structured data like text, images and trees, it works well.
- Complex problem can be solved with an appropriate kernel function of SVM.
- For high dimensional data, SVM is properly scalable.
- The risk of overfitting is less in SVM.

4.11.2.2.5 Disadvantages of SVM

Few disadvantages of SVM are given as follows [66]:

- Choosing a good kernel function for SVM is hard.
- For large datasets, it requires long training time.
- The final model is difficult to understand and interpret.

4.11.2.3 Logistic Regression (LR)

LR is a statistical method to describe data and to explain the relationship between one independent variable and one or more nominal, ordinal, interval or ratio-level independent variables [67]. It is a predictive analysis. It analyzes a dataset in which there are one or more independent variables that provides the outcome. The outcome is calculated with a dichotomous variable. This does not perform statistical classification but it can be used to make a classifier .

4.11.2.3.1 Mathematical Model of Logistic Regression

Logistic Regression provides the coefficients of a formula to predict a logit transformation of the probability of presence of the characteristic of interest.

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k \tag{30}$$

where p is the probability of presence of the characteristic of interest [67]. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence} \tag{31}$$

and
$$logit\ (p) = \ln(\frac{p}{1-p}) \tag{32}$$

4.11.2.3.2 Implementation of Logistic Regression in Weka

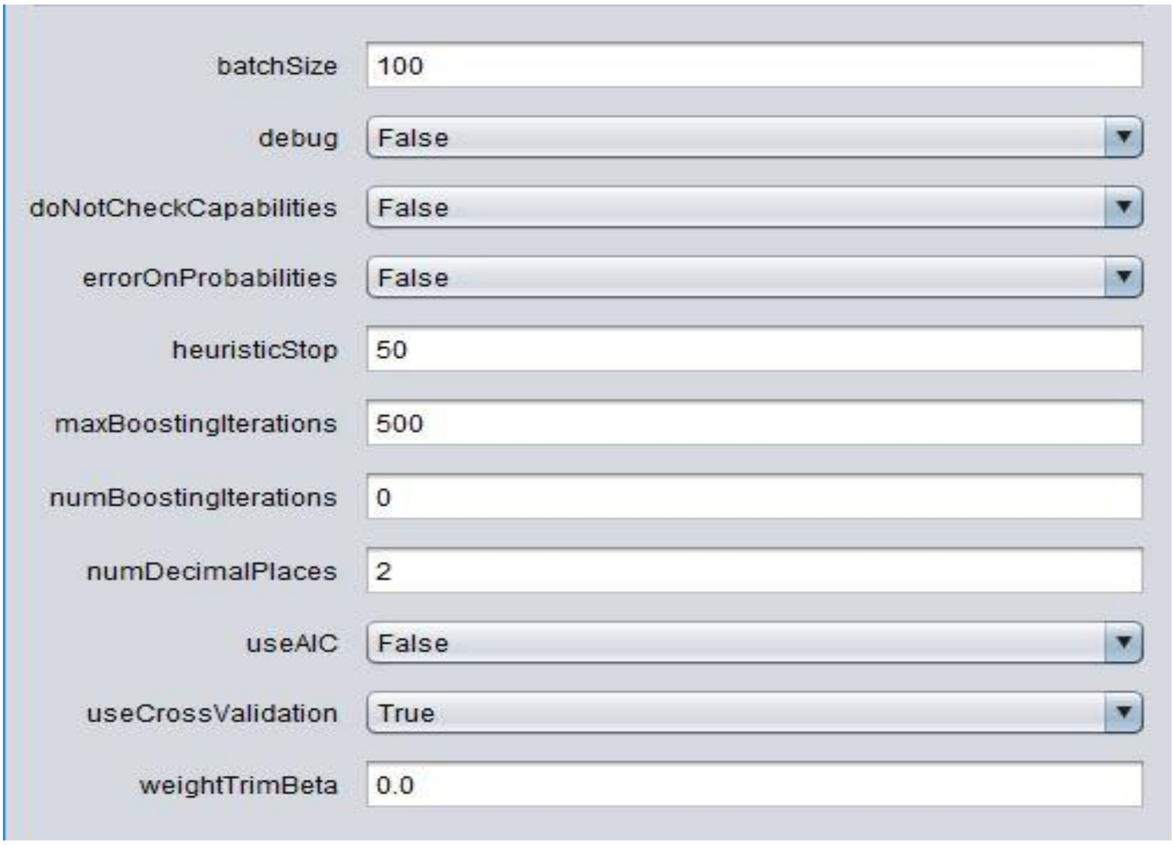
In Weka 3.9.2, first ‘Explorer’ application is selected to insert and classify the data. After inserting the data using ‘Preprocessing’ tab, ‘Classify’ is clicked to perform classification process on the dataset.

In ‘Classify’ window ‘Choose’ option is clicked to select the classifier. In ‘functions’ category, the ‘SimpleLogistic’ classifier is chosen. After this, ‘Percentage Split’ is set to 80% from the ‘Test Option’ which is shown just below the ‘Choose’ menu.

Then by clicking the ‘Start’, the LR classifier is used to classify the EOG dataset.

4.11.2.3.3 Properties of Logistic Regression

Properties of LR classifier is shown below:



batchSize	100
debug	False
doNotCheckCapabilities	False
errorOnProbabilities	False
heuristicStop	50
maxBoostingIterations	500
numBoostingIterations	0
numDecimalPlaces	2
useAIC	False
useCrossValidation	True
weightTrimBeta	0.0

Figure 4.9: Properties of Logistic Regression

From the properties of Logistic Regression shown in Figure 4.9, it is seen that batchsize is 100 and debug mode is kept ‘False’.

4.11.2.3.4 Advantages of Logistic Regression

The advantages of LR as classifier are given below [68]:

- For LR, the independent variables are not normally distributed or have equal varience in each group..
- Does not predict a linear relationship between independent and dependent variables.
- It can deal with nonlinear effects.
- Explicit interaction and power terms can be added.
- In LR, normally distributed errors are not assumed.
- The independents in LR does not need to be interval or unbounded.

4.11.2.3.5 Disadvantages of Logistic Regression

Disadvantages of LR are specified as follows [69]:

- It cannot identify correct independent variables.
- It has limited outcome variables.
- Independent observations are required for logistic regression.
- Sometimes overfitting of the model occurs in logistic regression.

4.11.2.4 K-Nearest Neighbour

K-nearest neighbour is a simple algorithm based on a similarity measure i.e. distance functions. It stores all available cases and classifies new cases. In statistical estimation and pattern recognition, KNN has been used frequently [70].

4.11.2.4.1 Mathematical Model of K-Nearest Neighbour

A case is classified by a majority vote of its neighbors with the case being assigned to the class most common among its K nearest neighbors. It is calculated by a distance function. If k=1, then the case is alloted to the class of its nearest neighbor [71].

Some distance functions are given below:

$$Euclidean\, function = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{33}$$

$$Manhattan\, function = \sum_{i=1}^{k}|x_i - y_i| \tag{34}$$

$$Minkowski\, function = (\sum_{i=1}^{k}(|x_i - y_i|)^q)^{\frac{1}{q}} \tag{35}$$

These distance functions are only valid for continous variables. For categorical variables, Hamming distances are used. Hammning distance is given as:

$$D_H = \sum_{i=1}^{k}|x_i - y_i| \tag{36}$$

if $x = y \rightarrow D = 0\, and\, x \neq y \rightarrow D = 1$

4.11.2.4.2 Implementation of K-Nearest Neighbor in Weka

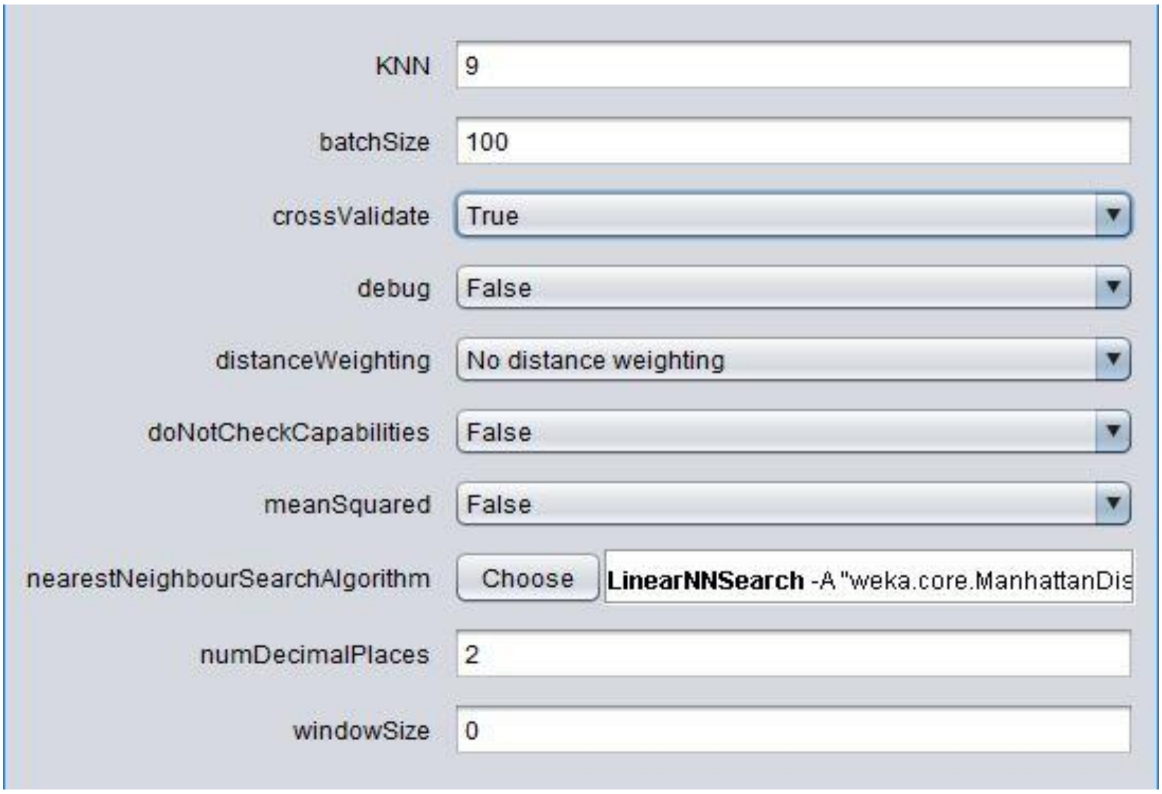
In Weka 3.9.2, first ‘Explorer’ application is selected to insert and classify the data. After inserting the data using ‘Preprocessing’ tab, ‘Classify’ is clicked to perform classification process on the dataset.

In ‘Classify’ window ‘Choose’ option is clicked to select the classifier. In ‘lazy’ category, the ‘lBk’ classifier is chosen which is the K-Nearest neighbor. After this, ‘Percentage Split’ is set to 80% from the ‘Test Option’ which is shown just below the ‘Choose’ menu.

Then by clicking the ‘Start’, the KNN classifier is used to classify the EOG dataset.

4.11.2.4.3 Properties of K-Nearest Neighbor

The properties of KNN classifier is displayed below:



KNN	9
batchSize	100
crossValidate	True
debug	False
distanceWeighting	No distance weighting
doNotCheckCapabilities	False
meanSquared	False
nearestNeighbourSearchAlgorithm	Choose **LinearNNSearch** -A "weka.core.ManhattanDis
numDecimalPlaces	2
windowSize	0

Figure 4.10: Properties of KNN

In Figure 4.10 it is seen that 9 classes are selected for KNN classifier and batch size is kept 100. The debug mode is chosen as ‘False’ and nearest neighbour search algorithm is chosen as “LinearNNSearch’.

4.11.2.4.4 Advantages of K-Nearest Neighbor

The advantages of KNN are given as follows [72] [73]:

- KNN classifier is very simple and intuitive.
- It has no assumptions.
- Training step is not required for this.
- For multiple class problem, it is very easy to implement.
- It can be used for both classification and regression.

4.11.2.4.5 Disadvantages of K-Nearest Neighbor

The disadvantages of KNN are describes below [72] [73]:

- It is a slow algorithm.
- For large number of variables, KNN struggles to predict the output of new data.
- KNN needs homogeneous features.
- KNN is not applicable for imbalanced data.
- Missing value problem cannot be dealt by this classifier.

4.11.2.5 Random Forest

Random Forest is an ensemble learning method for classification and regression. It creates the forest with a number of trees. It contructs a multitude of decision trees and shows the class as output which is the mode of the classes or mean prediction of the individual trees [74]. As they utilize a set of results to make a final decision, they are called Ensemble techniques.

4.11.2.5.1 Mathematical Model of Random Forest

In random forest, feature importance is estimated as the decrease in node impurity weighted by the probability of reaching that node. The higher the value the more important the feature. For each decision tree, Scikit-learn calculates a nodes importance using Gini importance, assuming only two child nodes [75]:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{37}$$

Where

$ni_j = the\ importance\ of\ node\ j$

$w_j = weighted\ number\ of\ samples\ reaching\ node\ j$

$C_j = the\ impurity\ value\ of\ node\ j$

$left\ (j) = child\ node\ from\ left\ split\ on\ node\ j$

$right\ (j) = child\ node\ from\ right\ split\ on\ node\ j$

The importance for each feature on a decision tree is then measured as:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k\in all\ nodes} ni_k} \tag{38}$$

Where

$fi_i = the\ importance\ of\ feature\ i$

$ni_j = the\ importance\ of\ node\ j$

These are noramlized between 0 and 1 by dividing by the sum of all feature importance values:

$$norm\ fi_i = \frac{fi_i}{\sum_{j\in all\ features}\ fi_j} \tag{39}$$

The final feature importance is it's average over all the trees. It is given as:

$$RF\ fi_i = \frac{\sum_{j\in all\ trees} fi_{ij}}{T} \tag{40}$$

Here,

RF fi_i = the importance of feature I calculated from all trees in the Random Forest model

norm fi_ij = the normalized feature importance for i in tree j

T = total number of trees.

Spark measures a feature's importance by summing the gain, scaled by the number of samples passing through the node [75]:

$$fi_i = \sum_{j:nodes\ j\ splits\ on\ feature\ i} s_i C_j \tag{41}$$

Where,

$fi_i = the\ importance\ of\ feature\ i$

$s_j = number\ of\ smaples\ reaching\ node\ j$

$C_j = the\ impurity\ value\ of\ node\ j$

At random forest level, the feature importance for each tree is normalized in relation to the tree:

$$norm\ fi_i = \frac{fi_i}{\sum_{j\ \in all\ features} fi_j} \tag{42}$$

Where,

$norm\ fi_i = the\ normalized\ importance\ of\ feature\ i$

$fi_i = the\ importance\ of\ feature\ i$

From each tree, feature importance values are normalized as :

$$RF\ fi_i = \frac{\sum_j norm\ fi_{ij}}{\sum_{j\ \in all\ features,\ \ k\ \in all\ trees} norm\ fi_{jk}} \tag{43}$$

Where,

RF fi_i = the importance of feature I calculated from all trees in the Random Forest model

norm fi_{ij} = the normalized feature importance of i in tree j

4.11.2.5.2 Implementation of Random Forest in Weka

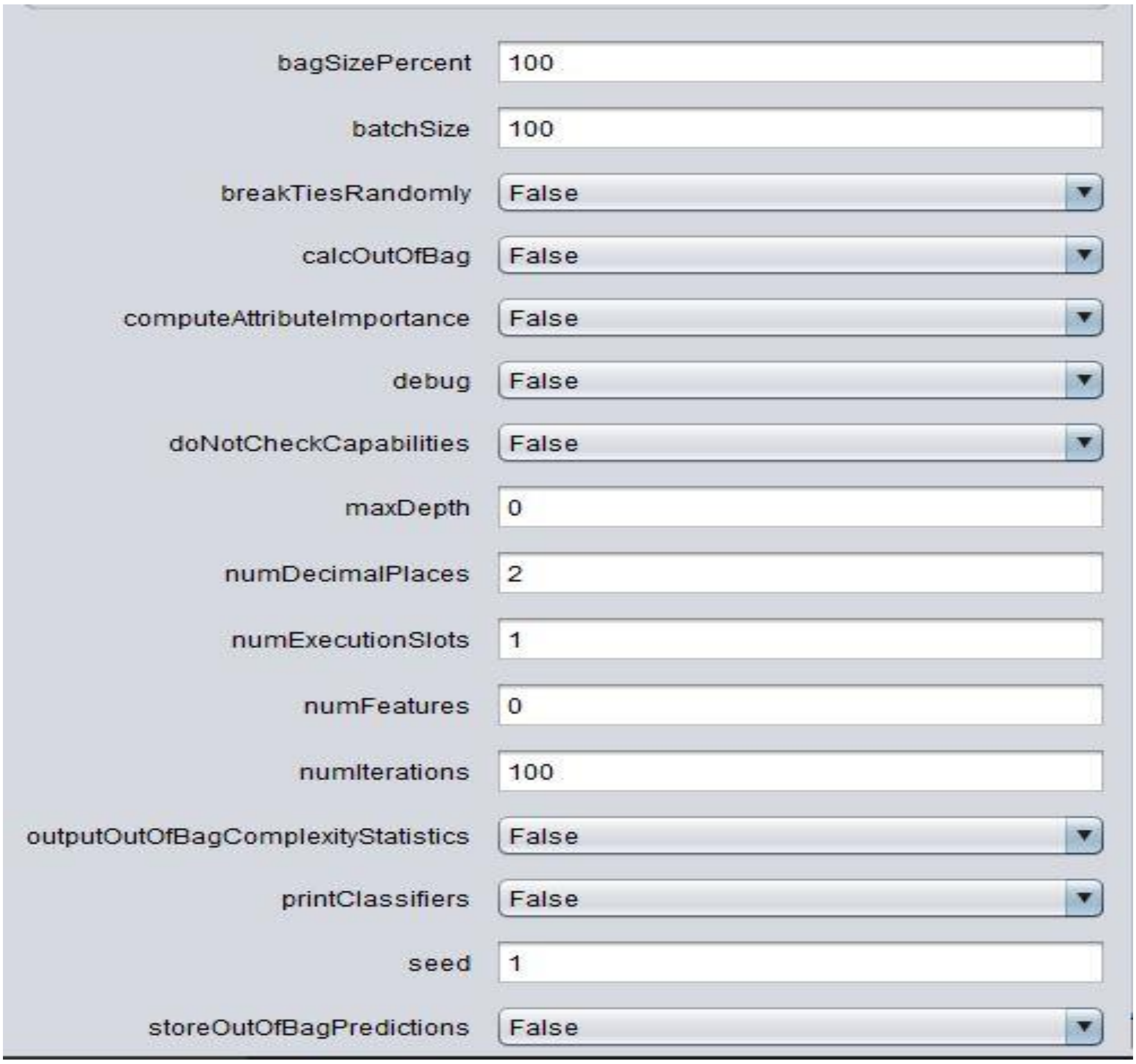
In Weka 3.9.2, first ‘Explorer’ application is selected to insert and classify the data. After inserting the data using ‘Preprocessing’ tab, ‘Classify’ is clicked to perform classification process on the dataset.

In ‘Classify’ window ‘Choose’ option is clicked to select the classifier. In ‘trees’ category, the ‘RandomForest’ classifier is chosen. After this, ‘Percentage Split’ is set to 80% from the ‘Test Option’ which is shown just below the ‘Choose’ menu.

Then by clicking the ‘Start’, the Random Forest classifier is used to classify the EOG dataset.

4.11.2.5.3 Properties of Random Forest

Properties of Random Forest are given below:



bagSizePercent	100
batchSize	100
breakTiesRandomly	False
calcOutOfBag	False
computeAttributeImportance	False
debug	False
doNotCheckCapabilities	False
maxDepth	0
numDecimalPlaces	2
numExecutionSlots	1
numFeatures	0
numIterations	100
outputOutOfBagComplexityStatistics	False
printClassifiers	False
seed	1
storeOutOfBagPredictions	False

Figure 4.11: Properties of Random Forest

In Figure 4.11 the properties of Random Forest are shown. It is seen from the Figure that the batchsize is kept 100 and debug mode is kept ‘False’ for Random Forest classifier.

4.11.2.5.4 Advantages of Random Forest

Several advantages of random forest classifier are listed below [76]:

- For large datasets, it gives a highly accurate output.
- It can process thousands of input variables without variable deletion.

- It can estimate missing data effectively.

- It can balance error in class population unbalanced datasets.

4.11.2.5.5 Disadvantages of Random Forest

Few disadvantages of RF classifier are listed as follows [76]:

- Random Forest has overfit issue for some datasets with noisy classification/regression tasks.

- For categorical variables with different number of levels, Random Forests are biased of those attributes which have more levels. Thus, the importance scores from Random Forest are not reliable for this type of data.

4.11.2.6 Bagging

A bagging is an ensemble learning techniques that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions to form a final prediction [77].

4.11.2.6.1 Mathematical Model of Bagging

The general technique is same as the training algorithm. A training set X = $x_1,\ldots, x_n$ is given with responses Y = $y_1,\ldots, y_n$. Bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples [78].

For b = 1,….., B :

a) Sample, with replacement, n training esamples from X, Y ; call these X_b, Y_b

b) Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on $x^{'}$.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{44}$$

To create better model performance, bagging procedure decreases the variance of the model without increasing the bias.

An estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from the individual regression trees on x$^{'}$.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b\,(x') - \hat{f})^2}{B-1}} \tag{45}$$

The number of samples/trees, B, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number trees B can be found using cross-validation, or by observing the out-of-bag error. The mean prediction error on each training sample x_i, using only the trees that did not have x_i, in their bootstrap sample. The training and test error tend to level off after some number of trees have been fit.

4.11.2.6.2 Implementation of Bagging in Weka

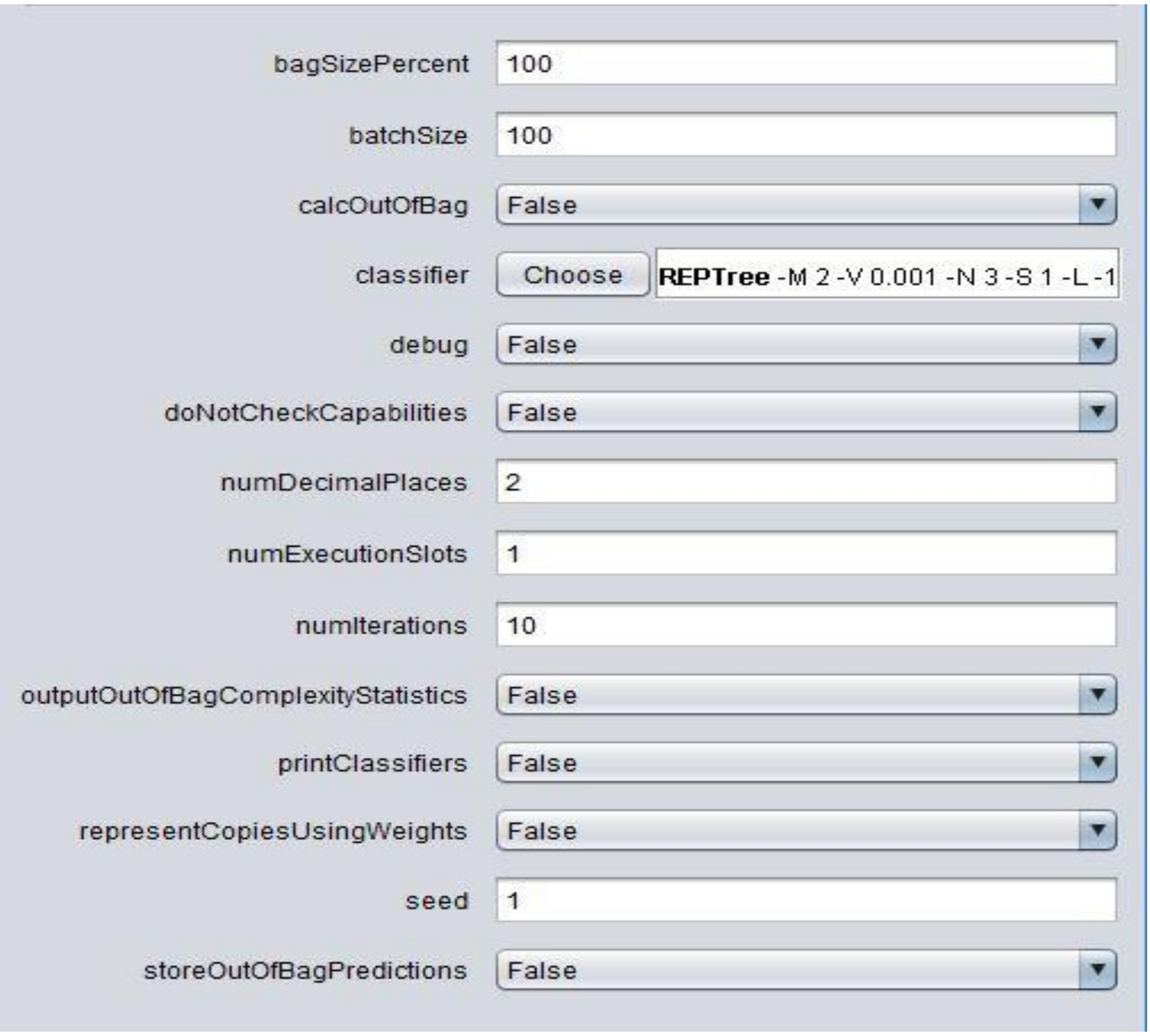
In Weka 3.9.2, first ‘Explorer’ application is selected to insert and classify the data. After inserting the data using ‘Preprocessing’ tab, ‘Classify’ is clicked to perform classification process on the dataset.

In ‘Classify’ window ‘Choose’ option is clicked to select the classifier. In ‘meta’ category, the ‘Bagging’ classifier is chosen. After this, ‘Percentage Split’ is set to 80% from the ‘Test Option’ which is shown just below the ‘Choose’ menu.

Then by clicking the ‘Start’, the Bagging classifier is used to classify the EOG dataset.

4.11.2.6.3 Properties of Bagging

Properties of bagging are given below:



bagSizePercent	100
batchSize	100
calcOutOfBag	False
classifier	Choose REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1
debug	False
doNotCheckCapabilities	False
numDecimalPlaces	2
numExecutionSlots	1
numIterations	10
outputOutOfBagComplexityStatistics	False
printClassifiers	False
representCopiesUsingWeights	False
seed	1
storeOutOfBagPredictions	False

Figure 4.12: Properties of Bagging

The properties of Bagging is shown in Figure 4.12. The batchsize is 100 and debug mode is kept ‘False’. As for the classifier option in bagging properties ‘REPTree’ is selected.

4.11.2.6.4 Advantages of Bagging

The advantages of bagging are given below [79]:

- It improves the accuracy and stability of machine learning algorithms which has been used in statistical classification and regression.
- Variance is reduced by this classifier and it also helps to avoid overfitting.

4.11.2.6.5 Disadvantages of Bagging

Few disadvantages of bagging are given below [79]:

- It can mildly degrade the performance of sTable methods such as k-nearest neighbors.

4.11.3 Underfitting and Overfitting in Machine Learning

In machine learning, there are some issues which occur while using the classifiers. These issues lead to poor predictions on new datasets. These issues are: Underfitting and Overfitting. Detailed description of these two issues are given below:

4.11.3.1 Underfitting

When a satistical model or machine learning algorithm cannot capture the underlying trend of the data it is said to have underfitting. If it occurs that means the model or the algorithm does not fit the data well enough. When there is not enough data to build an accurate model or to build a linear model with a non-linear data this usually occurs [80].

4.11.3.1.1 How to Avoid Underfitting

There are various ways to avoid underfitting issues in machine learning. They are given below [80]:

- Resampling: It is the process of repeatedly drawing samples from a dataset and refitting a given model on each sample with the goal of learning more about the fitted model.

- It can be avoided by using more data.

- By reducing the features using feature selection, this can be avoided.

4.11.3.2 Overfitting

When a statistical model or machine learning algorithm captures the noise of the data, overfitting occurs. In overfitting, the model is trained with a lot of data. After getting trained with so much of data, the model starts learning from the noise and inaccurate data entries in the dataaset. Then the data is not categorized correctly by the model, because of too much details and noise [80].

4.11.3.2.1 How to Avoid Overfitting

The commonly methods of avoiding overfitting are listed below:

- Cross-validation: 5-fold cross validation is used as a standard way to find out of sample prediction error.

- Early Stopping: The guidance cab be provided by it's rules as to how iterations which can be run before learner begins to overfit.

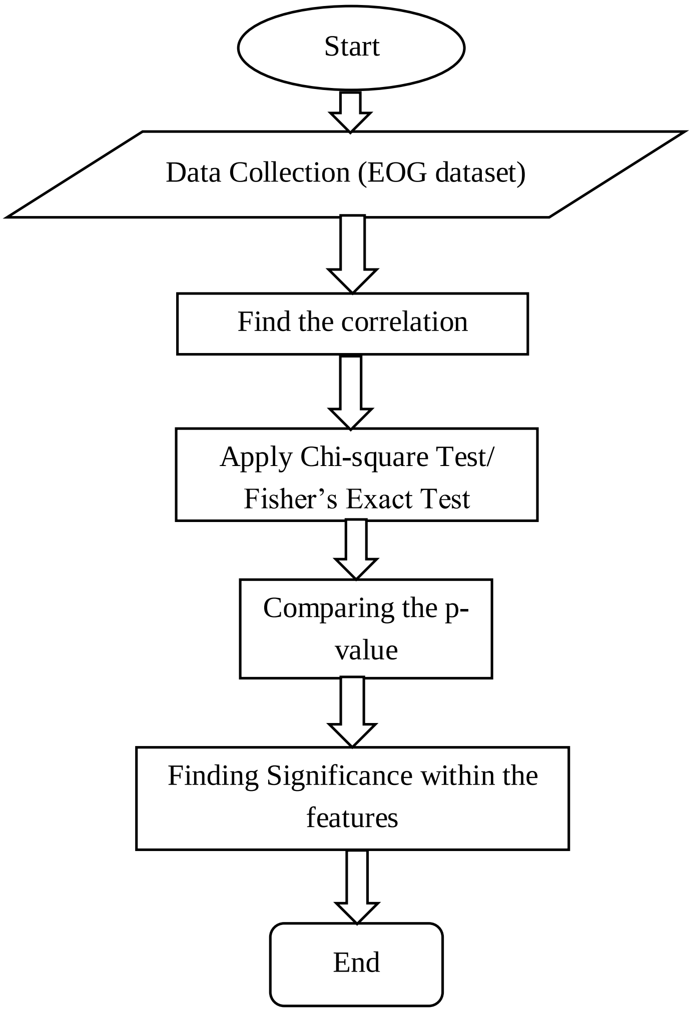
- Pruning: While building related models, pruning is used exclusively. It simply removes the nodes which add little predictive power for the problem in hand.

- Regularization: It gives a cost term for bringing in more features with the objective function. It attempts to push coefficients for many variables to zero and lessens the cost term.

In this experiment, underfitting occurs on the EOG dataset. To avoid this issue, resampling method has been used [80].

4.12 Correlation Method

Correlation in machine learning is the procedure to find the mutual relationship or association between quantities. Same 11 features are selected for all 7 subjects in the experiment. IBM SPSS Statistics 25 software has been used to show the correlation between the class and the features. Steps of correlation procedure are shoowed in the following flowchart in Figure 4.13:



Start
Data Collection (EOG dataset)
Find the correlation
Apply Chi-square Test/ Fisher's Exact Test
Comparing the p-value
Finding Significance within the features
End

Figure 4.13: Flowchart of Correlation Method

Description of correlation method is given below:

4.12.1 Software Used for Correlation

SPSS statistics is a software package which is used for interactive, statistical analysis.It is originally stood for Statistical Package for the Social Sciences (SPSS). There are several statistics included in this software:

- Descriptive Statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics
- Bivariate Statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Non parametric tests, Bayesian
- Prediction for numerical outcomes: Linear regression
- Prediction for identifying groups: Factor Analysis, Cluster Analysis (two-step, k-means, hierarchical)
- Geo spatial analysis, simulation
- R extension (GUI), Python

SPSS statistics can read and write data from ASCII text files, other statistical packages, spreadsheets and databases. It also can read and write to external relational database Tables via ODBC (Open Database Connectivity) and SQL (Structured Query Language) [81].

4.12.2 Ranker

Ranker is a scheme of arranging each feature in order. It ranks each feature sequentially based on their evaluations and it also removes the lower ranked attributes by itself. It is possible to set a custom threshold for deleting the lower ranked features. It is also feasible to specify particular features that need to be there in the list whether it is in lower or upper rank [36].

4.12.2.1 Information Gain Attribute Evaluation

With respect to the class InfoGainAttributeEval measures the information gain to evaluate the worth of an attribute. InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute) [88]. It will help to determine the most correlated features in the dataset. In Weka 3.9.2, under ‘Select Attributes’ option information gain of the each attribute can be calculated using ‘Ranker method’.

4.13 Procedure of Correlation Method

Chi-square test / Fisher's Exact test is performed to find the correlation between the features and the class. First, the dataset is opened in IBM SPSS Statistic 25 software.Then by clicking the 'Analyze' option, 'Descriptive Statistics' is chosen. From that, 'Crosstabs' is clicked to use the correlatiion tests. The necessary options are selected from 'Crosstab' window. The detailed description about chi-square test and Fisher's Exact test are given below:

4.13.1.1 Chi-Square Test

Chi-Square (χ^2) statistic is a test that calculates expectations comparison with actual observed data. That data must be random, raw, mutually exclusive, drawn from independent variables and drawn from a large enough sample. The test determine if there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. It is actually hypothesis test where the sampling distribution of the test statistic is a Chi-squared distribution when the null hypothesis is true [82].

4.13.1.1.1 Mathematical Model of Chi-Square Test

Let us assume that n observations in a random sample from a population are classified into k mutually exclusive classes with respective observed numbers x_i (for i=1,2,….,k) and a null hypothesis gives the probability p_i that an observation falls into the ith class. So, we have the expected numbers $m_i = np_i$ for all i, where

$$\sum_{i=1}^{k} p_i = 1 \tag{46}$$

$$\sum_{i=1}^{k} m_i = n \sum_{i=1}^{k} p_i = \sum_{i=1}^{k} x_i \tag{47}$$

Under the circumstance of the null hypothesis being correct, as $n \to \infty$ the limiting distribution of the quantity given below is χ^2 distribution [83].

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{{x_i}^2}{m_i} - n \tag{48}$$

Pearson considered the case where the expected numbers depended on the parameters that had to be measured from the sample with the notation of m_i being the true expected numbers and m_i' being the calculated expected numbers, the difference is given as:

$$\chi^2 - \chi'^2 = \sum_{i=1}^{k} \frac{{x_i}^2}{m_i} - \sum_{i=1}^{k} \frac{{x_i}^2}{m_i'} \tag{49}$$

4.13.1.2 Fisher’s Exact Test

Fisher’s exact test is a statistical test. It determines if there are nonrandom associations between two categorical variables.

4.13.1.2.1 Mathematical Model of Fisher’s Exact Test

Let us consider two variable X and Y, with m and n observed states. An $m \times n$ matrix is formed in which the entries a_{ij} represent the number of observations where x = i and y = j. The row and columns sums R_i and C_j are calculated. The total sum is given as:

$$N = \sum_i R_i = \sum_j C_j \tag{50}$$

The conditional probability of getting the actual matrix is measured given the particular row and column sums:

$$P_{cutoff} = \frac{(R_1!\,R_2!\ldots\ldots R_n!)\,(C_1!\,C_2!\ldots\ldots C_n!)}{N\,!\,\prod_{ij} a_{ij}!} \tag{51}$$

This is a multivariate generalization of the hypergeometric probability function.

To compute the P-value of the test, the Tables must be ordered by some criterion which measures dependence, and those tabels that represent equal or greater deviation from independence than the observed Table which are the ones whose probabilities are added together. The test is most commonly applied to 2 X 2 matrices. For Tables larger than 2 X 2, the test cannot be performed [84].

4.13.1.3 Measuring P-Value:

IBM SPSS will be used to find out the P-Value. It defines the probability or chance of rejecting a null hypothesis if it is true. SPSS statistical software has packages that can make the calculation of any parameter. It has two hypotheses. One is null hypothesis which can be composite or simple and another one is alternate hypothesis. The significant level will be defined to 5%. The probability of type 1 is the l.o.s and is denoted as,

P (Rejecting the null hypothesis / It is true)

l.o.s. = Assigned the risk of rejecting the null hypothesis (if it is true)

p-value = Observed the risk of rejecting the null hypothesis (if it is true)

Therefore, the decision criterion can be found as,

If p-value < l.o.s., will reject the null hypothesis

If p-value > l.o.s., will not reject the null hypothesis

Thus the P-Value is measured.

CHAPTER V

RESULTS & DISCUSSION

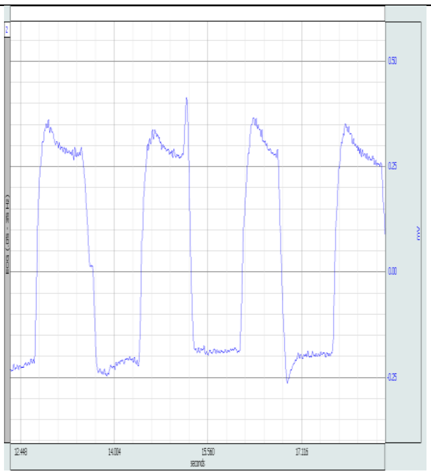
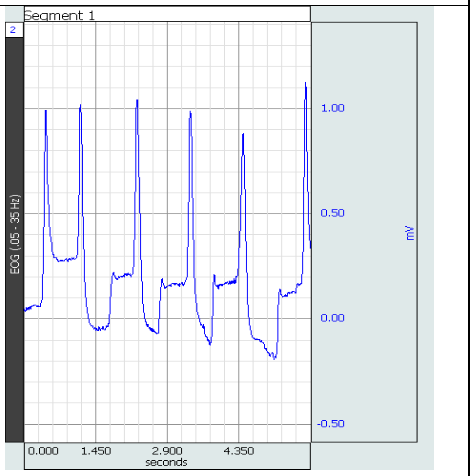
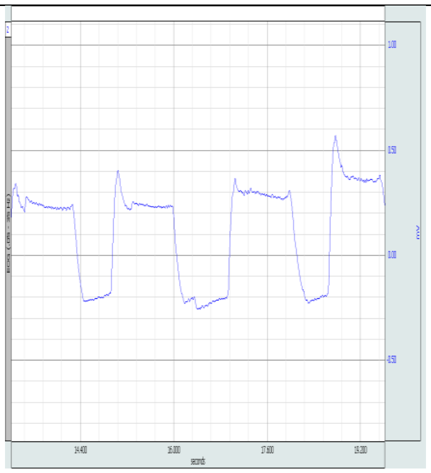
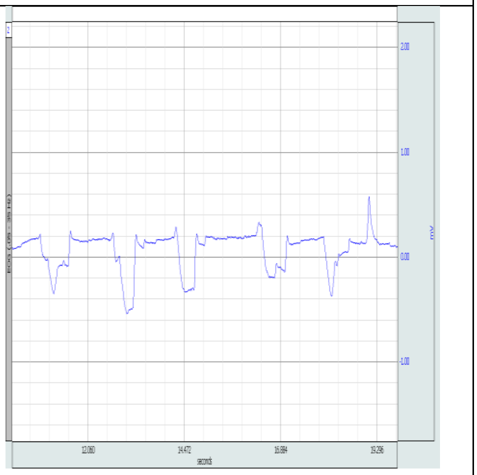
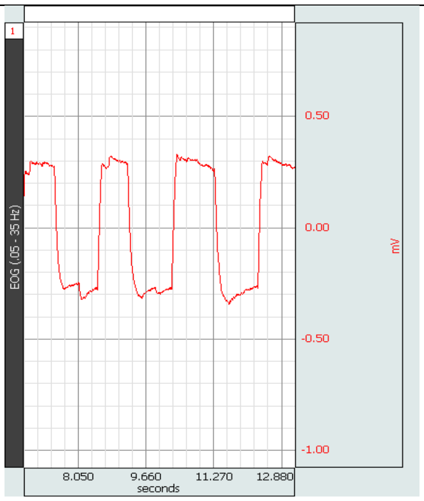
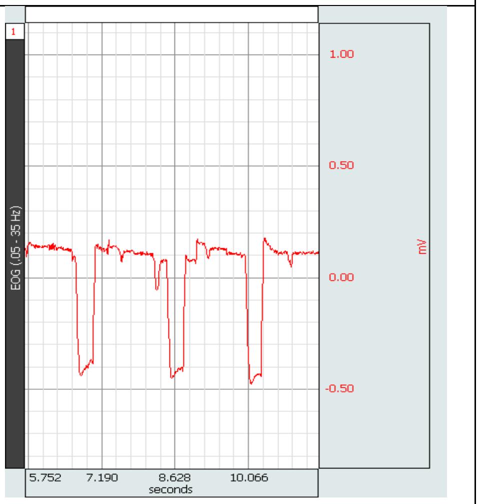
5.1 Introduction

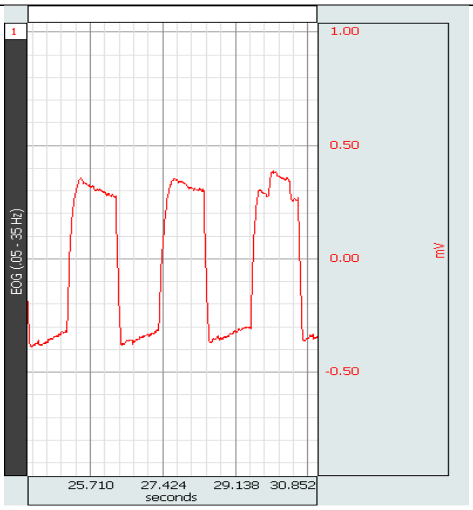
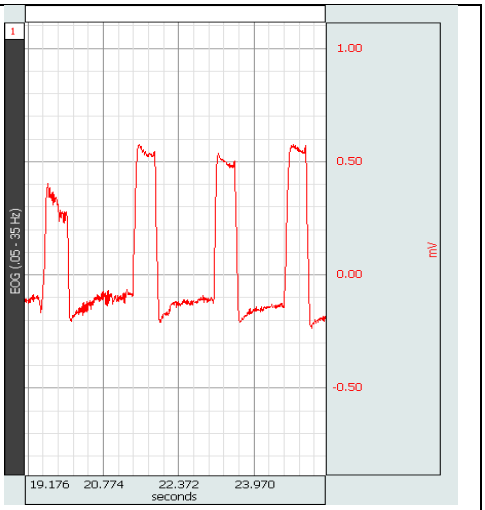
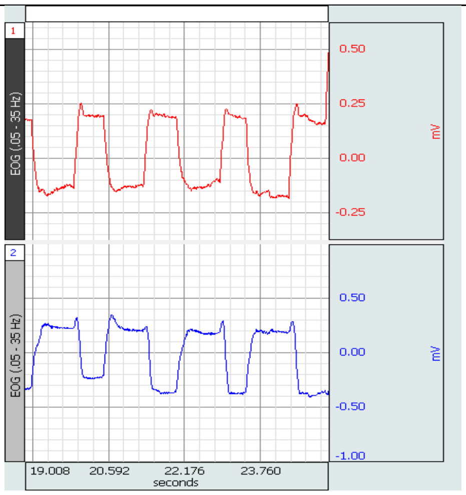
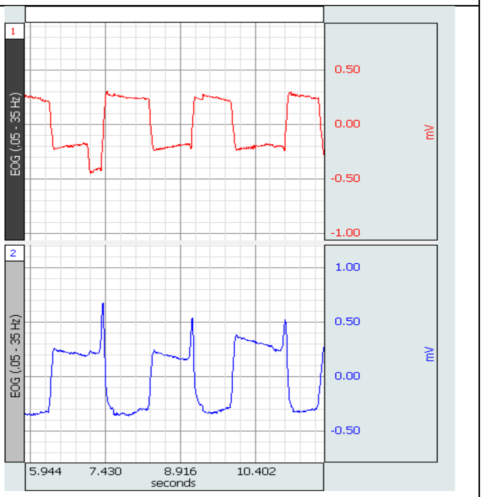
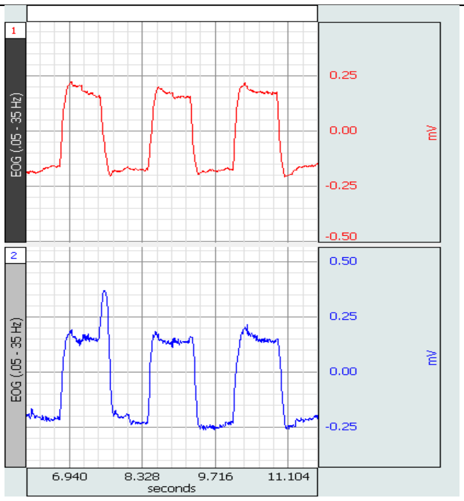
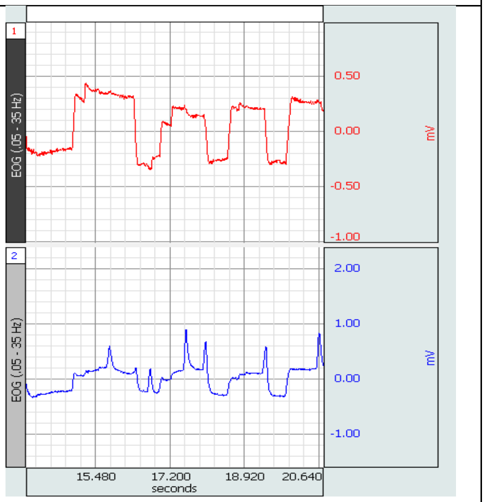
This chapter summarizes the experimental and calculated results of all test perform in this study. Test results of different machine learning classification algorithms applied on the EOG dataset are presented in this chapter. As from the classification result, it is seen that Naïve Bayes has accuracy of 30.7692%, SVM has accuracy of 30.7692%, Logistic Regression has accuracy of 53.8462%, KNN has accuracy of 7.6923%, Random Forest has accuracy of 84.615% and Bagging has accuracy of 92.31%. The comparison of these classifiers is also presented in this chapter. The comparison is done based one the correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error, relative absolute error, TP rate, FP rate, precision, recall, F-measure, MCC, ROC area and PRC area. As from the comparison result, it is seen that bagging has the highest accuracy and KNN has the lowest accuracy among all the classifiers. The correlation result of the features with classes is also showed here. From the correlation result, it is seen that mean of channel 1 and channel 2 are the most significant features among all the features present in the EOG dataset. These two features are directly related to the classes in the dataset.

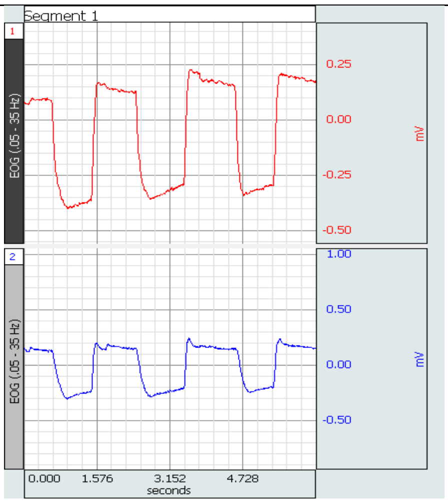
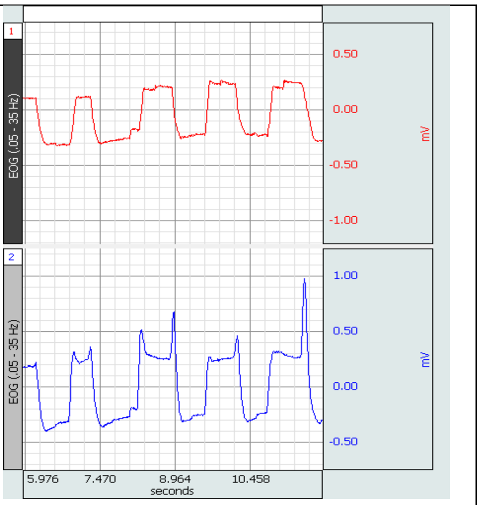
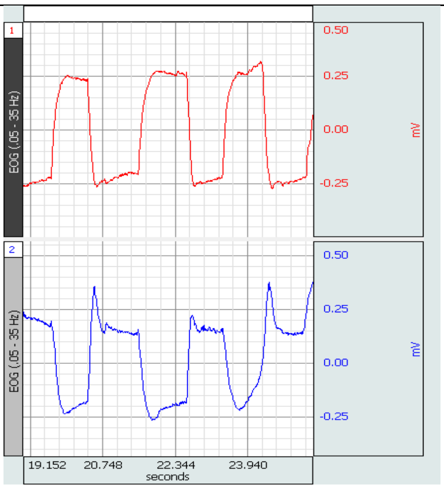
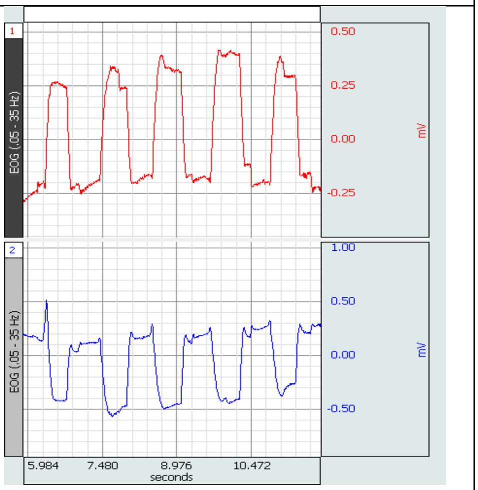
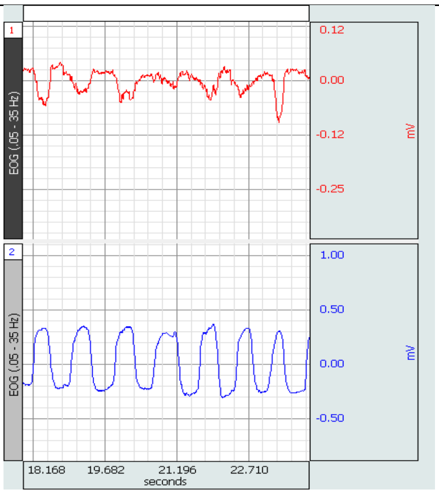
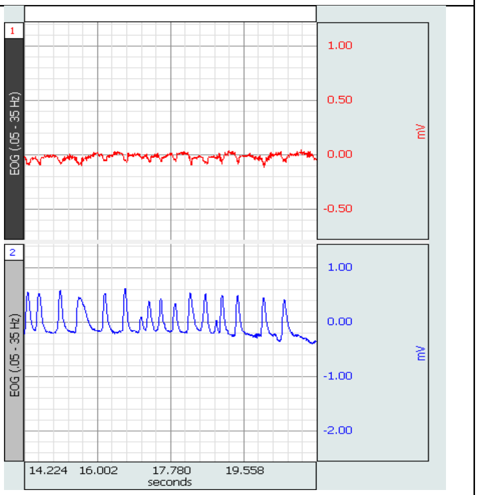
5.2 Typical EOG Signal for Normal and Squint Eye Subject

Eyeball movement signal is acquired using the Biopac MP36 Acquisition Unit. This acquisition unit comes with Biopac Student Lab Software. In that software the EOG signal can be seen in a sinusoidal form. It can be saved for further analysis. In the following table typical EOG data of normal and squint eye subject are shown:

Table 5.1: Typical EOG Signal

Direction	Normal Subject	Squint Eye Subject
Up		
Down		
Left		

Right		
Upleft		
Upright		

Downleft		
Downright		
Blink		

5.3 Subject Information Related to EOG Signal

The subject information such as mean, standard deviation, kurtosis, skewness, variance are related directly to the EOG signal. These infomations of the subject are extracted from the EOG signal using MATLAB library functions. These are the important features which have direct impact on the EOG dataset. Average for all subject informations are shown in the following Table 5.2:

Table 5.2: Average Values of Subject Information

Direction	Mean ± SD Average		Kurtosis Average		Skewness Average		Variance Average	
	EOG1	EOG2	EOG1	EOG2	EOG1	EOG2	EOG1	EOG2
Up	-0.026± 0.060	0.072± 0.261	3.792	3.093	-0.360	0.367	0.004	0.081
Down	0.024± 0.054	0.060± 0.229	6.529	4.022	-0.131	-0.076	0.004	0.062
Left	-0.088± 0.283	-0.073± 0.142	1.635	6.152	-0.160	0.804	0.085	0.037
Right	0.033± 0.298	0.002± 0.107	1.479	5.855	0.134	0.860	0.095	0.016
Upleft	-0.013± 0.234	0.007± 0.267	1.611	1.846	0.103	-0.193	0.059	0.083
Upright	-0.026± 0.275	0.011± 0.283	1.424	2.638	-0.068	-0.134	0.083	0.123
Downleft	-0.005± 0.242	0.014± 0.233	1.637	2.120	-0.216	-0.328	0.062	0.060
Downright	0.010± 0.271	-0.015± 0.200	1.357	1.686	0.159	-0.184	0.077	0.044
Blink	0.016± 0.138	-0.019± 0.351	7.879	3.133	0.860	0.267	0.025	0.133

5.4 Classification Results

Different classification algorithms are applied on the EOG data taken from 7 subjects aged between 22 to 48 years. Out of 7 subjects 4 are male and 3 are female. In the dataset there are total 16 instances which are classified with the help of the classificaton algorithms. There are two types of dataset used for the classification. They are: raw dataset and dataset prepared after performing feature extraction. Time taken in Weka 3.9.2 for both the dataset are given as follows:

Time taken for Raw Dataset: 10-30 minutes

Time taken for Dataset after feature extraction: 1-5 minutes

The result of naïve bayes, support vector machine, logistic regression, k-nearest neighbor, random forest and bagging are given below:

5.4.1 Naïve Bayes Classifier Result

The result of naïve bayes classifier is given in the following Table 5.3:

Table 5.3: Detailed Accuracy by Naïve Bayes Classifier

Parameters	Using Raw Data	After Feature Extraction
Correctly classified instances	61.54%	30.7692%
Incorrectly classified instances	38.46%	69.2308%
Kappa Statistic	0.5578	0.22
Mean Absolute Error	0.0994	0.1624
Root mean squared Error	0.2834	0.3818
Relative absolute Error	49.74%	81.2109%
Root relative squared error	89.01%	119.8783%
TP Rate	0.615	0.308
FP Rate	0.04	0.08
Recall	0.615	0.308
ROC Area	0.799	0.799

From Table 5.3, it is seen that naïve bayes algorithm has correctly classified 30.7692% instances. So it can be determined that the accuracy of naïve bayes classifier is 30.7692%.

5.4.2 Support Vector Machine (SVM) Classifier Result

SVM classifier result is given as follows:

Table 5.4: Detailed Accuracy by SVM Classifier

Parameters	Using Raw Data	After Feature Extraction
Correctly classified instances	53.85%	30.7692%
Incorrectly classified instances	46.15%	69.2308%
Kappa Statistic	0.4658	0.2041
Mean Absolute Error	0.1819	0.1828
Root mean squared Error	0.2962	0.2978
Relative absolute Error	90.97%	91.4452%
Root relative squared error	93.02%	93.5244%
TP Rate	0.538	0.308
FP Rate	0.07	0.116
Recall	0.538	0.308
ROC Area	0.677	0.753
PRC Area	0.529	0.520

From Table 5.4 it is determined that the accuracy of support vector machine algorithm is 30.7692%. And it is incorrectly classified 69.2308% instances.

5.4.3 Logistic Regression Classifier Result

The result of logistic regression is given below:

Table 5.5: Detailed Accuracy by Logistic Regression Classifier

Parameters	Using Raw Data	After Feature Extraction
Correctly classified instances	38.46%	53.8462%
Incorrectly classified instances	61.54%	46.1538%
Kappa Statistic	0.3113	0.473
Mean Absolute Error	0.1679	0.1123
Root mean squared Error	0.2943	0.2746
Relative absolute Error	84.00%	56.1495%
Root relative squared error	92.43%	86.2439%
TP Rate	0.358	0.538
FP Rate	0.06	0.057
Recall	0.358	0.538
ROC Area	0.591	0.838
PRC Area	0.458	0.705

From Table 5.5 it can be seen that the accuracy of logistic regression classifier is 53.8462%. And this classifier is incorrectly classified 46.1538% instances.

5.4.4 K-Nearest Neighbor Classifier Result

The result of KNN classifier is given below:

Table 5.6: Detailed Accuracy by KNN Classifier

Parameters	Using Raw Data	After Feature Extraction
Correctly classified instances	15.38%	7.6923%
Incorrectly classified instances	84.62%	92.3077%
Kappa Statistic	0.0272	-0.047
Mean Absolute Error	0.2008	0.2013
Root mean squared Error	0.3225	0.3245
Relative absolute Error	100.43%	100.6713%
Root relative squared error	101.27%	101.9141%
TP Rate	0.154	0.077
FP Rate	0.126	0.125
Recall	0.154	0.077
ROC Area	0.47	0.415
PRC Area	0.16	0.222

In Table 5.4, it is seen that the correctly classified instances of k-nearest neighbor is 7.6923% which is the accuracy of that classifier. And the incorrectly classified instances for SVM is 92.3077%.

5.4.5 Random Forest Classifier Result

Random Forest classifier result o the EOG dataset is given as below:

Table 5.7: Detailed Accuracy by Random Forest Classifier

Parameters	Using Raw Data	After Feature Extraction
Correctly classified instances	84.62%	84.62%
Incorrectly classified instances	15.38%	15.38%
Kappa Statistic	0.8169	0.8207
Mean Absolute Error	0.1411	0.1322
Root mean squared Error	0.2438	0.2315
Relative absolute Error	70.58%	66.1228%
Root relative squared error	76.57%	72.7092%
TP Rate	0.846	0.846
FP Rate	0.037	0.014
Precision	0.846	0.949
Recall	0.846	0.846
F-measure	0.846	0.872
MCC	0.809	0.865
ROC Area	0.924	0.948
PRC Area	0.882	0.913

From Table 5.7 it can be determined that the accuracy of random forest classifier is 84.62%. And the random forest classifer is incorrectly classified 15.38% instances.

5.4.6 Bagging Classifier Result

The of Bagging classification algorithm is given as follows:

Table 5.8: Detailed Accuracy by Bagging Classifier

Parameters	Using Raw Data	After Feature Extraction
Correctly classified instances	92.31%	92.31%
Incorrectly classified instances	7.69%	7.69%
Kappa Statistic	0.9091	0.9091
Mean Absolute Error	0.0701	0.0981
Root mean squared Error	0.1525	0.1831
Relative absolute Error	35.09%	49.0556%
Root relative squared error	47.88%	57.5001%
TP Rate	0.923	0.923
FP Rate	0.014	0.014
Precision	0.949	0.949
Recall	0.923	0.923
F-measure	0.923	0.923
MCC	0.915	0.915
ROC Area	0.943	0.988
PRC Area	0.89	0.969

The accuracy of bagging classifier is 92.31% which is displayed in Table 5.8. From this Table it is also seen that the incorrectly classified instances of bagging classifier is 7.69%.

5.5 Comparison of Different Classification Algorithms

In this research, six different machine learning algorithms are used on the EOG dataset to classify the eyeball movement directions. The comparison among these classification algorithms is performed to show the best algorithm for EOG data. The comparison is done based on correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error, root mean squared error, relative absolute error, root relative squared error. Here the comparison among the algorithms is mainly focused on the correctly classified instances. This parameter gives information about the accuracy of the algorithms. This accuracy is showed in percentage. The higher the percentage , the accuracy of the algorithm is higher which proves the effeectiveness of that classifier on the dataset. The following Table is shown the comparison among the algorithms:

Table 5.9: Comparison Among Different Classifiers

Parameters	Naïve Bayes Classifier	Support Vector Machine Classifier	Logistic Regression Classifier	K-nearest Neighbor Classifier	Random Forest Classifier	Bagging Classifier
Correctly classified instances (Accuracy)	30.7692%	30.769%	53.8462%	7.6923%	84.62%	92.31%
Incorrectly classified instances	69.2308%	69.23%	46.1538%	92.31%	15.38%	7.69%
Kappa Statistic	0.22	0.2041	0.473	-0.047	0.8207	0.9091
MAE	0.1624	0.1828	0.1123	0.2013	0.1322	0.0981
RMSE	0.3818	0.2978	0.2746	0.3245	0.2315	0.1831
RSE	81.21%	91.445%	56.15%	100.67%	66.122%	49.055%
RRSE	119.878%	93.524%	86.2439%	101.91%	72.71%	57.5%
TP Rate	0.308	0.308	0.538	0.077	0.846	0.923
FP Rate	0.080	0.116	0.057	0.125	0.014	0.014
Recall	0.308	0.308	0.538	0.077	0.846	0.923
ROC Area	0.778	0.753	0.838	0.415	0.948	0.988
PRC Area	0.569	0.520	0.705	0.222	0.913	0.969

From the Table 5.9, it is seen that the best algorithm for EOG dataset is ‘Bagging’. It has accuracy of 92.31% which is better than all other algorithms applied on the EOG dataset. As it is observed from this Table, when the accuracy increases the kappa statistic also increases whereas the mean absolute error and the relative absolute error decrease. ‘Bagging’ classifier has kappa statistic value of 0.9091 which is more than all other algorithms’ kappa statistic value. And for the mean absolute error and the relative absolute error, ‘Bagging’ has the lowest value of 0.0701 and 35.0884% respectively.

5.6 Comparison with Other Researches:

In the literature review, some researches related to EOG signal classification using machine learning algorithms have been discussed. In the following Table, a comparison is shown between the proposed method and the other researches which are mentioned in the literature review section.

Table 5.10: Comparison between the proposed method and other researches

Reference No.	Authors Name	Year of Publication	Number of Features	Number of Directions for Eyeball Movement	Used Algorithms	Average Weighted Accuracy
51	L. Qin et al	2018	10	4	SVM and ANN	SVM- 66.5% ANN- 69.75%
53	L. Deng et al	2009	Not Defined	4	Fuzzy Logic	Fuzzy Logic- 90%
55	D. Ramkumar et al	2016	16	11	TDNN and FFNN	TDNN- 91.48% FFNN- 94.18%
56	P. Syal et al	2017	16	2	KNN, SVM and DT	KNN- 99.87% SVM- 99.2% DT- 95.4%
The Proposed Method			16	9	Naïve Bayes, SVM, LR, KNN, RF and Bagging	Naïve Bayes- 61.54% SVM- 53.85% LR- 38.46% KNN- 15.38% RF- 84.62% Bagging- 92.31%

From Table 5.10, it is seen that in the proposed method more algorithms have been applied for EOG data. Even though there is a compromise in the accuracy, this method has helped to identify the suiTable algorithms for EOG data.

5.7 Features List

In this research, 16 features are selected for the subjects to prepare the EOG dataset. These features are selected based on eyesight and physical conditions of the subject. The selected features are: mean, standard deviation, kurtosis, skewness, variance of channel 1 (horizontal channel) data and channel 2 (vertical channel) data, age, gender, height, weight, blood group, eyesight problem, color blind, sleep duration, left/right handed, blood pressure, heart disease. All the features are shown in Table 5.11.

Table 5.11: Selected Features List

Features		Subcategory	Data Description
Mean	Channel 1	Minimum: -0.105	Information Gain: 2.7090647189604646
		Maximum: 0.037	
	Channel 2	Minimum: -0.232	Information Gain: 2.7624608609968297
		Maximum: 0.082	
Age		Minimum: 23	Information Gain: 0
		Maximum: 48	
Gender		Male: 36	57.15%
		Female: 27	42.85%
			Information Gain: -0.000000000000001332
Height (cm)		Minimum: 157	Information Gain: 0
		Maximum: 175	
Weight (kg)		Minimum: 40	Information Gain: 0
		Maximum: 70	
Blood Group		O+: 45	71.44%
		A+: 9	14.28%
		B+: 9	14.28%
			Information Gain: 0
Eyesight Problem		Yes: 18	28.57%
		No: 45	71.43%
			Information Gain: 0

Color Blind		Yes: 0	0%
		No: 63	100%
			Information Gain: 0
Sleep Duration (Hrs)		Minimum: 7	Information Gain: 0
		Maximum: 8	
Left/Right Handed		Left Handed: 9	14.285%
		Right Handed: 54	85.715%
			Information Gain: -0.000000000000000888
Blood Pressure		Yes: 18	28.57%
		No: 45	71.43%
			Information Gain: 0.000000000000000444
Heart Disease		Yes: 0	0%
		No: 63	100%
			Information Gain: 0
Kurtosis	Channel 1	Maximum:1.128	Information Gain: 0.6558374337286752
		Minimum:14.605	
	Channel 2	Maximum:1.243	Information Gain: 0
		Minimum:15.458	
Skewness	Channel 1	Maximum:-3.123	Information Gain: 0
		Minimum:2.388	
	Channel 2	Maximum:-2.243	Information Gain: 0
		Minimum:3.131	
Variance	Channel 1	Maximum:0.001	Information Gain: 0.7399545301668979
		Minimum:0182	
	Channel 2	Maximum:0.002	Information Gain: 0
		Minimum:0.586	
Standard Deviation	Channel 1	Maximum:0.025	Information Gain: 0.7399545301668979
		Minimum:0.427	
	Channel 2	Maximum:0.044	Information Gain: 0
		Minimum:0.765	

In Table 5.9, mean, standard deviation, percentage, information gain, variance, skewness and kurtosis values of all the featurs are presented. As from information gain and other parameters, the features can be ranked as follows:

1. Channel 2 mean
2. Channel 1 mean
3. Channel 1 Standard deviation

4. Channel 1 Variance
5. Channel 1 Kurtosis
6. Blood pressure
7. Blood Group
8. Weight (kg)
9. Color blind
10. Height (cm)
11. Age
12. Eyesight problem
13. Channel 2 Standard Deviation
14. Sleep duration (Hrs)
15. Channel 2 Variance
16. Heart Disease
17. Channel 1 Skewness
18. Channel 2 Kurtosis
19. Channel 2 Skewness
20. Left/right handed and
21. Gender.

From this Table, it can be said that channel 2, channel 1 and blood pressure are more related to the direction (class) whereas left/right handed and gender features are not significant compared to other features.

5.8 Correlation Results

Correlation results are showed based on Chi-square test / Fisher's Exact test performed on the processed EOG dataset. Correlation between the features and the classes (directions) are done based on these two tests.. The result of correlation is given as follows:

Table 5.12: Correlation result between the class and the features

Serial No.	Features	Different Measures	Asymptotic Significance (P-value)
1.	**Channel 1 Mean**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
2.	**Channel 2 Mean**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
3.	Age	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
4.	Gender	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
5.	Height (cm)	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
6.	Weight (kg)	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
7.	Blood group	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
8.	Eyesight problem	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
9.	Colorblind	Chi-Square Test	Cannot be computed
		Directional Measure	Cannot be computed
		Symmetric Measure	Cannot be computed

Serial No.	Features	Different Measures	(P-value)
10.	Sleep Duration (Hrs)	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
11.	Left/Right Handed	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
12.	Blood Pressure	Chi-Square Test	1.00
		Directional Measure	Cannot be computed
		Symmetric Measure	1.00
13.	Heart Disease	Chi-Square Test	Cannot be computed
		Directional Measure	Cannot be computed
		Symmetric Measure	Cannot be computed
14.	**Channel 1 Kurtosis**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
15.	**Channel 2 Kurtosis**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
16.	**Channel 1 Skewness**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
17.	**Channel 2 Skewness**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
18.	**Channel 1 Variance**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
19.	**Channel 2 Variance**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
20.	**Channel 1 Std. Deviation**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392
21.	**Channel 2 Std. Deviation**	Chi-Square Test	.392
		Directional Measure	**.002**
		Symmetric Measure	.392

As from the Table 5.12, it is seen that the P-value of Channel 1 and Channel 2 mean, Channel 1 and Channel 2 kurtosis, Channel 1 and Channel 2 skewness, Channel 1 and Channel 2 variance, Channel 1 and Channel 2 standard deviation are 0.002. For the medical data when the P-value of a feature is less is than 0.05, it is said to be significant for the class. In this EOG dataset, mean, standard deviation, kurtosis, skewness and variance of channel 1 and channel 2 are the significant features for the class (direction) whose asymtotic value are below 0.05. From this Table it is observed that directional measure value of some features such as age, gender, height, weight, bloodgroup, eyesightproblem, left/right handed, blood pressure and heart disease could not be computed as the asymptotic standard error of those features are 0. For these features, the chi-square test and symmetric measure value are 1. And for colorblind and heart disease, the chi-square test and symmetric measure value also could not be computed as those two features are constant. By cross checking the significant features in IBM SPSS Statistics 25 Software with Weka 3.9.2 based on information gain it is seen that the following features are the most significant:

1. Channel 2 Mean

2. Channel 1 Mean

3. Channel 1 Standard Deviation

4. Channel 1 Variance

5. Channel 1 Kurtosis

6. Blood Pressure

All other features are not that significant to the EOG dataset.

5.9 Graphical Representation

Graphical representation is very important part of a research. To understand the result of the research at a glance this grraphical represemtation helps a lot. In this experiment, mainly the graphical representations of the correlation result are shown. The graphical representations of the class versus the features are displayed as follows:

5.9.1 Class Vs Channel 1 Mean

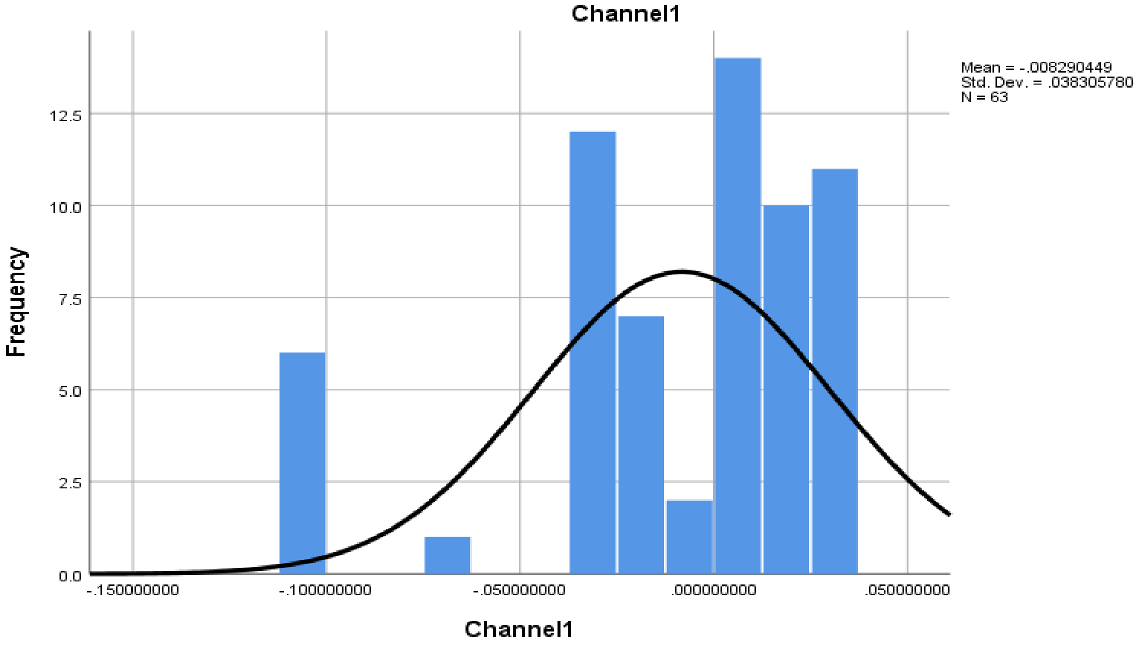


Figure 5.1: Histogram of Class Vs Channel 1 Mean

From Figure 5.1, the mean is calculated as 0.008290449 and the standard deviation is 0.038305780. As from the normal distribution curve, it is seen that most of the values are placed on the right side of the histogram.

5.9.2 Class Vs Channel 2

Class vs Channel 2 histogram is given below:

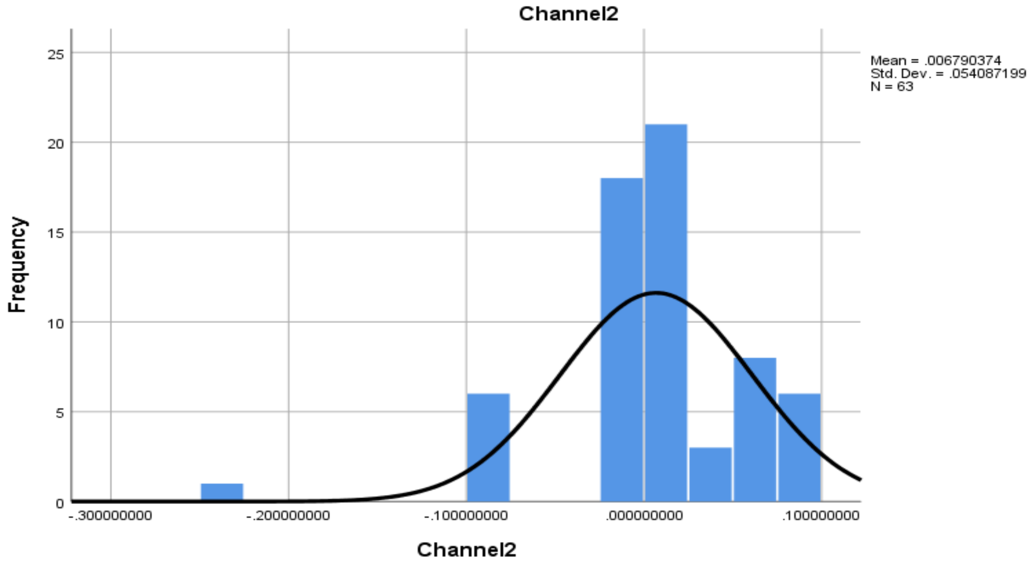


Figure 5.2: Histogram of Class Vs Channel 2 Mean

It is seen from the histogram in Figure 5.2 that the mean is 0.006790374 and the standard deviation is 0.054087189. The normal distribution curve shows that the most values are on the right side of the graph.

5.9.3 Class Vs Age

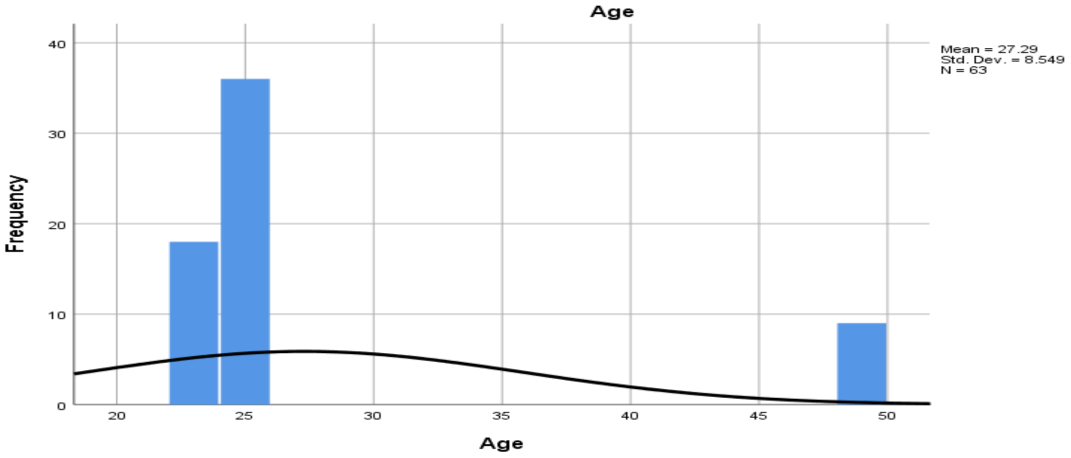


Figure 5.3: Histogram of Class Vs Age

From the histogram shown is Figure 5.3, the values are more frequent in the left side of the graph.

5.9.4 Class Vs Gender

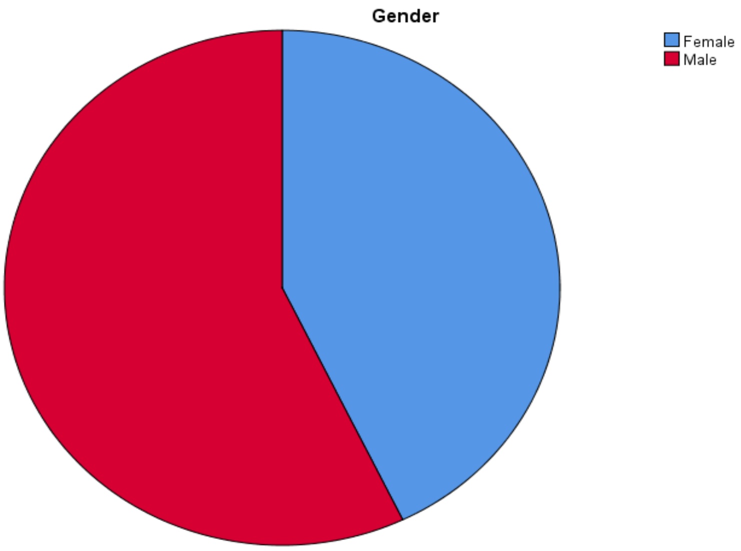


Figure 5.4: Pie chart of Class Vs Gender

From the pie chart shown in Figure 5.4, the male percentage is more in gender than the female for the class.

5.9.5 Class Vs Height (cm)

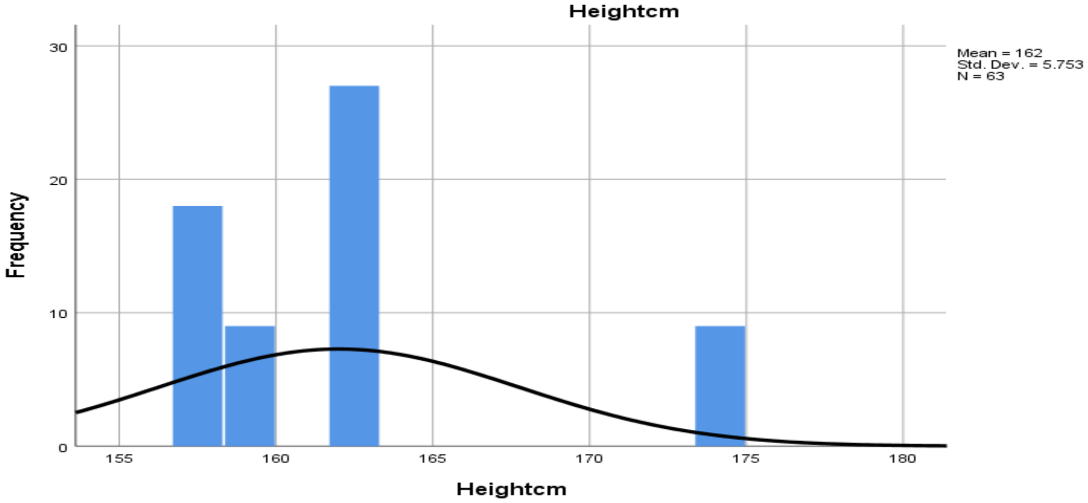


Figure 5.5: Histogram of Class Vs Height (cm)

In Figure 5.5, the histogram of Class Vs Heightcm is shown. The most of values are present on the left side of the graph which is based on the normal distribution curve.

5.9.6 Class Vs Weight (kg)

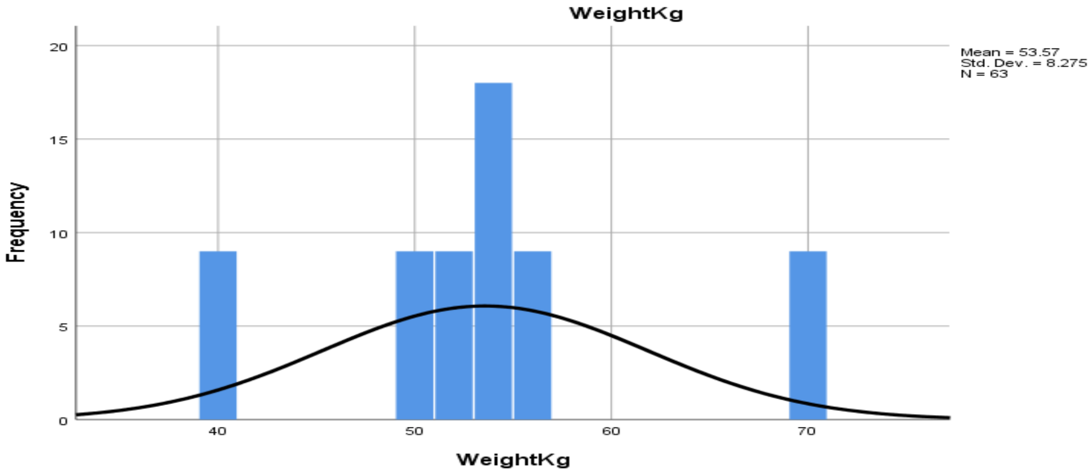


Figure 5.6: Histogram of Class Vs Weight (kg)

In Figure 5.6, the histogram of class vs weightkg is shown. From the Figure it is seen that the most of the values are present in the center of the graph as per the normal distribution curve. The range is between 50 to 60kg. The mean is calculated as 53.57 and the standard deviation is 8.275.

5.9.7 Class Vs Blood Group

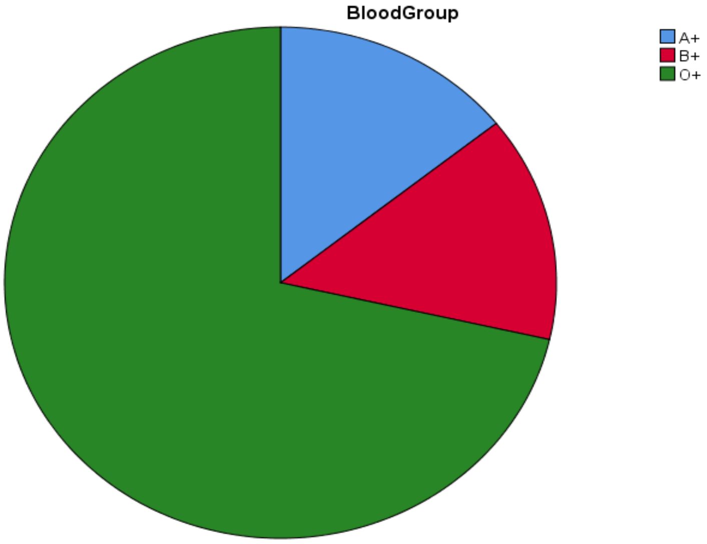


Figure 5.7: Pie chart of Class Vs Blood Group

From the pie chart of class vs bloodgroup shown in Figure 5.7, it is seen that O+ blood group has occupied more than 70% of the pie chart and the blood group of A+ or B+ has occupied the equal percentage in the pie chart.

5.9.8 Class Vs Eyesight problem

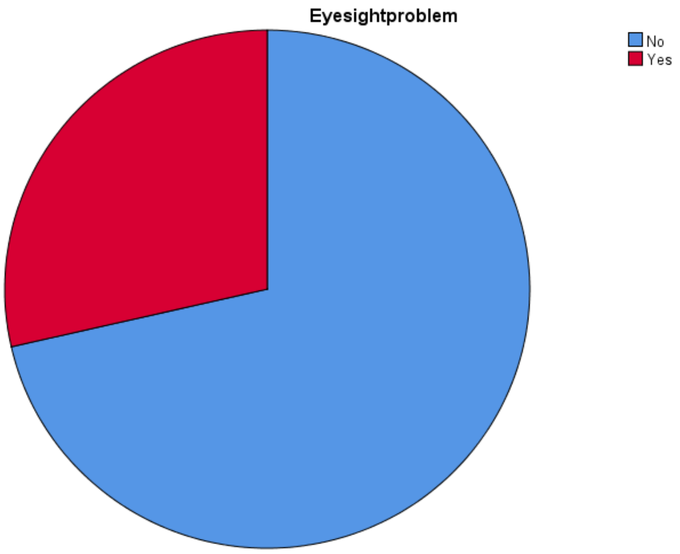


Figure 5.8: Pie chart of Class Vs Eyesight problem

In Figure 5.8, the pie chart of class vs eyesightproblem is shown. It is seen from the Figure that more than 70% of the subjects do not have eyesight problem whereas less than 30% have eyesight problem.

5.9.9 Class Vs Colorblind

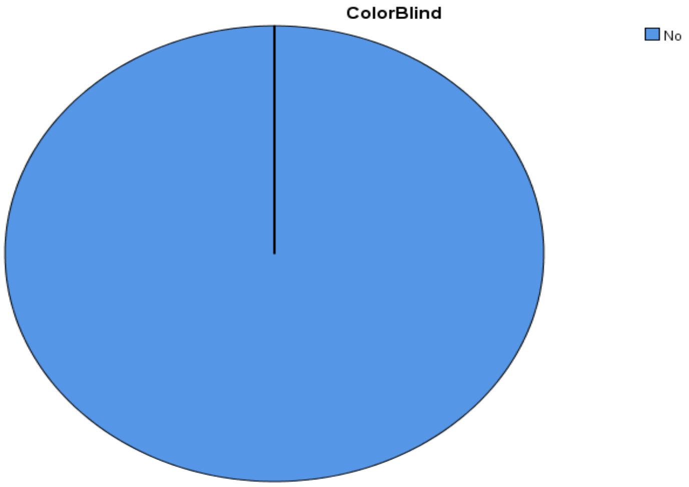


Figure 5.9: Pie chart of Class Vs Colorblind

From the pie chart of class vs colorblind shown in Figure 5.9, it is seen that in this experiment there are no colorblind subject.

5.9.10 Class Vs Sleep Duration (Hrs)

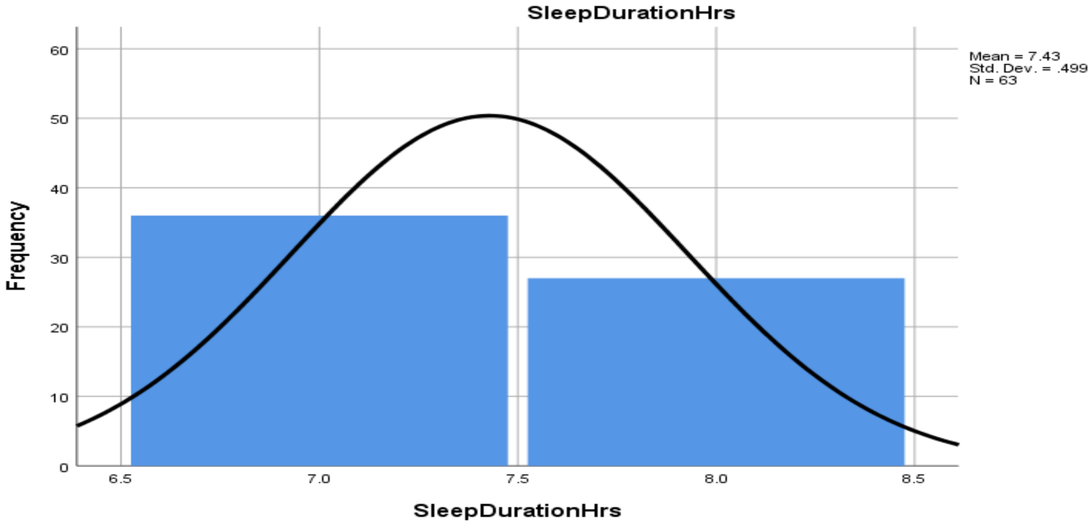


Figure 5.10: Histogram of Class Vs Sleep Duration (Hrs)

In Figure 5.10, the histogram of class vs sleepdurationhrs is shown. From the Figure, it is seen that the most of the values are present at the left side of the graph as per the normal distribution curve.

5.9.11 Class Vs Left/Right Handed

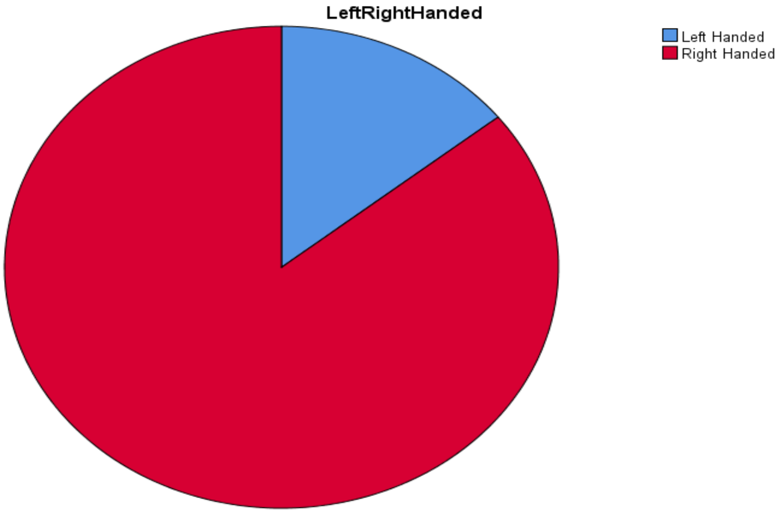


Figure 5.11: Pie chart of Class Vs Left/Right Handed

In Figure 5.11, the pie chart of class vs leftrighthanded is shown. From the pie chart, it is seen that 85% subjects are right handed and 15% subject is left handed.

5.9.12 Class Vs Blood Pressure

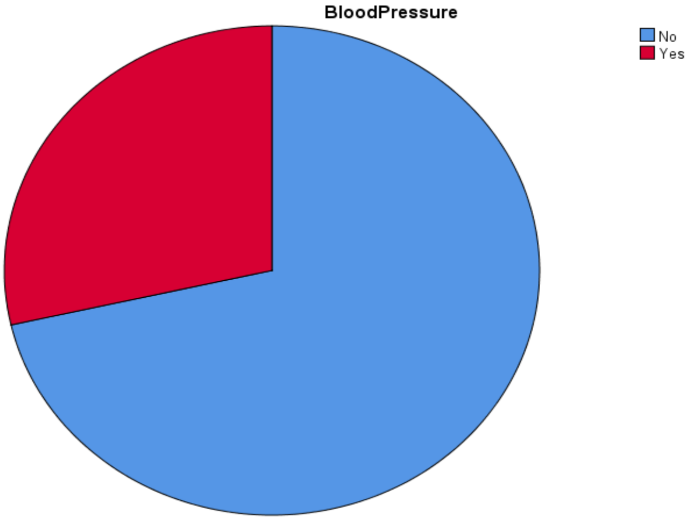


Figure 5.12: Pie chart of Class Vs Blood Pressure

The pie chart of class vs bloodpressure shown in Figure 5.12, more than 70% of the subjects do not have blood pressure but less than 30% subjects have blood pressure.

5.9.13 Direction Vs Heart Disease

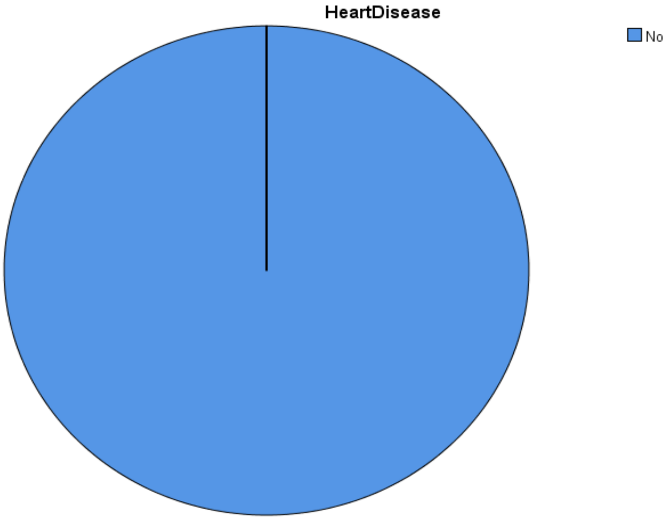


Figure 5.13: Pie chart of Class Vs Heart Disease

From the Figure 5.13, the pie chart of class vs heartdisease is shown. It is seen from the pie chart that there is no subject with heart disease for each class.

5.9.14 Class Vs Channel 1 Kurtosis

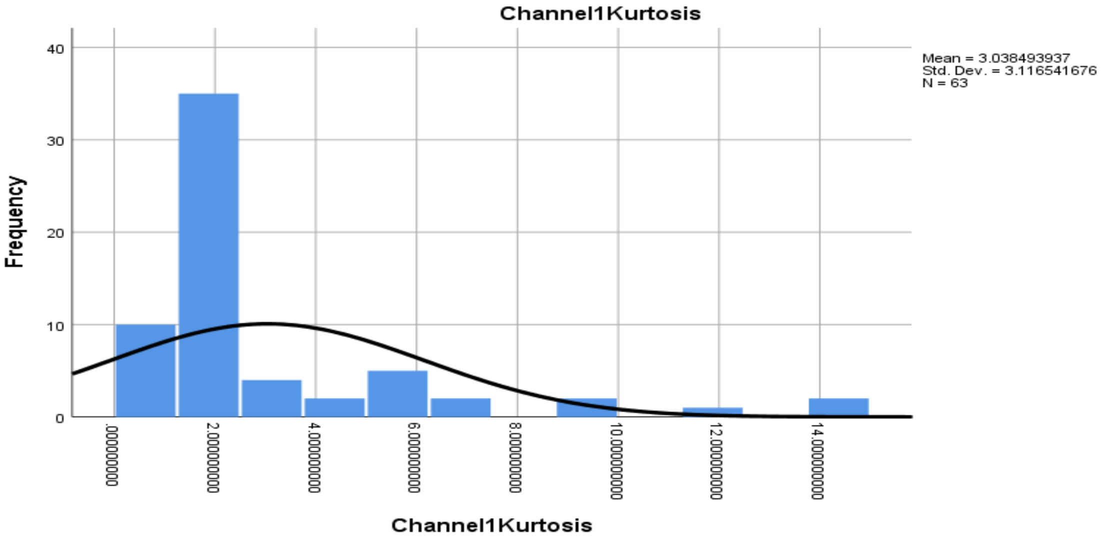


Figure 5.14: Histogram of Class Vs Channel 1 Kurtosis

From the histogram of Figure 5.14 it is seen that most of the values are left side of the value 8. It can be said from the normal curve that channel 1 kurtosis values are more frequent before the value 8.

5.9.15 Class Vs Channel 2 Kurtosis

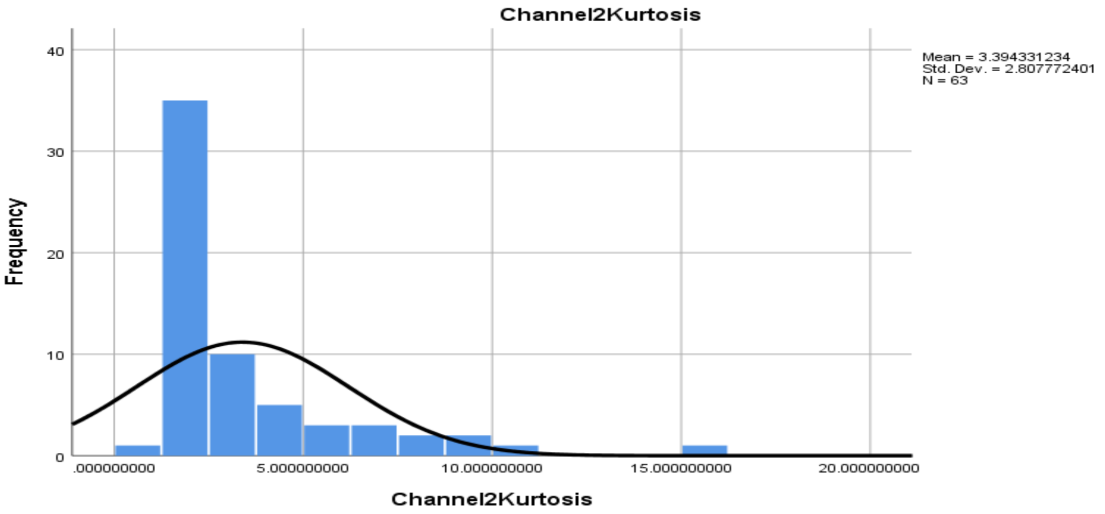


Figure 5.15: Histogram of Class Vs Channel 2 Kurtosis

From the normal curve in Figure 5.15 it is seen that the values are more to the left side of the graph. The values are more frequent before channel 2 kurtosis value reaches 10.

5.9.16 Class Vs Channel 1 Skewness

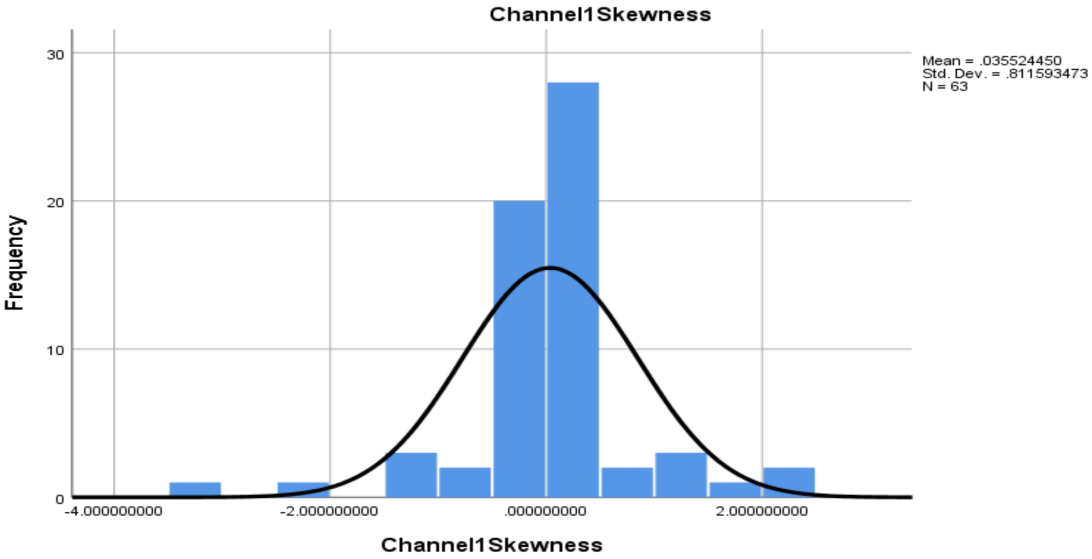


Figure 5.16: Histogram of Class Vs Channel 1 Skewness

In Figure 5.16, the histogram of class vs channel 1 skewness is shown. It is seen that the values are more frequent slightly at the right side of the graph where the channel 1 skewness value is .0000 .

5.9.17 Class Vs Channel 2 Skewness

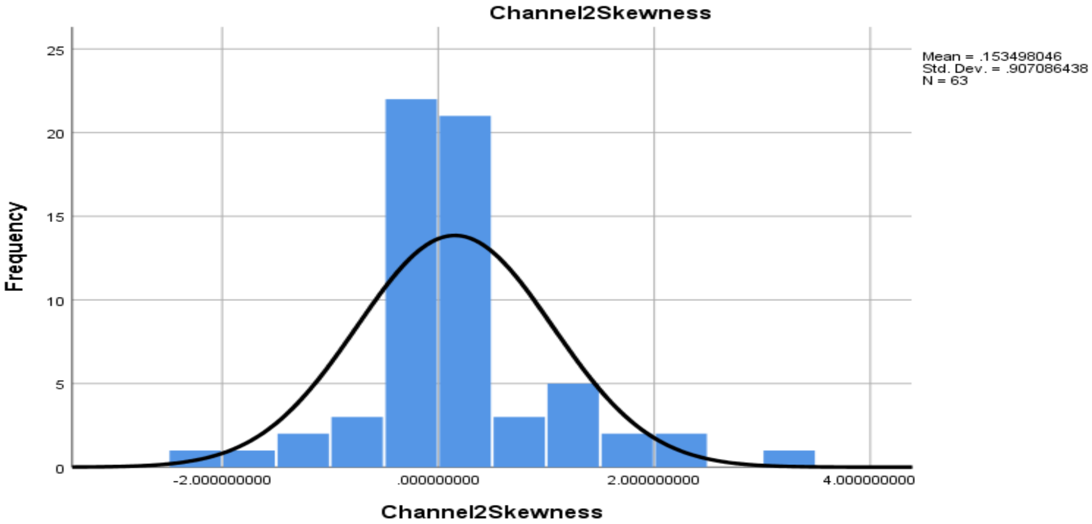


Figure 5.17: Histogram of Class Vs Channel 2 Skewness

In Figure 5.17, it is seen from the histogram of class vs channel 2 skewness that the values are more frequent slightly at left side of the graph.

5.9.18 Class Vs Channel 1 Variance

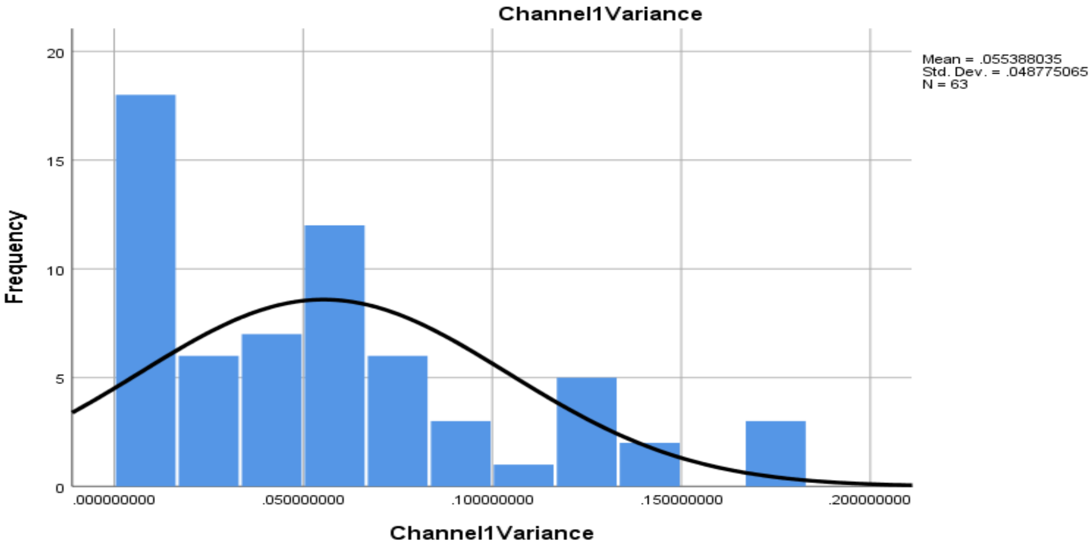


Figure 5.18: Histogram of Class Vs Channel 1 Variance

In Figure 5.18, the histogram of class vs channel 1 variance is displayed. From the normal curve it is understood that the values are more frequent to left side of the graph.

5.9.19 Class Vs Channel 2 Variance

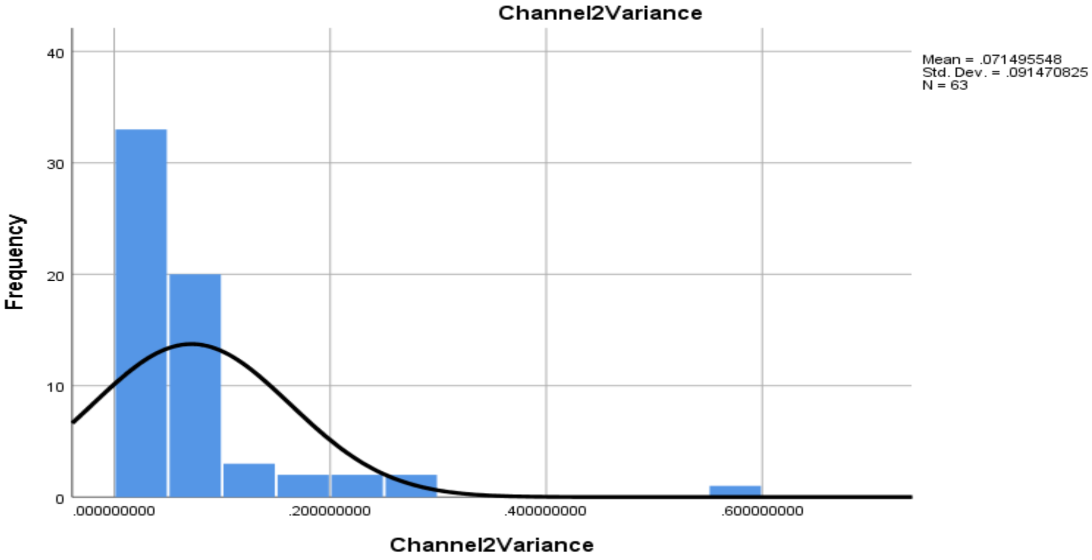


Figure 5.19: Histogram of Class Vs Channel 2 Variance

From Figure 5.19, it is seen that the values are more frequent at the left side of the graph. Also the mean and standard deviation of channel 2 variance are .071 and .091 respectively.

5.9.20 Class Vs Channel 1 Standard Deviation

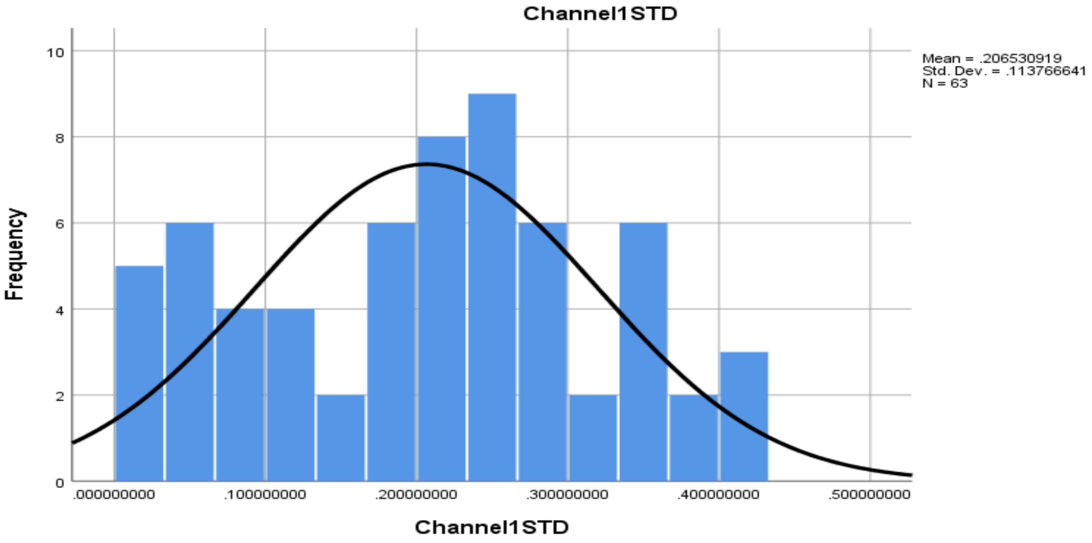


Figure 5.20: Histogram of Class Vs Channel 1 Standard Deviation

In Figure 5.20, the histogram of class vs channel 1 standard deviation is shown. It is seen that the values are more frequent at the center of the graph.

5.9.21 Class Vs Channel 2 Standard Deviation

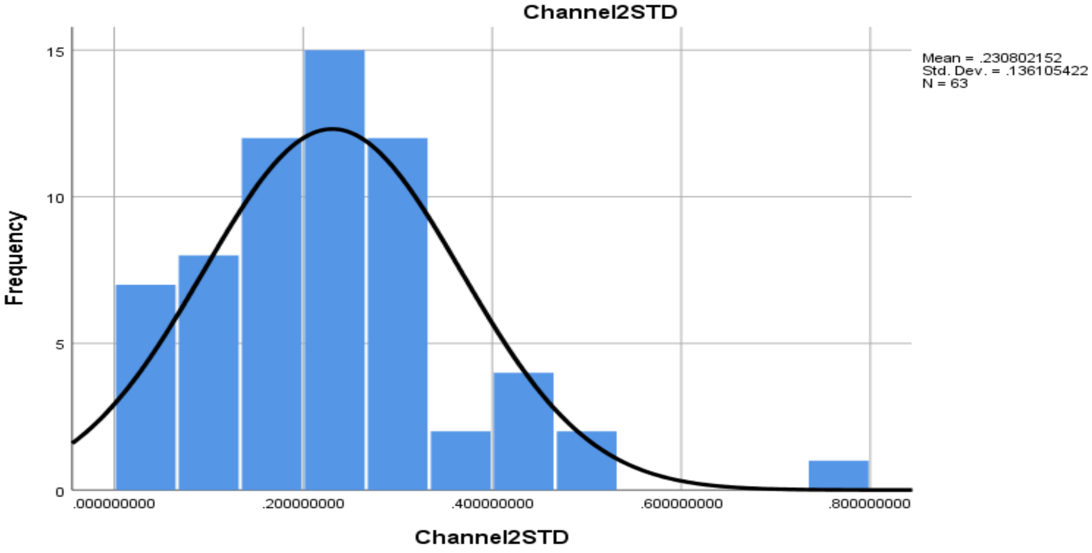


Figure 5.21: Histogram of Class Vs Channel 2 Standard Deviation

In Figure 5.21, the histogram of class vs channel 2 standard deviation shows that the values are more frequent at the left side of the graph. When the channel 2 standard deviation value is .2 it is more frequent compared to other values.

CHAPTER VI

CONCLUSION & RECOMMENDATION

6.1 Conclusion

The work presented in this thesis has been corncerned with acquiring and classifying the data of eyeball movement directions. Data acquisition is done using Biopac Student Lab Software. 7 subjects are told to look at 9 different directions. With the help of Biopac MP3X acquisition unit and surface electrodes, data are taken from all the subjects. For each direction, 15 trials have been performed for 20 sec to get the accurate data.

In this research, different directions of eyeball are classified using machine learning algorithms. EOG data is used because eyeball movement is present even if the person is full-body paralyzed. Six different algorithms are used to classify the directions of the eyeball. Total 9 directions are classified for 7 subjects. By combining all the data, an EOG dataset has been prepared to do the classification for developing multiple class HCI system. In this experiment, all the directions are considered as classes. With the help of Weka 3.9.2 software, the EOG dataset is properly classified into multiple classes. By comparing among the classification algorithms, KNN has the lowest accuracy of 7.6923% and Bagging has the highest accuracy of 92.31%. Random Forest also performs well on this dataset. It has 84.61% accuracy. So, it is seen that Bagging is the best suited algorithm for classifying the different directional movement of the eyeball.

As for correlation, Chi-square test/ Fisher's Exact test has been performed. The correlation is found between 16 features and 9 classes. Correlation is actually done to find the significant features which is related to the class. It also shows which feature has impact on the class. Depending on the p-value the features are correlated with the classes. In this experiment, mean of Channel 1 and Channel 2 are the features on which the directional eyeball movements are directly depended. Standard deviation, variance and kurtosis of channel 1 and blood pressure have some significance for the class (direction). Other features are not that significant. Mainly based on the chi-square test this correlation is done. Fisher's exact test is functional for only 2 X 2 classes. But in this dataset, there are total 9 classes. So, the Fisher's Exact test cannot be performed on this dataset. In this experiment, Chi-square test results are cross checked with ranker method in Weka 3.9.2 which is done based on informaton gain of the features. That cross checked results are the final result for correlation technique.

There are some limitations of this experiment. Proper data acquisition is one of the major limitations for this experiment. Because finding a subject is pretty difficult. Electrode placement on the subject is another issue. The exact position of the electrodes on the human skull is hard to locate for acquiring the EOG data. There is a chance of motion artifact while acquiring the data cause the surface electrode used in this experiment are sensitive to the motion. So the subject should not move his/her head while performing the experiment. To obtain the corner directional data such as upleft, upright, downleft and downright is another problem. As these data have both horizontal and vertical channel value. So the gazing has to be accurate and precise while obtaining the corner directional data. Preprocessing the data is another issue in this experiment. This can be solved by using the Microsoft Excel to prepare the proper EOG dataset. Some data are misclassified due to the limitations of the classification algorithms. In correlation method, fisher's exact test could not be performed as there are 9 classes in the dataset. Fisher's exact test can be performed for 2x2 matrix. So, in this experiment only the chi-square test is performed to find the correlation between the features and the classes.

6.2 Fulfilling the Goal of the Thesis

The main goal of this research was to develop a multiple class HCI system which has been achieved using machine learning algorithms on the EOG dataset. Applying different classification algorithms on this type of data is a new approach. In previous researches, only 4 basic eyeball directions have been considered whereas in this experiment 9 different directional eyeball movements are considered and classified. With this all the eyeball movement directions are covered. It really helps to make the HCI system more sTable and accurate.

Previous investigations on EOG dataset are done using neural network, SVM or KNN classifier. In this proposed method, six machine learning algorithms are used to classify the data for developing the multiple class HCI system. After comparing the algorithms, the best suited classifier is found for EOG dataset which is Bagging.

In correlation method, the features are properly correlated with the classes. From chi-square test result, it is shown that Channel 1 and Channel 2 are the most significant features of the prepared EOG dataset for eyeball movement.

6.3 Recommendations for Future Work

In this research various algorithms are used to classify the EOG data and significant features are found applying the correlation. For developing hands-free HCI devices, the proposed method can be very useful. The outcome of this research can help not only the paralyzed people but also the general people with some modifications. The following recommendations are suggested on the basis of the thesis results:

- EEG data can be used instead of EOG data. EEG is the electrical activity of human brain. HCI devices can be controlled using EEG. If this data is processed properly, just by thinking a device can be controlled easily.
- An intelligent system can be developed using machine learning algorithm. This system can predict the data and can automatically control the HCI device. This type of device can be very useful to make a fully automated workstation.
- Multiple class HCI can be implemented in the hospital. EOG controlled wheelchair can be very useful to transport patient from one place to another inside the hospital.
- Various other machine learning algorithms can be applied to improve the accuracy of EOG controlled HCI devices. This can stabilize the system.
- Obstacle detection system can be implemented in the HCI device. This will help the system to avoid any collision with other objects.

References

[1] G. Chao, "Human-Computer Interaction: Process and Principles of Human-Computer Interface Design", in *2009 International Conference on Computer and Automation Engineering*, Bangkok, Thailand, 2009, pp. 230-233.

[2] P. Alafaireet, "Graphic User Interface: Needed Design Characteristics for Successful Physician Use", in *2006 ITI 4th International Conference on Information & Communications Technology*, Cairo, Egypt, 2006, pp. 1-1.

[3] M. Mustafa, "Importance of Human-Computer Interaction", Usman Institute of Technology, 2017.

[4] P. Paul, A. Kumar and M. Ghosh, "Human Computer Interaction and its Types: A Types", Bengal Engineering and Science University, Howrah, West Bengal, India, 2018.

[5] K. Kozlikova, "Biological Signals-Biosignals", Bratislava,Slovakia, 2011.

[6] H. AlJobouri, İ. Çankaya and H. Jaber, "Biosignal Processing, Medical Imaging and fMRI (BSPMI) Software Package Based on MATLAB GUI for Education and Research", *International Journal of Scientific Research in Information Systems and Engineering (IJSRISE)*, vol. 1, no. 2, pp. 35-43, 2015. [Accessed 16 June 2019].

[7] C. Harland, T. Clark and R. Prance, "Electric potential probes - new directions in the remote sensing of the human body", *Measurement Science and Technology*, vol. 13, no. 2, pp. 163-169, 2001. Available: 10.1088/0957-0233/13/2/304 [Accessed 16 June 2019].

[8] J. Kim, S. Kwon, S. Seo and K. Park, "Highly Wearable Galvanic Skin Response Sensor using Flexible and Conductive Polymer Foam", in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, USA, 2014, pp. 6631-6634.

[9] H. Rongen, V. Hadamschek and M. Schiek, "Real Time Data Acquisition and Online Signal Processing for Magnetoencephalography", in *14th IEEE-NPSS Real Time Conference, 2005*, Stockholm, Sweden, 2005, p. 3 pp.

[10] M. Porcheron, J. Fischer, S. Reeves and S. Sharples, "Voice Interfaces in Everyday Life", in *2018 Conference on Human Factors in Computing Systems*, Montreal, Canada, 2018, pp. 1-12.

[11] Y. Bai and W. Hsu, "An Improvement Design of a Four-quadrant and Voice Interaction User Interface of a Smartphone for the Visually Impaired User", in *2016 IEEE 5th Global Conference on Consumer Electronics*, Kyoto, Japan, 2016, pp. 1-2.

[12] J. Seong and Y. Choi, "Design and Implementation of User Interface through Hand Movement Tracking and Gesture Recognition", in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, South Korea, 2018, pp. 552-555.

[13] J. Kim, J. Park, H. Kim and C. Lee, "HCI (Human Computer Interaction) Using Multi-Touch Tabletop Display", in *2007 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, BC, Canada, 2007, pp. 391-394.

[14] W. Buxton, R. Hill and P. Rowley, "Issues and techniques in touch-sensitive Tablet input", *Proceedings of the 12th annual conference on Computer graphics and interactive techniques - SIGGRAPH '85*, vol. 19, no. 3, pp. 215-224, 1985. Available: 10.1145/325334.325239 [Accessed 17 June 2019].

[15] M. Reaz, M. Hussain and F. Mohd-Yasin, "Techniques of EMG signal analysis: detection, processing, classification and applications", *Biological Procedures Online*, vol. 8, no. 1, pp. 11-35, 2006. Available: 10.1251/bpo115 [Accessed 17 June 2019].

[16] I. Moon, M. Lee, J. Chu and M. Mun, "Wearable EMG-based HCI for Electric-Powered Wheelchair Users with Motor Disabilities", in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Barcelona, Spain, 2005, pp. 2649 - 2654.

[17] B. Champaty, J. Jose, K. Pal and A. Thirugnanam, "Development of EOG Based Human Machine Interface control System for Motorized Wheelchair", in *2014 Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD)*, Kottayam, India, 2014, pp. 1-7.

[18] M. Chowdhury et al., "Simplistic Approach to Design a Prototype of an Automated Wheelchair Based on Electrooculography", in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, Rajshahi, Bangladesh, 2018, pp. 1-4.

[19] M. Chowdhury, M. Mollah, M. Raihan, A. Ahmed, M. Halim and M. Hossain, "Designing a Cost Effective Prototype of an Automated Wheelchair Based on EOG (Electrooculography)", in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018, pp. 1-4.

[20] J. Kumar and P. Bhuvaneswari, "Analysis of Electroencephalography (EEG) Signals and Its Categorization–A Study", *Procedia Engineering*, vol. 38, pp. 2525-2536, 2012. Available: 10.1016/j.proeng.2012.06.298 [Accessed 17 June 2019].

[21] H. Park, B. Myung and S. Yoo, "Power consumption of wireless EEG device for Bel application", in *2013 International Winter Workshop on Brain-Computer Interface (BCI)*, Gangwo, South Korea, 2013, pp. 100-102.

[22] T. Doyle, Z. Kucerovsky and W. Greason, "Design of an Electroocular Computing Interface", in *2006 Canadian Conference on Electrical and Computer Engineering*, Ottawa, Ont., Canada, 2006, pp. 1458 - 1461.

[23] D. Betancourt and C. del Rio, "Study of the human eye working principle: an impressive high angular resolution system with simple array detectors", *Fourth IEEE Workshop on Sensor Array and Multichannel Processing, 2006.*, pp. 93-97, 2006. Available: 10.1109/sam.2006.1706098 [Accessed 17 June 2019].

[24] C. Willoughby, D. Ponzin, S. Ferrari, A. Lobo, K. Landau and Y. Omidi, "Anatomy and physiology of the human eye: effects of mucopolysaccharidoses disease on structure and function - a review", *Clinical & Experimental Ophthalmology,* vol. 38, pp. 2-11, 2010. Available: 10.1111/j.1442-9071.2010.02363.x [Accessed 17 June 2019].

[25] D. Purves, G. Augustine, D. Fitzpatrick, W. Hall, A. LaMantia and L. White, *Neuroscience,* 3rd ed. Sunderland, Massachusetts.: Sinauer Associates, Inc., 2012.

[26] T. Foulsham, "Eye movements and their functions in everyday tasks", *Eye,* vol. 29, no. 2, pp. 196-199, 2014. Available: 10.1038/eye.2014.275 [Accessed 17 June 2019].

[27] C. Kavitha and G. Nagappan, "Sensing and Processing of EOG Signals to Control Human Machine Interface System", *International Journal of Science, Engineering and Technology Research (IJSETR),* vol. 4, no. 5, pp. 1330-1336, 2015. [Accessed 17 June 2019].

[28] H. Singh and J. Singh, "A Review On Electrooculography", *International Journal of Advanced Engineering Technology,* vol. 3, no. 4, pp. 1-8, 2012. [Accessed 17 June 2019].

[29] "What is Machine Learning? A definition - Expert System", *Expertsystem.com*, 2019. [Online]. Available: https://www.expertsystem.com/machine-learning-definition/. [Accessed: 17- Jun- 2019].

[30] "List of Machine Learning Algorithms with Details [2018 Updated]", *New Tech Dojo,* 2018. [Online]. Available: https://www.newtechdojo.com/list-machine-learning-algorithms/. [Accessed: 17- Jun- 2019].

[31] R. Dalinina, "Introduction to Correlation", *Datascience.com,* 2017. [Online]. Available: https://www.datascience.com/blog/introduction-to-correlation-learn-data-science-tutorials. [Accessed: 17- Jun- 2019].

[32] T. Okazaki and T. Hase, "A Hand-free Device Operation Method for Home Appliances", in *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE),* Tokyo, Japan, 2014, pp. 136 - 137.

[33] S. Bou-Ghazale and A. Asadi, "Hands-Free Voice Activation of Personal Communication Devices", in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100),* Istanbul, Turkey, 2000, pp. 1735 - 1738 vol.3.

[34] H. Heidenreich, "What are the types of machine learning?", *Towards Data Science,* 2018. [Online]. Available: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f. [Accessed: 17- Jun- 2019].

[35] M. Bramer, *Principles of Data Mining*. London, United Kingdom: Springer, 2007.

[36] I. Witten, E. Frank and M. Hall, *Data mining Practical Machine Learning Tools and Techniques,* 3rd ed. Burlington, Mass.: Morgan Kaufmann Publishers, 2011, pp. 244-322.

[37] T. Gunasegaran and Y. Cheah, "Evolutionary Cross Validation", in *2017 8th International Conference on Information Technology (ICIT),* Amman, Jordan, 2017, pp. 89 - 95.

[38] R. Pontius and M. Millones, "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment", *International Journal of Remote Sensing*, vol. 32, no. 15, pp. 4407-4429, 2011. Available: 10.1080/01431161.2011.552923 [Accessed 17 June 2019].

[39] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", *Climate Research*, vol. 30, pp. 79-82, 2005. Available: 10.3354/cr030079 [Accessed 17 June 2019].

[40] C. Chen, J. Twycross and J. Garibaldi, "A new accuracy measure based on bounded relative error for time series forecasting", *PLOS ONE*, vol. 12, no. 3, p. e0174202, 2017. Available: 10.1371/journal.pone.0174202 [Accessed 17 June 2019].

[41] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", Flinders University of South Australia, Adelaide,Australia, 2007.

[42] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. Available: 10.1016/j.patrec.2005.10.010 [Accessed 17 June 2019].

[43] J. Brownlee, "How to Calculate Correlation Between Variables in Python", *Machine Learning Mastery*, 2018. [Online]. Available: https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/. [Accessed: 17- Jun- 2019].

[44] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442-451, 1975. Available: 10.1016/0005-2795(75)90109-9 [Accessed 17 June 2019].

[45] "Likelihood Ratios - CEBM", *CEBM*, 2019. [Online]. Available: https://www.cebm.net/2014/02/likelihood-ratios/. [Accessed: 17- Jun- 2019].

[46] J. Sindik and N. Vidak, "Uncertainty coefficient as a method for optimization of the competition system in Table-tennis leagues in "sokaz"", Sport Science 2, Croatia, 2013.

[47] S. Hong, H. Kim, S. Lee and Y. Moon, "Secure Multiparty Computation of Chi-Square Test Statistics and Contingency Coefficients", in *2017 IEEE 3rd International Conference On Big Data Security On Cloud (Bigdatasecurity), IEEE International Conference On High Performance and Smart Computing (Hpsc), And IEEE International Conference On Intelligent Data And Security (Ids)*, Beijing, China, 2017, pp. 53 - 56.

[48] D. Rumsey, "What a p-Value Tells You About Statistical Data - dummies", *dummies*, 2019. [Online]. Available: https://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/. [Accessed: 17- Jun- 2019].

[49] D. Mathew, M. Hajj and M. Abri, "Human-Computer Interaction (HCI): An overview", in *2011 IEEE International Conference on Computer Science and Automation Engineering*, Shanghai, China, 2011, pp. 99-100.

[50] S. Jambukia, V. Dabhi and H. Prajapati, "Classification of ECG signals using Machine Learning Techniques: A Survey", in *2015 International Conference on Advances in Computer Engineering and Applications*, Ghaziabad, India, 2015, pp. 714 - 721.

[51] L. Qi and N. Alias, "Comparison of ANN and SVM for classification of eye movements in EOG signals", *Journal of Physics: Conference Series*, vol. 971, p. 012012, 2018. Available: 10.1088/1742-6596/971/1/012012 [Accessed 18 June 2019].

[52] S. Aungsakul, A. Phinyomark, P. Phukpattaranont and C. Limsakul, "Evaluating Feature Extraction Methods of Electrooculography (EOG) Signal for Human-Computer Interface", *Procedia Engineering*, vol. 32, pp. 246-252, 2012. Available: 10.1016/j.proeng.2012.01.1264 [Accessed 18 June 2019].

[53] L. Deng, C. Hsu, T. Lin, J. Tuan and Y. Chen, "EOG-Based Signal Detection and Verification for HCI", in *2009 International Conference on Machine Learning and Cybernetics*, Hebei, China, 2009, pp. 3342 - 3348.

[54] C. Kavitha and G. Nagappan, "Sensing and Processing of EOG Signals to Control Human Machine Interface System", *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 4, no. 5, pp. 1330-1336, 2015. [Accessed 18 June 2019].

[55] D. Ramkumar, D. Kuma and G. Emayavaramban, "EOG Signal Classification Using Neural Network for Human Computer Interaction", *International Journal of Computer Technology and Applications*, vol. 9, pp. 223-231, 2016. [Accessed 18 June 2019].

[56] P. Syal and P. Kumari, "Comparative Analysis of KNN, SVM, DT for EOG based Human Computer Interface", in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, Mysore, India, 2017, pp. 1023 - 1028.

[57] W. Hamalainen, "Efficient discovery of the top-K optimal dependency rules with Fisher’s exact test of significance", in *2010 IEEE International Conference on Data Mining*, Sydney, NSW, Australia, 2010, pp. 196 - 205.

[58] L. Pirhaji et al., "The performances of the chi-square test and complexity measures for signal recognition in biological sequences", *Journal of Theoretical Biology*, vol. 251, no. 2, pp. 380-387, 2008. Available: 10.1016/j.jtbi.2007.11.021 [Accessed 18 June 2019].

[59] A. López, F. Ferrero, M. Valledor, J. Campo and O. Postolache, "A Study on Electrode Placement in EOG Systems for Medical Applications", in *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Benevento, Italy, 2016, pp. 1-5.

[60] "Biopac Student Lab PRO Manual", *Biopac.com*, 2010. [Online]. Available: https://www.biopac.com/wp-content/uploads/BSL-PRO-3_7-Manual.pdf. [Accessed: 18- Jun- 2019].

[61] A. Unnam, "Weka – Graphical User Interference Way to Learn Machine Learning", *Analytics Vidhya*, 2019. [Online]. Available: https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning. [Accessed: 18- Jun- 2019].

[62] M. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier", in *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, Bangalore, India, 2016, pp. 1-5.

[63] S. Taheri and M. Mammadov, "Learning the naive Bayes classifier with optimization models", *International Journal of Applied Mathematics and Computer Science*, vol. 23, no. 4, pp. 787-795, 2013. Available: 10.2478/amcs-2013-0059 [Accessed 18 June 2019].

[64] P. Kaviani and M. Dhotre, "Short Survey On Naive Bayes Algorithm", *International Journal of Advance Engineering and Research Development*, vol. 4, no. 11, 2017. Available: 10.21090/ijaerd.40826.

[65] D. Srivastava and L. Bhambhu, "Data Classification Using Support Vector Machine", *Journal of Theoretical and Applied Information Technology*, vol. 12, no. 1, pp. 1-7, 2010. [Accessed 18 June 2019].

[66] L. Auria and R. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis", *SSRN Electronic Journal*, 2008. Available: 10.2139/ssrn.1424949 [Accessed 18 June 2019].

[67] H. Park, "An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain", *Journal of Korean Academy of Nursing*, vol. 43, no. 2, pp. 154-164, 2013. Available: 10.4040/jkan.2013.43.2.154 [Accessed 18 June 2019].

[68] V. Fang, "Advantages and disadvantages of logistic regression", *Victor Fang's Computing Space*, 2019. [Online]. Available: https://victorfang.wordpress.com/2011/05/10/advantages-and-disadvantages-of-logistic-regression/. [Accessed: 18- Jun- 2019].

[69] N. Robinson, "The Disadvantages of Logistic Regression", *The Classroom | Empowering Students in Their College Journey*, 2018. [Online]. Available: https://www.theclassroom.com/disadvantages-logistic-regression-8574447.html. [Accessed: 18- Jun- 2019].

[70] I. Gazalba and N. Reza, "Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification", in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, 2017, pp. 294 - 298.

[71] A. Kataria and M. Singh, "A Review of Data Classification Using K-Nearest Neighbour Algorithm", *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 6, pp. 354-360, 2013. [Accessed 18 June 2019].

[72] S. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", *International Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605-610, 2019. [Accessed 18 June 2019].

[73] "Pros and Cons of K-Nearest Neighbors - From The GENESIS", *From The GENESIS*, 2018. [Online]. Available: https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/. [Accessed: 18- Jun- 2019].

[74] E. Saifutdinova, D. Dudysova´, L. Lhotska´, V. Gerla and M. Macasˇ, "Artifact Detection in Multichannel Sleep EEG using Random Forest Classifier", in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018, pp. 2803 - 2805.

[75] S. Ronaghan, "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark", *Medium*, 2018. [Online]. Available: https://medium.com/@srnghn/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3. [Accessed: 18- Jun- 2019].

[76] D. Jain, "What are the advantages and disadvantages for a random forest algorithm?", *Quora*, 2018. [Online]. Available: https://www.quora.com/What-are-the-advantages-and-disadvantages-for-a-random-forest-algorithm/. [Accessed: 18- Jun- 2019].

[77] G. Fumera, F. Roli and A. Serrau, "A Theoretical Analysis of Bagging as a Linear Combination of Classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1293-1299, 2008. Available: 10.1109/tpami.2008.30 [Accessed 18 June 2019].

[78] P. Bühlmann, "Bagging, Boosting and Ensemble Methods", *Handbook of Computational Statistics*, pp. 985-1022, 2011. Available: 10.1007/978-3-642-21551-3_33 [Accessed 18 June 2019].

[79] A. Jurek, Y. Bi, S. Wu and C. Nugent, "A survey of commonly used ensemble-based classification techniques", *The Knowledge Engineering Review*, vol. 29, no. 5, pp. 551-581, 2013. Available: 10.1017/s0269888913000155 [Accessed 18 June 2019].

[80] H. Jabbar and R. Khan, "Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study)", *Computer Science, Communication and Instrumentation Devices*, pp. 163-172, 2014. Available: 10.3850/978-981-09-5247-1_017 [Accessed 18 June 2019].

[81]"SPSS Software", *Ibm.com*, 2019. [Online]. Available: https://www.ibm.com/analytics/spss-statistics-software. [Accessed: 18- Jun- 2019].

[82] "Using Chi-Square Statistic in Research - Statistics Solutions", *Statistics Solutions*, 2019. [Online]. Available: https://www.statisticssolutions.com/using-chi-square-statistic-in-research/. [Accessed: 18- Jun- 2019].

[83] R. Plackett, "Karl Pearson and the Chi-Squared Test", *International Statistical Review / Revue Internationale de Statistique*, vol. 51, no. 1, pp. 59-72, 1983. Available: 10.2307/1402731.

[84] F. YATES, "Tests of Significance for 2 X 2 Contingency Tables", *Journal of the Royal Statistical Society*, vol. 147, no. 3, p. Journal of the Royal Statistical Society, 1984. [Accessed 18 June 2019].

[85] K. More, M. Raihan, A. More, S. Padule and S. Mondal, "A12176 Smart phone based “heart attack” risk prediction; innovation of clinical and social approach for preventive cardiac health", *Journal of Hypertension*, vol. 36, p. e321, 2018. Available: 10.1097/01.hjh.0000549311.94493.c1 [Accessed 12 August 2018].

[86] S. Lei, "A Feature Selection Method Based on Information Gain and Genetic Algorithm", in *2012 International Conference on Computer Science and Electronics Engineering*, Hangzhou, China, 2012, pp. 1-4.

[87] D. Joanes and C. Gill, "Comparing measures of sample skewness and kurtosis", *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 1, pp. 183-189, 1998. Available: 10.1111/1467-9884.00122.

[88] "InfoGainAttributeEval", *sourceforge*, 2019.

Appendix (Sample of the EOG dataset)

Channel 1 Mean	Channel 2 Mean	Age	Gender	Height (cm)	Weight (Kg)	Blood Group	Eyesight problem	Color Blind	Sleep Duration (Hrs)	Left/Right Handed	Blood Pressure	Heart Disease	Direction
-0.02959	0.076045	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Up
0.027016	0.064166	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Down
-0.10404	-0.09249	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Left
0.03603	-0.00273	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Right
-0.01291	0.006883	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Upleft
-0.03387	-0.00035	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Upright
0.004276	0.015436	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Downleft
0.008609	-0.02428	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Downright
0.016339	0.016479	25	Male	175	50	O+	Yes	No	7	Left Handed	No	No	Blink
-0.02745	0.07551	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Up
0.028027	0.061349	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Down
-0.10054	-0.09233	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Left
0.036985	-0.00212	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Right
-0.01274	0.006262	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Upleft

-0.03224	-0.00032	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Upright
0.004319	0.01583	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Downleft
0.008234	-0.02252	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Downright
0.016365	0.016313	24	Male	159	54	A+	No	No	7	Right Handed	No	No	Blink
-0.02839	0.075976	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Up
0.028069	0.06372	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Down
-0.10145	-0.09194	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Left
0.035833	-0.00245	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Right
-0.01273	0.006168	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Upleft
-0.03258	-0.00032	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Upright
0.004295	0.015816	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Downleft
0.008923	-0.02333	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Downright
0.016269	0.01632	24	Female	157	52	O+	Yes	No	8	Right Handed	Yes	No	Blink
-0.02587	0.07554	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Up
0.02019	0.06172	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Down
-0.10519	-0.09282	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Left

0.035684	-0.00282	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Right
-0.01258	0.006339	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Upleft
-0.03205	-0.00034	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Upright
0.004464	0.015971	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Downleft
0.008854	-0.02139	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Downright
0.016051	0.016122	48	Male	162	70	O+	No	No	8	Right Handed	Yes	No	Blink
-0.02874	0.075877	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Up
0.028491	0.063461	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Down
-0.10179	-0.09188	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Left
0.035551	-0.00287	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Right
-0.01228	0.006575	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Upleft
-0.03134	-0.00038	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Upright
0.004326	0.015408	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Downleft
0.008916	-0.02273	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Downright
0.016166	0.016414	24	Male	162	54	O+	No	No	8	Right Handed	No	No	Blink
-0.02728	0.07374	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Up

0.027627	0.062948	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Down
-0.10026	-0.09065	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Left
0.034402	-0.00291	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Right
-0.01263	0.006283	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Upleft
-0.03236	-0.00033	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Upright
0.004305	0.015386	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Downleft
0.008997	0.023737	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Downright
0.016826	0.016724	23	Female	157	55	B+	No	No	7	Right Handed	No	No	Blink
-0.01578	0.057249	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Up
0.009425	0.04642	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Down
-0.00645	0.035886	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Left
0.02313	0.035557	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Right
-0.02053	0.016464	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Upleft
0.010424	0.081914	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Upright
-0.06467	0.004705	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Downleft
0.022472	-0.01892	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Downright

0.020109	-0.23169	23	Female	162	40	O+	No	No	7	Right Handed	No	No	Blink

List of Publications

[1] M. Chowdhury et al., "Simplistic Approach to Design a Prototype of an Automated Wheelchair Based on Electrooculography", in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, Rajshahi, Bangladesh, 2018, pp. 1-4.

[2] M. Chowdhury, M. Mollah, M. Raihan, A. Ahmed, M. Halim and M. Hossain, "Designing a Cost Effective Prototype of an Automated Wheelchair Based on EOG (Electrooculography)", in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018, pp. 1-4.