ANALYSIS OF DEMOGRAPHIC AND CLINICAL DATA OF PATIENTS TO DETERMINE THE EFFECTIVE MARKERS OF BIPOLAR DISORDER

By

UTTARA V. TIPNIS

Bachelor of Engineering in Mechanical Engineering

Pune University

Pune, India

2013

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July, 2016

STATISTICAL ANALYSIS OF DEMOGRAPHIC AND CLINICAL DATA OF PATIENTS TO DETERMINE THE MARKERS OF BIPOLAR DISORDER

Thesis Approved:

Dr. Sunderesh Heragu

Thesis Adviser

Dr. Manjunath Kamath

Dr. Dursun Delen

Name: UTTARA V. TIPNIS

Date of Degree: JUNE 2016

Title of Study: STATISTICAL ANALYSIS OF DEMOGRAPHIC AND CLINICAL DATA OF PATIENTS TO DETERMINE THE MARKERS OF BIPOLAR DISORDER

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract:

Bipolar disorder is also known as manic depression. Patients with this disease exhibit symptoms of depression and mania or hypomania in a cyclical manner. While depression symptoms are relatively easy to detect, manic and hypomanic symptoms are not. As a result, patients with bipolar disorder are either misdiagnosed as suffering from just depression or are diagnosed late – typically five to ten years from the onset of the disorder. The mood of a bipolar patient misdiagnosed as having depression and treated only for that condition can become elevated to a state of hypermania (Matza et al., 2005 and Charney et al., 2003). A late diagnosis can worsen the bipolar condition as well. Thus, a misdiagnosis or late diagnosis could aggravate the symptoms, and may require the bipolar patient to be hospitalized. This situation is made worse by the presence of psychological and/or physiological comorbidities that commonly coexist in patients with bipolar disorder.

In this thesis, we apply logistic regression models, decision trees, and artificial neural networks to detect the existence of bipolar disorder in a patient with that disease at an early stage by analyzing their clinical and sociodemographic data, including comorbidities and prescribed medication. The goal is to apply the aforementioned three techniques to detect the existence of bipolarity in a patient with reasonable accuracy, so that he or she may be presented to psychiatrists for further medical diagnosis and treatment. The techniques will also help in screening out the patients needing treatment for other psychiatric disorders, e.g., major depression, that have symptoms similar to bipolar disorder. We use clinical and demographic data from Cerner Health Facts® database and the techniques identify the variables that can help detect bipolarity. We compare the three techniques relative to their effectiveness in detecting bipolar patients for the dataset used in this thesis. Based on the Cerner database, our study also finds that some of the variables identified in the literature as effective predictors of bipolar disorder are not as effective or do not have the same relationship with bipolar disorder.

.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

Bipolar disorder, also known as manic depression, is a psychological disorder that is characterized by drastic mood swings. The patient goes through a cycle of alternating elevated mood (called mania or hypomania, depending on the severity), depression, or a mixed state, which may impair his or her social and work life. Nearly all the patients suffering from manic depression have a comorbid psychological condition, such as anxiety disorder, social phobia, eating disorders, substance abuse, attention deficit hyperactivity disorder (ADHD), premenstrual dysphoric disorder, or panic disorder (Krishnan, 2005). This comorbidity makes it difficult for healthcare providers to diagnose this illness. Following is the description of different mood states in a patient with bipolar disorder (Das et al., 2005 and Slomka et al., 2012):

1.1.1 Symptoms of Mania

1) At least a week of extremely elevated mood and/or irritability
2) Speaking excessively in a rapid and uninterruptible manner
3) Racing thoughts and easily getting distracted
4) Agitation

5) Excessive risk taking or impulsive behavior, such as hypersexuality, drug or alcohol abuse, and unwarranted spending of money.
6) Impaired judgment
7) Decreased need for sleep
8) Psychosis, delusions, hallucinations (in extreme cases).

Mania impairs a patient's ability to work or socialize. Manic patients must be institutionalized to treat this condition and also to prevent the patient from harming themselves or others. If untreated, this condition can last for months.

1.1.2 Symptoms of Hypomania

1) At least four days of elevated mood and/or irritability
2) Increased productivity
3) Increased creativity
4) Decreased need for sleep
5) Poor judgment
6) Increased energy levels
7) Hyperactivity.

Hypomania, as the name suggests, is a mild form mania that does not result in a need for the person to be institutionalized. Because the symptoms include increased productivity and creativity, a hypomanic episode may actually be enjoyable for the person experiencing it. The symptoms of hypomania do not include psychosis and they do not impair the person's ability to socialize or work.

1.1.3 Symptoms of Depression

1) Decreased energy levels
2) Persistent sadness and feelings of guilt, isolation, loneliness, self-loathing, apathy, indifference, anger, and hopelessness.
3) Increased need for sleep
4) Change in appetite
5) Excessive fatigue
6) Loss of interest in usually enjoyable activities
7) Social anxiety
8) Chronic psychosomatic pain
9) Loss of concentration
10) Irritability
11) Recurrent suicidal thoughts
12) Loss of interest in sexual activities
13) Excessive crying (with or without cause)
14) A general negative outlook on life
15) Psychosis, delusions, hallucinations (in extreme cases).

Depression symptoms are easy to observe and the disease can be easily detected by a family member, doctor, or nurse. If untreated, this state can last for months. It hampers a person’s ability to work and socialize.

1.1.4 Symptoms of Mixed State

A person going through a mixed episode experiences the symptoms of both mania and depression at the same time. During an episode, a patient may feel extremely energetic and sad at the same time. The risk of suicide is the highest during a mixed episode.

1.1.5 Subtypes of Bipolar Disorder

The following four subtypes of bipolar disorder have been defined by American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-V):

- *Bipolar I disorder:* Characterized by at least one manic episode and depression.
- *Bipolar II disorder:* Characterized by one or more hypomanic episodes and at least one major depressive episode. It is often misdiagnosed as major depressive disorder due to the difficulty in identifying hypomania.
- *Cyclothymia:* Characterized by hypomanic and depressive episodes, that are not as severe as major depression.
- *Bipolar disorder NOS (Not Otherwise Specified):* When the symptoms do not fit any of the above subtypes, the patient is diagnosed as having bipolar disorder NOS.

Rapid cycling bipolar disorder is a severe form of this illness where the symptoms of major depression, mania, hypomania, and hypomania all occur within a year.

1.1.6 Diagnosis

Bipolar disorder cannot be diagnosed by a blood test or a brain scan, but these tests are carried out to rule out the possibility of other conditions that may be causing mood swings, including a thyroid condition, human immunodeficiency virus (HIV), syphilis, or epilepsy. The doctor typically conducts a physical examination, an interview, and lab tests in order to diagnose this illness. Family history of mental illnesses and medical and behavioral history of the patient are discussed in the interview. The family members, friends, and people who come in contact with the patient on a daily basis are also interviewed by the doctor to know more about the patient's behavior.

1.1.7 Treatment

The treatment options for bipolar disorder typically consist of medication (for mood stabilization or sleep), psychotherapy (Cognitive Behavioral Therapy, family-focused therapy, psychoeducation, interpersonal and social rhythm therapy), or electroconvulsive therapy.

If left untreated, bipolar disorder can worsen over time, with an increase in the frequency of episodes and their severity. Misdiagnosis of bipolar disorder as major depression may lead to further problems, as the medications meant for the two illnesses are different. If a bipolar person with a manic/hypomanic episode is treated with antidepressants, chances are that the symptoms of mania/hypomania will worsen.

1.2 Problem statement

Misdiagnosis or late diagnosis of bipolar disorder is very common especially in cases where the patients exhibit depression and hypomania symptoms. A delay in the correct diagnosis of bipolar disorder exacerbates the patient's condition and they begin exhibiting symptoms of major depression (MD) and hypomania. The presence of MD symptoms further delays the correct diagnosis of the bipolar disorder (Dean et al., 2004 and Sanchez-Moreno et al., 2009). Moreover, the treatment for major depression only aggravates the symptoms for bipolar disorder, as the medication works towards getting the patient out of depression, which results in the patient becoming manic when he or she is already in a normal or hypomanic mood state. This misdiagnosis generally occurs due to differences in judgements of different medical professionals and/or incorrect or incomplete information provided by the patients and their family members in psychiatric interviews conducted for diagnosis. Misdiagnosed bipolar patients have been found to have received inappropriate and expensive treatment, which only aggravates the symptoms of this disorder (Matza et al., 2005 and Charney et al., 2003).

The objective of this thesis is to apply two statistical models, namely, logistic regression analysis and decision trees, as well as the artificial neural network technique to detect a patient with bipolar disorder with reasonable accuracy at an early stage by analyzing the patient's clinical and sociodemographic data. Using an available dataset, we compare the performance of these three techniques relative to their efficacy in detecting bipolar disorder in patients. Patients who are determined by the technique to have this disorder can be presented to a psychiatrist for further screening and diagnosis to confirm the presence of this disorder. On the other hand, patients who are determined not to have the bipolar disorder can be presented to psychiatrists for detecting other psychological disorders that they might have, for appropriate treatment. We have access to data from

Cerner Health Facts® database, and we will use that to confirm or refute the results of a few other studies in the literature that attempt to identify predictors of the bipolar disorder.

CHAPTER II

LITERATURE REVIEW

In this chapter, we review the existing literature on the detection of bipolar disorder including various methods to detect the presence of bipolar and other psychological disorders in individuals. The papers in this chapter can be classified into the following broad categories:

1) data collection and analysis of various clinical parameters such as heart rate variability (HRV), electrocardiography (ECG), electroencephalography (EEG), polysomnography of bipolar patients to determine the autonomic nervous system (ANS) symptoms of the disorder
2) speech analysis for diagnosing various psychological disorders
3) statistical analysis of clinical and sociodemographic data of bipolar patients to determine the predictors of this disease

2.1 Analysis of autonomic nervous system symptoms of bipolar disorder

Migliorini et al. (2012) present the relationship between linear and non-linear physiological parameters during sleep and several mood states in a patient with bipolar disorder. The study was conducted using ECG, EEG, electrooculography, and electromyography data collected per the PSYCHE (PerSonalized monitoring sYstems for Care in mental Health, a personalized, pervasive, cost-effective, and multi-parametric data acquisition system used for mood disorder management) project protocol from one bipolar patient (37 year old bipolar woman who has undergone electroconvulsive therapy) and eight healthy control subjects. The bipolar patient also had to go undergo a mood assessment by a psychologist at intervals of one week during the course of the study in different mood states. The control group was made up of women in the age group of 18 to 45 years on oral contraception, without a history or comorbidity of psychiatric disorders, and without the need for any heavy medication.

From the collected data, several linear parameters were obtained by calculation. Among these, the time-domain parameters were the mean and standard deviation of the RR intervals (also known as heart rate variability, or HRV, where R is the point corresponding to the peak of the QRS complex of the ECG wave) and the root mean square of successive differences between subsequent intervals. Frequency-domain parameters (very low frequency or VLF, low frequency or LF, and high frequency or HF) were calculated by assigning bands of frequency to each parameter and then counting the number of RR intervals that match each frequency. The frequency-domain parameters are used for computing the balance between the sympathetic nervous system (part of autonomic nervous system that activates the flight-or-flight response) and the vagus nerve, which together control the heart’s activity. From the polysomnography results, the sleep stages of the subjects were classified as non-REM (rapid eye movement), REM, and WAKE. This was done by studying the body movements and by using the linear and non-linear parameters obtained from the RR intervals.

The results of this study show that the REM sleep percentage, mean and standard deviation of the RR intervals, Lempel-Ziev complexity, and the sample entropy for the bipolar patient are significantly and consistently different from the normal ranges of these parameters in all the healthy test subjects during all the mood states.

Valenza et al. (2014) present research aimed at detecting bipolar disorder by taking into consideration its psycho-physiological markers. They focus on ECG, respirogram, and body posture information of subjects to detect a pattern of objective physiological parameters that support diagnosis of the mood state. Mood recognition is modeled as an intra-subject evaluation modeled on Markov chains. This study makes use of the PSYCHE platform as well for data acquisition. All the 8 subjects for the study were 18-65 years of age, had low suicidal tendencies, had no cognitive impairment, did not have substance abuse problems, and needed a change in their current treatment.

The mood states of the patients were evaluated by psychologists as per DSM-IV (Diagnostic and Statistical Manual of Mental Disorders – IV, published by the American Psychiatrist Association and used as a standard criteria for the classification of mental disorders). With the help of the data collected through the PSYCHE system from the patients, the researchers then diagnosed their mood at that time by modelling the dataset as standard and then as Markov chains. It was found that the overall accuracy of Markov models was higher than standard models. The paper suggests that this novel system can be used to detect the mood states better in bipolar patients, leading to better management of this illness and also to possibly decide the course of the disorder.

Voss et al. (2006) reviewed the available literature on bipolar disorder and explained why it is necessary to study autonomic changes in this mental illness. They also mention that there is a dearth of literature describing the effect of manic episodes in bipolar disorder on ANS. According to the authors, autonomic dysfunction in depressed patients has attracted more attention as the risk of cardiovascular morbidity and mortality is relatively high in them. It has also been observed that patients who suffer depression after a myocardial infarction (i.e., heart attack) have a low long-term survival rate. The paper then goes on to describe how HRV, blood pressure variability (BPV), and baroreflex sensitivity (BRS) are measured.

Voss et al. (2006) also discuss a study that compared the HRV, BPV, BRS, and NLD of unmedicated and medicated severely depressed patients (major depression; DSM-IV) over a 30 minute time period at rest by means of the task force monitor developed by CNSystems, Graz, Austria. Some of the patients who were studied were on selective serotonin reuptake inhibitors (SSRIs, antidepressants), while the others were not. Age and gender were the other control parameters. The study found that the HR of unmedicated patients was significantly higher. Further results show that depressed and unmedicated patients are no different from the controls in terms of time and frequency domain parameters of HRV, BPV, and BRS.

The authors conclude that more research is needed to determine the effect of manic episodes on ANS activity, and that both HRV and BPV measured in time and frequency-domain as well as with nonlinear techniques are important for this purpose. They also suggest that circadian rhythm, sleep cycles, pupil reactivity, and pain perception are also important parameters that change in bipolar patients depending upon the mood state. The authors also observe that it is necessary to study subjects in the absence of any medication, including antidepressants, neuroleptics, and benzodiazepine, as they have vagolytic effects (inhibit the action of vagus nerve on heart, gastrointestinal tract, and other organs) on HRV and BPV.

Koukopoulos et al. (2013) describe a research conducted to determine the effect of the two major course sequences, i.e., mania/hypomania following episodes of major depression (DMI) and mania/hypomania preceding episodes of depression (MDI), on how patients respond to lithium treatment. The research focused on studying computerized clinical records and life-charts of 855 bipolar-I and bipolar-II (BD-I and BD-II) patients.

The research described in this paper has its motivation in a study by Kukopulos et al. (1975) that first identified the difference between DMI and MDI sequences, and their different prevalence amongst BD-I and BD-II patients. This study is also motivated by the findings of Baldessarini et al. (2010a, 2010b, 2010c, 2012a) who indicate that future morbidity of depression or mania is often dependent upon the nature of the first episode and predominance of either mania/hypomania or depression.

Koukopoulos et al. (2013) studied computerized medical records, life-charts, and long-term assessments of adult BD-I and BD-II patients who sought treatment at the Lucio Bini Mood Disorders Centers in Rome and Caligari, Italy between 1990 and 2012. Data was collected from subjects by means of interviews, diagnostic assessments, treatment, and follow-ups conducted by the researchers and the course sequences were decided based on life-charts of the subjects. Treatment plans for patients were decided on a case-by-case basis and generally included the use of lithium carbonate, anticonvulsants, antipsychotic agents, and counselling. Statistical analysis was carried out on the collected data to determine the differences between subjects with DMI versus MDI course sequence.

Koukopoulos et al. (2013) found out that DMI and MDI have a very similar prevalence amongst the subjects. The paper also concludes that the course sequences in BD-I and BD-II are markedly different, owing to the prevalence of mania in the former and hypomania in the latter. It was shown that DMI subjects were more likely to be diagnosed as BD-II, whereas MDI patients were more likely to be diagnosed with BD-I. Rapid cycling was also found to be more prevalent amongst BD-II patients. The research also found that DMI sequence is more frequent in women than in men, and that the onset of illness occurs earlier in MDI subjects. The authors also note that DMI patients' response to mood stabilizers and lithium during long-term treatment is significantly lower than those of MDI patients.

Greco et al. (2012) describe a research focused on detecting the effect of mood states in bipolar patients on their electrodermal activity (ED), also known as skin conductance (SC). SC is basically the electrical conductance of skin that varies depending on sweating, which is controlled by the ANS. The ANS dysfunction causes mood changes in bipolar patients, which can be physically measured and used to predict what mood state the subject is experiencing at the moment. This research is also a part of the European project PSYCHE.

Three patients without suicidal tendencies, delusions, and hallucinations were selected and were screened with a psychiatric interview to be classified as euthymic, depressed, or in a mixed state. They were also given a questionnaire to determine their level of anxiety.

Then, the subjects were put through an experimental procedure where they first had to rest for five minutes with their eyes closed, then for five minutes with their eyes open. After this, the subjects watched an IASP (International Association for the Study of Pain) presentation for six minutes and then a TAT (Thematic Apperception Test) presentation for four minutes. The patients were put through this protocol over a period of 75 days in order to induce an emotional response, which would result in a change in the SC. The experiment was conducted each time the subject underwent a mood change, which was decided based on a clinical evaluation. The results of this experiment show that the data acquired from each patient was significantly different for each mood state.

2.2 Speech analysis for diagnosing psychological disorders

Vanello et al. (2012) describe an algorithm that assesses the mood state of bipolar patients by studying their speech patterns. The research is motivated from previous literature that describes how rhythm, stress, intonation, and pitch of speech are different in depressed patients when compared to control subjects. This work specifically studies the difference in mean and standard deviation of pitch and jitter of speech, and the speed at which a person talks, in control and bipolar subjects. This work is also a part of the European project PSYCHE.

Vanello et al.'s research is classified into three different phases: In the first phase, the algorithm was tested on a speech database (CMU Arctic Database); in the second phase, the algorithm was applied to patients' speech recordings; while the third phase assessed mood states according to the speech pattern. The results of this work show that the mean and standard deviation of pitch and the jitter in speech change for each subject based on the mood state he or she is in at the time of the recording. They are not consistent for all the subjects; however, for some subjects the mean and standard deviation are lower when depressed, while they are higher for the others. The authors mention that patients' anxiety level must be taken into consideration, because that can also affect the speech.

de Jong and Wempe (2009) present a computer program written in the software Praat ("talk" in Dutch; developed by Paul Boersma and David Weenink in 2007) aimed at detecting syllable nuclei (the peak of a syllable, generally located at the vowel) and calculating the interval between consecutive nuclei, in order to calculate speech rate as the number of syllables per unit time. The program first extracts the intensity of speech and then marks the possible syllables (peaks above a specified threshold above the median). Actual syllable nuclei are then extracted from this data, from which the speech rate is calculated. The authors indicate that this technology can be used to detect the mood state a bipolar person is in (based on the speech rate and pattern), as a patient will have running speech when manic/hypomanic and slower speech when he or she is depressed. It is also useful in assessing the extent to which the patient is depressed or manic, depending on the speech rate.

Cannizzaro et al. (2004) describe research aimed at replicating a prior published result (Stassen et al., 1998) that indicates the speed and pitch at which a person talks is closely related to his or her mood state. The study was performed on audio and video recordings of interviews of five men and two women, which are used as standard instructional tapes for training psychologists. Since their recording, these tapes have been rated for the level of depression of the interviewed patients by several psychologists. These ratings have been used by the authors as standards with which to compare the results obtained from their tests. The authors first calculated the speech rate of the subjects from the number of syllables per unit time. Then, the percentage of time for which the subjects took a pause while speaking was estimated by summing up all the silences of less than 250 ms, subtracting them from 1, and then multiplying the result by 100. The pauses that lasted for more than 250 ms were interpreted as the time when the subject was thinking. Pitch variation of the subjects was calculated by taking a ratio of the standard deviation of the fundamental frequency (85-180 Hz for adult males and 165-255 Hz for adult females) with respect to the mean of fundamental frequency for each subject.

Cannizzaro et al. (2004) indicate that speech rate and pitch variation are inversely proportional to the level of depression (the scale on HDRS), and that the percent pause time and HDRS scale are related only to a certain extent. This means that the speech rate and pitch variation can be safely used to determine the extent to which a person is depressed, while the percent pause time is not a very reliable indicator of depression.

2.3 Statistical analysis of clinical and sociodemographic data of bipolar patients to determine its predictors

Todder et al. (2004) discuss the use of chaos theory and nonlinear analysis to study HRV of euthymic bipolar patients. The data set is made up of 39 subjects and 39 controls, all in the same age group, although the authors used only the ECG strips free of any aberrations, noise, or premature ventricular beat. In the first step, a set of multi-dimensional vectors was built out of the RR intervals. This is known as reconstruction of the attractor, where attractor refers to a set of multi-dimensional variables towards which the chaotic system tends to evolve. After this, Cao's method (Cao, 1997) that uses the false nearest neighbor concept is used in order to estimate the minimum embedding dimension (MED) of the phase space where the attractor is supposed to unfold. Calculation of the largest Lyapunov exponent (LLE) for the embedding dimension was done after this. LLE is the quantity characterizing the rate of separation of close trajectories in a phase space that grow exponentially with time.

In the next step, the HRV data was transformed into symbolic representation (consisting of 4 basic symbols) based on the mean and standard deviation of the time series. A set of three symbols is called a ‘word’, and a different distribution of these was obtained for each subject. Shannon entropy for each subject’s distribution was calculated based on the words.

Poincare plot is another nonlinear analysis tool used by the Todder et al (2004). This is a diagram of RRI(i) plotted as a function of RRI(i - ʎ), where ʎ is a predefined delay that mostly takes a value of 1. These plots are analyzed by calculating the standard deviations of the distances of the RR(i) from the lines y=x and y= -x+2×RRI_mean, which are called SD1 and SD2. This is done with HRV analysis software.

The control and subject groups were compared on the basis of t-tests on the previously calculated data. This comparison indicated that there is no significant difference in the values of MED, LLE, SD1, SD2, and SD1/SD2 for the bipolar group and the control group. The authors conclude that these results may be due to the methods used to calculate these values, as a previous research by Cohen et al. (2003) found a tendency towards vagal predominance in euthymic bipolar patients.

Mariani et al. (2012) discuss research aimed at assessing the mood state of bipolar patients based on their HRV. The data for this research was obtained through twelve bipolar type I and II patients. The selection of candidates was done in such a manner as to include only non-suicidal and non-delusional patients, without any sleep-related or neurological comorbidity. The subjects had to periodically visit the psychologist for mood assessment. Their ECG, respiration, and body activity data was collected on the night following this visit for about 7 hours.

From the collected HRV signals, linear (mean and standard deviation of RR intervals in the QRS complex, which is a combination of three graphical deflections seen on a typical ECG, and root mean square of differences between consecutive RR intervals) and non-linear time domain features, and spectral features were extracted. The sleep profile for each patient was assessed automatically with an ANN model that was trained on the 102 datasets obtained from the control group. This neural network classified the sleep stages into wake, REM, and non-REM. Statistical tests (Kolmogorov-Smirnov test to establish that the data is not normally distributed, and Kruskal-Wallis one-way analysis and multiple comparison t-test in order to determine what differentiates the bipolar patients from the control subjects) were also conducted separately on the data classified according to the depression-mania and anxiety levels.

The mean and standard deviations of the data were plotted separately for patients classified per the depression-mania levels and anxiety levels. These plots show that the mean and standard deviation of RR intervals in the QRS complex are significantly lower during mixed and hypomanic mood states. Differences were also observed in the non-linear features extracted from the bipolar and control subjects. It was also observed from the sleep profiles of bipolar patients that they have a significantly higher percentage of REM sleep than the control subjects, and also that the sleep efficiency is lower during mixed and hypomanic mood states.

Mariani et al. (2012) conclude by saying that the tests confirm that bipolar patients in hypomanic or mixed mood state, or with high levels of anxiety, have a lower HRV than the control group. The authors also state that bipolar patients can be differentiated from normal subjects based on their sleep patterns, as they show a lower REM percentage.

Inoue et al. (2015) describe a study conducted on patients who have had a major depressive episode to determine the prevalence of bipolar disorder amongst them. This study also focuses on other factors, such as suicidal tendencies, early onset of a major depressive episode, antidepressant-related switch to mania/hypomania, mixed depression, or two or more mood episodes within a year to determine if they can be effectively used as predictors of bipolar disorder.

The study was conducted across Japan at 23 psychiatric facilities from April to June 2013. For the purpose of this study, the investigators interviewed 448 patients, 114 of which were diagnosed with bipolar disorder, between the ages of 20 and 65 years who have presented with a major depressive episode using the guidelines of DSM-IV-TR using the Mini – International Neuropsychiatric Interview (MINI) to determine whether there have been any previous mood episodes and the presence of any psychiatric comorbidities, such as anxiety disorder. The subjects were also asked to complete three questionnaires designed to determine how severe the major depressive episode was, the health-related quality of life, and the frequency of trauma/abuse in the subject's childhood or adolescence respectively. The exclusion criteria was schizophrenia, psychotic disorders, high risk of suicide attempt, general medical condition induced mood disorder, and substance-induced mood disorder.

The gathered statistical data was analyzed on SAS 9.3 using Fisher's exact test or Wilcoxon test with a two-sided significance level of 5%. The goal of this analysis was to make a comparison between the demographic and clinical characteristics of bipolar disorder and major depressive disorder. Univariate regression analyses were first carried out for the demographic and clinical characteristics in order to identify the predictors and then multivariate analysis was performed for the predictors that had a p value less than 0.001 and an odds ratio of 2. The performance of various permutations and combinations of predictors found out by the multivariate analysis was determined by receiver-operating characteristic (ROC) curve (plot of true-positive rate versus false-positive rate, or sensitivity versus 1 – specificity).

The results of this study identified antidepressant-related switch to mania/hypomania, mixed depression, two or more mood episodes within a year, early age at the onset of a major depressive episode, and suicidal tendencies as the five effective predictors of bipolar disorder.

Takeshima and Oka (2013) describe a study aimed at identifying effective predictors of bipolar II disorder or bipolar disorder not otherwise specified (collectively known as soft bipolarity). The investigators diagnosed 199 patients with soft bipolarity as per the guidelines of DSM-IV-TR standard diagnostic procedures with the patients and their significant other or family members. The characteristics collected for their analysis as predictors of soft bipolarity were family history of bipolarity, age at the onset of first major depressive episode, age at first visit, gender, recurrent major depressive episodes, psychotic features, atypical features, postpartum depression (in females), mixed depression, antidepressant-induced mania/hypomania, antidepressant wear-off, temperament (cyclothymic, hyperthymic, depressive, anxious, and irritable), and lack of response to three antidepressant trials.

The statistical analysis was carried out in two stages – univariate regression and multivariate regression, with two-sided significance level of 5%. Univariate logistic regression analysis was first carried out to determine the features that were significantly correlated with soft bipolarity. After this step, multivariate logistic regression analysis was carried out on different permutations and combinations of the predictors deemed effective from the univariate analysis. After this, an ROC curve was plotted to determine the performance of these predictors in the diagnosis of soft bipolarity.

The results of this study show that family history of soft bipolarity, recurrent major depressive episodes, cyclothymic temperament, early age at the onset of first major depressive episode (< 25 years), and mixed depression are effective predictors of soft bipolarity.

Xiang et al. (2013) discuss a study that was conducted as a part of the Diagnostic Assessment Services for People with Bipolar Disorders in China (DASP) undertaken by the Chinese Society of Psychiatry. Subjects were recruited at 13 psychiatric facilities across China, a total of 1487 patients. All the subjects studied were between 16 and 65 years of age and had a diagnosis of major depressive disorder as per DSM-IV or ICD-10

(10th version of the International Statistical Classification of Diseases and Related Health Problems) and were being treated for it. Subjects were excluded if they had a past diagnosis of bipolarity, any history of or ongoing medical conditions, depression caused due to another medical condition, or if they had had electroconvulsive therapy in the preceding month.

The selected patients were subjected to a clinical interview in order to collect their sociodemographic and clinical background, and were tested for bipolar disorder through the Chinese version of the Mini – International Neuropsychiatric Interview (MINI) Version 5.0.

The gathered data was statistically analyzed with SPSS Statistics 13.0, and the sociodemographic and clinical characteristics for major depressive disorder and bipolar disorder were compared by one-way analysis of variance using ANOVA or by a chi-square test. For the characteristics that these tests were significant, further multivariate logistic regression was carried out to compare major depressive disorder and bipolar disorder.

The results of this study show that 309 patients out of the 1487 screened were diagnosed with bipolar disorder, 118 with type I and 191 with type II. Compared to patients with major depressive disorder, BD patients were more likely to be male, have had more depressive episodes, had atypical depression, were suicidal, psychotic, had a history of psychiatric disorders, were younger, and had an early age at the onset.

2.4 Summary

The existing literature on the detection of bipolar disorder and other psychological disorders in individuals can be divided into three broad categories mentioned in Sections 2.1, 2.2, and 2.3. The papers discussed in Section 2.1 take a more traditional approach to diagnose bipolar disorder and focus on the autonomic nervous system symptoms of bipolar disorder, such as heart rate variability, sleep patterns, and electrocardiography. On the other hand, papers in Section 2.2 take a non-traditional approach and analyze speech patterns of subjects in order to diagnose different psychological disorders. This approach can also be used to determine the severity of the psychological disorder in the patients.

The methodology used in this thesis is more closely related to the papers discussed in Section 2.3. Note that these papers use statistical analysis of clinical and sociodemographic data of bipolar patients to determine the predictors of this disease. The previous research in this field has mostly focused on using such factors as family history of psychological disorders, age of onset of the bipolar disorder, presence of psychosis in the patient, number of depressive episodes in a year, HRV, etc. in statistical analysis in order to detect the presence of bipolar disorder. The data used for this thesis lacks much of this information, but is rich in other aspects, such as medication prescribed to the patient, comorbidities, and sociodemographic data – features that have not been fully explored in the existing literature.

The approach used by papers in Section 2.3 has been chosen for analysis in this thesis for the following reasons:

1) The literature discussed in Section 2.1 relies on extensive medical knowledge about the effects of bipolar disorder on ANS to detect this disorder in patients. Our intention is to develop a decision support tool that does not rely solely on input from medical experts, hence this approach is not used in our thesis.
2) The literature in Section 2.2 uses speech pattern analysis in order to detect bipolar disorder, along with other psychological disorders, in patients. The studies are based on the analysis of groups of patients from a specific region or country. This approach is not practical because the accents and speech patterns vary vastly across countries or even regions within countries, making the development of a tool that can correct for dissimilarities in language and accent rather difficult.
3) The methods in Sections 2.1 and 2.2 involve humans in diagnosing the bipolar disorder, resulting in different or contradicting assessments by medical professionals. This contributes to late diagnosis or misdiagnosis of patients with bipolar disorder. The effects of this have been discussed in Chapter 1. We therefore prefer to apply statistical methods to detect the existence of bipolar disorder in patients. Statistical analysis methods that have been discussed in Section 2.3 are objective methods in predicting inferences given a set of independent variables. Thus, we apply two statistical analysis methods and the artificial neural network technique for detecting the bipolar disorder.

CHAPTER III

DATA ACQUISITION AND DESCRIPTION

3.1 Data Acquisition

The data used in this thesis comes from the CERNER Health Facts® database. This is a database containing a large amount of data collected from patients since 2000 in a real-world health environment. This data is collected from patients who visited Cerner and non-Cerner healthcare facilities. The database has more than a hundred different columns with demographic and clinical data of the patients. For the purpose of this thesis, six months' worth of data (January to June 2013) for bipolar and non-bipolar patients between the ages of 15 and 65 was used. Based on a reading of the literature, see for example Inoue et al. (2015), Takeshima et al. (2013), and Xiang et al. (2013), we selected the following variables as those being the most relevant in a bipolar diagnosis. All the variables pertain to patient data – clinical or demographic.

1) *Patient_ID:* A unique identifier used within the Cerner Health Facts Data Warehouse. This is a sequential number created as new persons are introduced to the database.

2) *Encounter_ID:* A sequential number created every time a patient visits a Cerner facility.

3) *Diagnosis_description:* Description of the diagnosis of a patient.

4) *Diagnosis_ID:* The unique identifier attached with each *diagnosis_description.*

5) *Gender:* Gender of the patient.

6) *Marital_status:* Marital status of the patient.

7) *Race:* Race of the patient.

8) *Age:* Age of the patient.

9) *Urban_rural_status:* Binary indicator describing whether the patient is from a rural or urban background.

10) *Census_region:* Indicator describing which of the four census regions (Northeast, Midwest, South, or West) the patient belongs to.

11) *Event_description:* Description of each clinical event encountered by a patient (e.g., systolic and diastolic blood pressure, alcohol use, heart rate, pulse, etc.)

12) *Result_value_number:* Numerical result of the clinical event in *event_description.*

13) *Medication_generic_name:* Generic name of the medication prescribed to the patient.

These were the only variables used for the purpose of this study as the others in the database, e.g., hospital bed size or year in which the data was collected, do not pertain to the diagnosis of any disease, but are associated with socio-economic factors.

3.2 Data Preparation

Each *patient_ID* is associated with a different *diagnosis_description* and/or *event_description* in the database. The initial dataset used for this thesis has patients with *diagnosis_descriptions* associated with the diseases listed below:

1) Acute myocardial infarction
2) Anorexia nervosa
3) Attention deficit hyperactivity disorder (ADHD)
4) Bulimia nervosa
5) Cerebral artery occlusion
6) Depression
7) Diabetes
8) Panic disorder
9) Posttraumatic stress disorder (PTSD)
10) Anxiety

Out of this list, only ADHD, PTSD, depression, and panic disorder were retained because approximately 95% pf the patients had one or more of these diseases. The remaining six diseases were seen in only 5% of the dataset of patients we considered, which included patients with and without bipolar disorder. Hence, acute myocardial infarction, anorexia nervosa, bulimia nervosa, cerebral artery occlusion, diabetes, and anxiety were excluded from the analysis.

Along with the diseases listed above, the following medication was also included in the analysis:

1) Alprazolam
2) Amphetamine dextroamphetamine
3) Aripiprazole
4) Bupropion
5) Chlorpromazine
6) Citalopram
7) Desipramine
8) Duloxetine
9) Escitalopram
10) Fluoxetine’

11) Haloperidol
12) Lithium
13) Methylphenidate
14) Paroxetine
15) Pioglitazone
16) Sertraline
17) Trazodone
18) Valporic acid
19) Venlafaxine

The above medications were included in analysis because each one of them is used to treat one or more psychological or psychosomatic disorders. Thus, even if the medication does not cause bipolar or any other psychological disorder, there might be a strong correlation between the two that is worth exploring. Appendix A lists the names of all the diseases that each of these medications are used in the treatment of.

The next step in processing the data was to convert the raw data into binary or nominal variables that could be used in statistical analyses with ease. Also, as mentioned earlier, the raw data had multiple entries for the same *encounter_IDs* as each *encounter_ID* was associated with multiple medications prescribed to the patient, multiple diagnoses, and different clinical encounter such as heart rate, BMI, pulse, alcohol use in a day, number of tobacco packs per day, etc. Thus, the data was in a format that could not be used in analysis directly. Hence, binary columns were created for each of these factors and the cleaning was done such that there was only one data-point associated with each of the *encounter_IDs.* This was carried out using MATLAB and SAS Enterprise Guide 9.1. Appendix A lists all the columns (i.e., variables) that resulted from this step, along with their descriptions and possible values.

3.2 Data Description

Although the association of the clinical and demographic factors with bipolar disorder will be explored in further detail in the following chapter, it is useful to determine how the data is distributed in order to better understand it. It is also necessary to note that 52.4% of the patients in the entire dataset have been diagnosed with bipolar disorder, while 47.6% do not have the bipolar disorder. The following graphs and charts help explain aspects of the data used for analysis in this thesis.

3.2.1 Distribution of Bipolar Disorder with Comorbidities

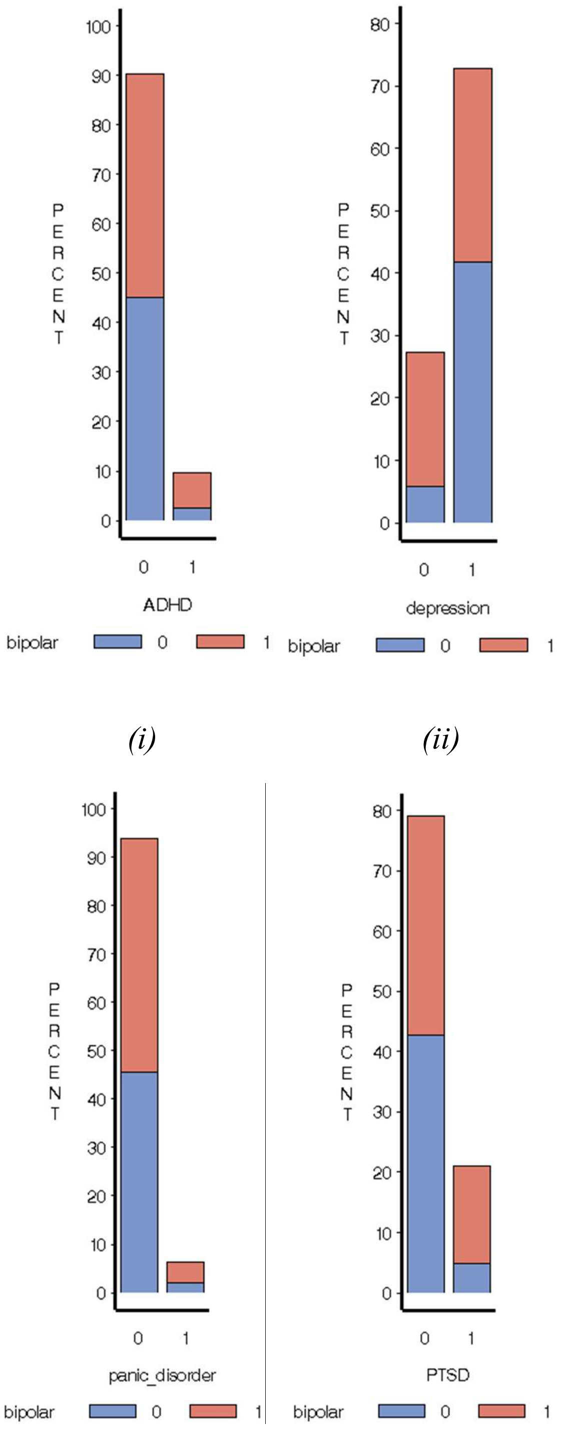
First, examine the distribution of bipolar diagnosis with the distribution of the four comorbidities that have been considered in our analysis – ADHD, depression, panic disorder, and PTSD. The vertical bars in Figure 3.1 (i) indicate that approximately 10% of the patients have been diagnosed with ADHD while the remaining 90% have not. Of these ADHD patients, approximately 70% also have a bipolar diagnosis. On the other hand, out of the 90% non-ADHD patients, there is an almost even distribution of bipolar and non-bipolar patients. Thus, we can see that the split between bipolar and non-bipolar patients in the subpopulation of the dataset with ADHD is unequal, while it is approximately equal in non-ADHD patients, leading us to believe that a comorbidity exists between bipolar disorder and ADHD.

As mentioned in Chapter 1, some of the symptoms for depression and bipolar disorder may be common, but they are two different diseases. However, in practice, due to late diagnosis or misdiagnosis, bipolar patients are often treated for depression, leading to further complications. The data in Figure 3.1 (ii) appears to corroborate these observations. It can be seen that approximately 72% of the patients in the dataset have been diagnosed with depression. Among the depressive patients, the distribution of bipolar and non-bipolar patients is roughly even. On the other hand, out of the remaining subjects who are not depressive, more than half are also bipolar patients.

Figure 3.1 (iii) shows the distribution of panic disorder and bipolar disorder. We can observe that the distribution of bipolar and non-bipolar patients is approximately even between both the subgroups, i.e., patients suffering from panic disorder and patients who do not have panic disorder. This suggests that panic disorder might not be a common comorbidity associated with bipolar disorder.

Lastly, Figure 3.1 (iv) illustrates the distribution of PTSD with bipolar disorder, and we can see that the number of patients suffering from PTSD as well as bipolar disorder is significantly higher than those suffering only from PTSD. On the other hand, the number of bipolar and non-bipolar patients is nearly equal amongst patients without PTSD. This skewed distribution of bipolarity among PTSD patients might be an indicator of a stronger correlation between these two psychological disorders.

From the above observations, it can be seen that ADHD, depression, and PTSD might have a correlation with the bipolar diagnosis of a patient. It can also be seen that whereas ADHD and PTSD might be common comorbidities occurring with bipolar disorder, it might be the exact opposite for depression. These claims cannot be conclusively made until further statistical analysis is carried out, which will be done in the next chapter.



Legend

0: Negative diagnosis

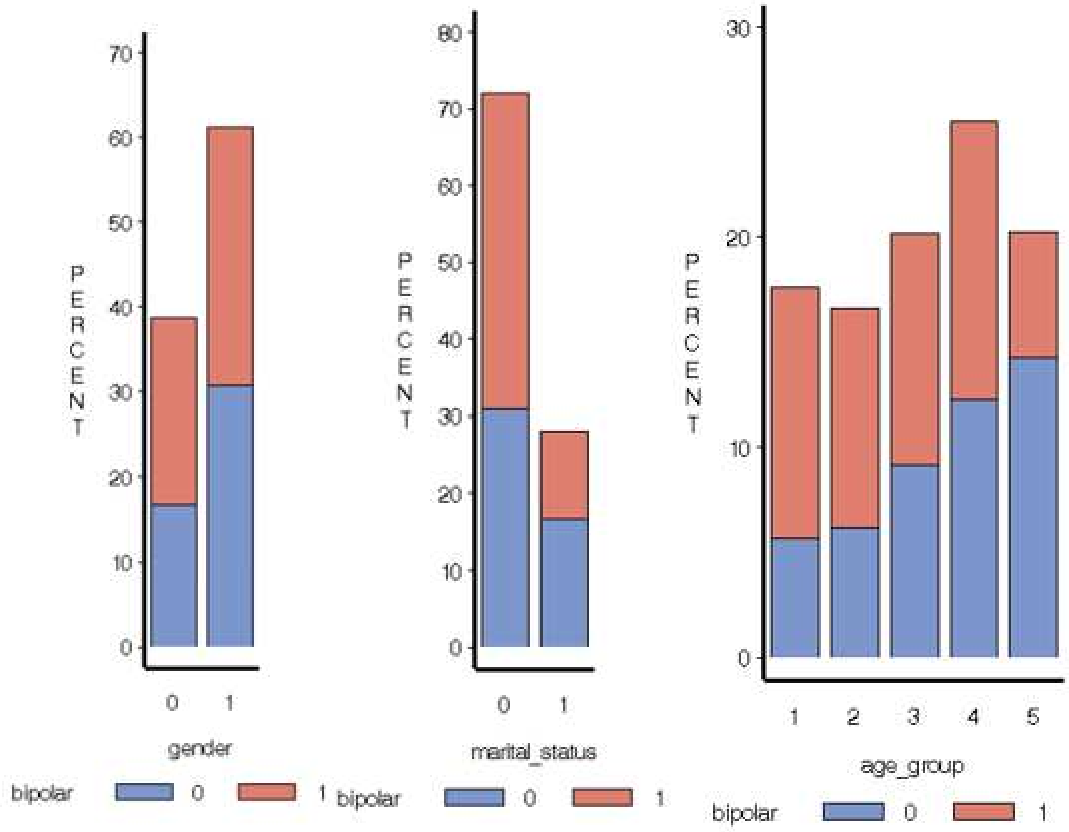
1: Positive diagnosis

(i) *(ii)*

(iii) *(iv)*

Figure 3.1 (From clockwise) Distribution of bipolar disorder with (i) ADHD, (ii) depression, (iii) panic_disorder, and (iv) PTSD

3.2.2 Distribution of Bipolar Disorder with Demographic Factors



Legend

0: Male	0: Single	1: 15-25, 2: 26-35, 3: 36-45
1: Female	1: Not single	4: 46-55, 5: 56-65
(i)	*(ii)*	*(iii)*

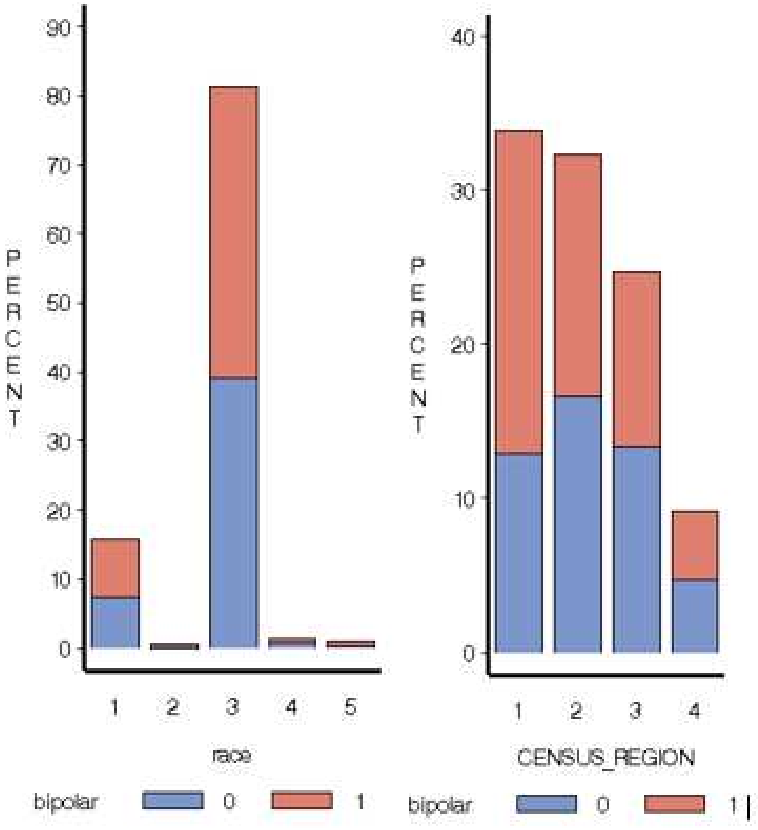
Figure 3.2 (From left) Distribution of bipolar disorder with (i) gender, (ii) marital_status, and (iii) age_group

Figure 3.2 shows the distribution of bipolar disorder with gender and marital status of the patients. It can be seen from Figure 3.2 (i) that the percentage of males (*gender* = 0) and females (*females* = 1) in the dataset is approximately 40% and 60%, respectively. It can also be observed that the number of bipolar and non-bipolar patients in both the genders is approximately the same, suggesting that there is no correlation between a person's gender and the existence of bipolar disorder.

Figure 3.2 (ii) shows the distribution of single and non-single (patients living with a partner or family) subjects in the dataset. Even here, it can be observed that the number of bipolar and non-bipolar patients is roughly equal for both sets of patients. This suggests that there is no correlation between a person's marital status and bipolar diagnosis.

Lastly, Figure 3.2 (iii) shows the distribution of bipolar disorder with age group, where *age_group* 1 represents patients 15-25 years of age, *age_group* 2 represents patients of 26-35 years age, and so on. It can be seen that the ratio of the percentages of bipolar to non-bipolar patients is approximately decreasing with age group, with the number of bipolar patients being higher in *age_group_1* and lower in *age_group_6*. This might be due to the fact that the age of onset of bipolar disorder is usually around puberty (Perlis et al., 2009). The lower percentage of bipolar disorder in higher age groups can be due to treatments sought for it at an early stage.

The literature that was studied for the purpose of this thesis does not mention a correlation between the race of a patient and bipolar diagnosis, but we chose to include this factor in the analysis to see whether it was relevant. As can be seen in Figure 3.3 (i), the maximum number of patients in the dataset Caucasian (*race* = 3), followed by African American (*race* = 1), Hispanic (*race* = 4), Native American (*race* = 5), and Asian (*race* = 2), in that order. This highly skewed distribution can be explained by the fact that the data used here was collected across USA and the data reflects the demographical variations. The number of bipolar and non-bipolar patients is almost equal amongst all the races, which suggests that there might not be a correlation between the race of a person and whether or not he or she will have bipolar disorder. This will be further explored in statistical analysis.



Legend

1: African American, 2: Asian,
3: Caucasian, 4: Hispanic,
5: Native American

(i)

1: Midwest, 2: Northeast,
3: South, 4: West

(ii)

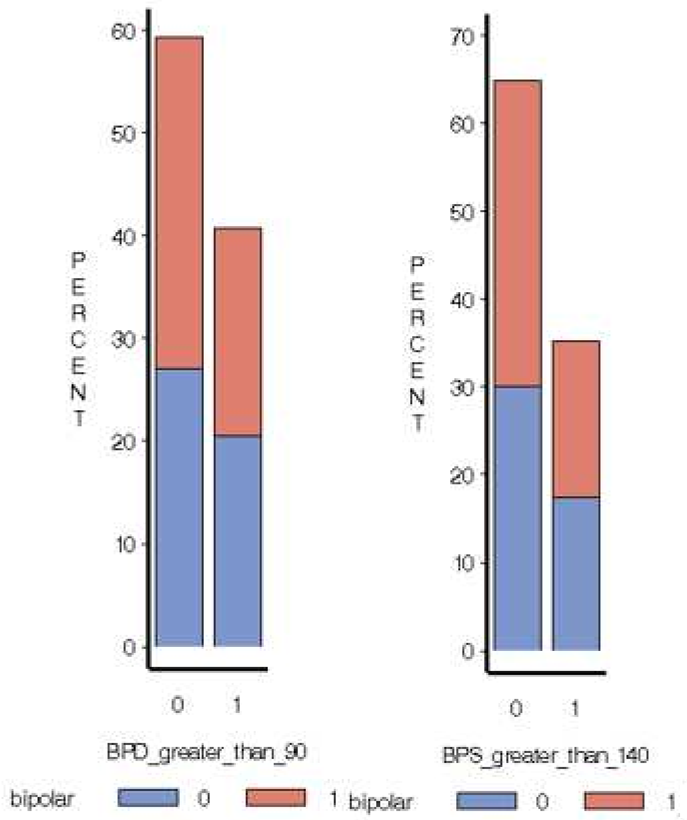
Figure 3.3 (From left) Distribution of bipolar disorder with (i) race and (ii) census_region

The decision to include census region that the patients belong to in the analysis was taken for the same reason as race. Figure 3.3 (ii) indicates that there might not be a correlation between the census region and a bipolar disorder, and that the distribution is similar over all four regions.

3.2.3 Bipolar Disorder and Manifestations of ANS Symptoms

As can be seen from the previous literature discussed in Section 2.1, the bipolar disorder affects the autonomous nervous system of a patient's body, which controls the involuntary processes in the human body, such as heartbeat and respiration. The effects of this manifest themselves through abnormalities in heart rate variability, respiratory, pulse, etc. Thus, it is necessary to consider these factors in the analysis as well. In this section, we examine the distributions of the indicators of abnormalities in systolic and diastolic blood pressure, and respiratory rate of patients with the distribution of bipolar disorder.

Figure 3.4 (i) and (ii) show the distributions of *BPD_greater_than_90* and *BPS_greater_than_140* with bipolar disorder distribution. A diastolic reading above 90 mmHg or a systolic reading above 140 mmHg are considered to be indicators of abnormally high blood pressure, also known as hypertension. From the figures, we can see that the number of patients with high diastolic and systolic blood pressure are almost equal. Also, the number of roughly equal within all four subgroups. This is suggestive of hypertension and bipolar disorder not being comorbidities.

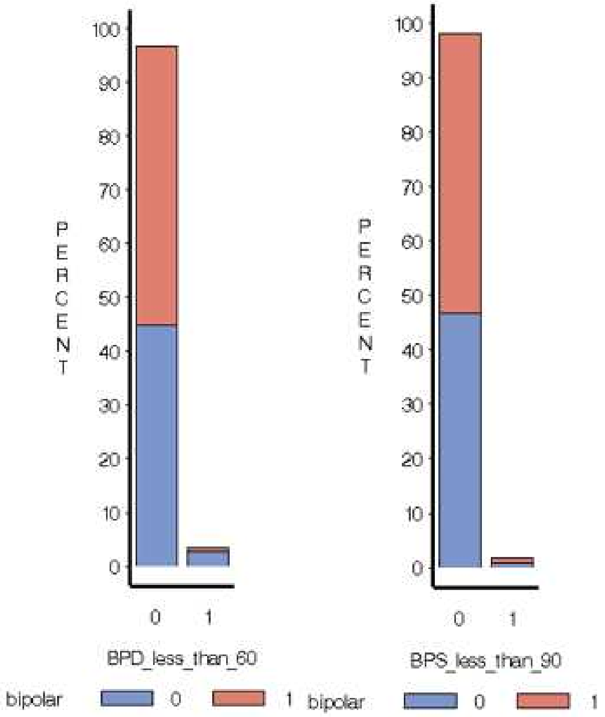


Legend

0: Diastolic BP < 90	0: Systolic BP < 140
1: Diastolic BP > 90	1: Systolic BP > 140
(i)	*(ii)*

Figure 3.4 (From left) Distribution of bipolar disorder with (i) BPD_greater_than_90 and (ii) BPS_greater_than_140

Similar to hypertension, it is also important to look at the distribution of hypotension (blood pressures below 60/90 mmHg) with that of bipolar disorder. This is done by examining the distributions of indicators *BPD_less_than_60* and *BPS_less_than_90*, shown in Figure 3.5 (i) and (ii) respectively. It can be seen from both the figures that the number of patients with abnormally low systolic and diastolic blood pressures is approximately less than 5%. Also, we can observe that the number of bipolar and non-bipolar patients is equal for patients whose blood pressures are not abnormally low. Observe that, among the patients with abnormally low systolic blood pressure, the number of bipolar and non-bipolar patients is equal as well. On the other hand, among patients with abnormally low diastolic blood pressure, the number of non-bipolar patients is much higher than those with this disease. This indicates that there might be a correlation between low diastolic blood pressure and bipolar disorder, which will be explored in detail in the next chapter.

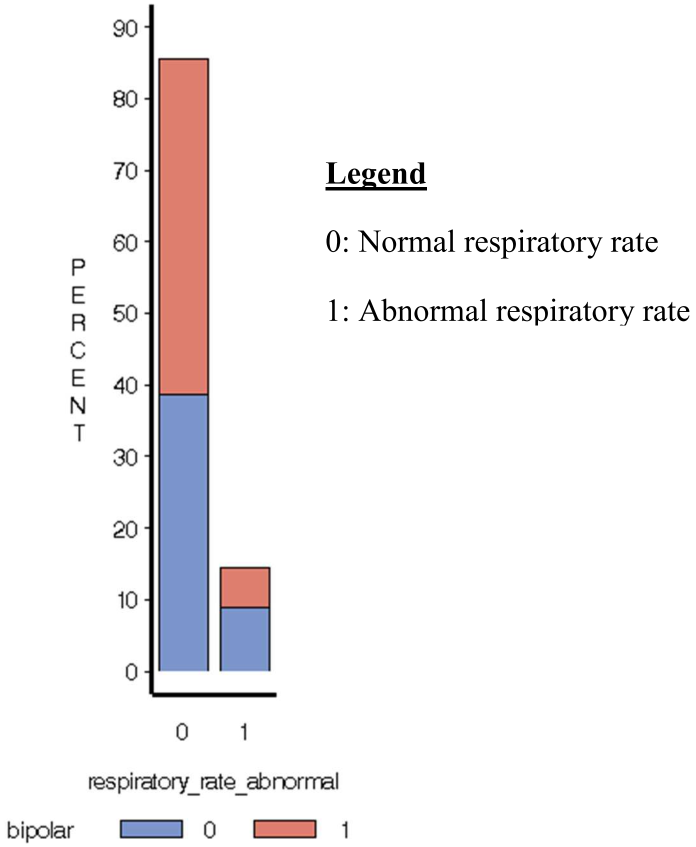


Legend

0: Diastolic BP > 60	0: Systolic BP > 90
1: Diastolic BP < 60	1: Systolic BP < 90
(i)	*(ii)*

Figure 3.5 (From left) Distribution of bipolar disorder with (i) BPD_less_than_60 and (ii) BPS_less_than_90

Lastly, let us examine the distribution of the indicator for abnormal respiratory rate with the distribution of bipolar disorder in Figure 3.6. It can be seen that approximately 15% of the patients in the dataset have an abnormally high or low respiratory rate. Also, the number of bipolar and non-bipolar patients is not equal for either subgroup. The number of bipolar patients is higher amongst patients with normal respiratory rate, while it is lower for the other group. This suggests that bipolar patients might be more likely to have a normal respiratory rate, which will be analyzed in further detail in the next chapter.



Legend

0: Normal respiratory rate

1: Abnormal respiratory rate

Figure 3.6 Distribution of bipolar disorder with respiratory_rate_abnormal

3.2.4 Distribution of Bipolar Disorder with Prescribed Medication

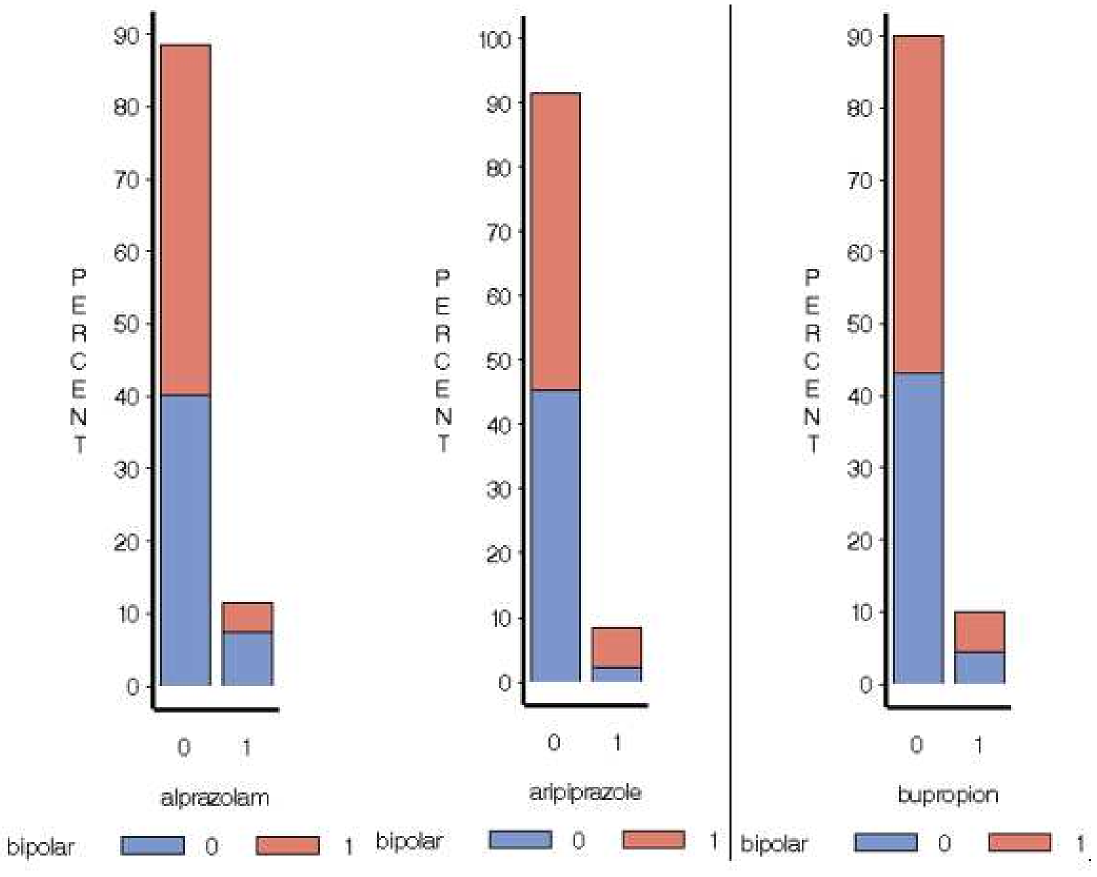
In this section, we will examine the distribution of the bipolar disorder among patients taking specific prescribed medicines. As discussed in Section 3.1, apart from four disorders (ADHD, depression, panic disorder, and PTSD) all the other diseases had to be excluded because they were a very small percentage of the dataset. This is why we believe it is worth looking at the medication prescribed to the patients. Most of the medications considered here are used in the treatment of multiple psychiatric and psychosomatic diseases (Appendix A), including the ones that had to be dropped off while cleaning and preparing the data for analysis. Thus, the prescriptions patterns, in conjunction with bipolar diagnosis, may provide valuable insights into whether or not the diseases treated by them have a correlation with the bipolar disorder.

We can see in Figure 3.7 (i) that the number of bipolar patients who have been prescribed alprazolam, which is used to treat anxiety and panic disorder, is lower than the number of non-bipolar patients taking this drug. On the other hand, the numbers are almost equal amongst patients not taking this drug. This might be an indication of anxiety and panic disorder not being usual comorbidities associated with the bipolar disorder. We also saw in Figure 3.1 (iii) that there did not seem to be any significant skew in the split between patients amongst those with the panic disorder. Thus, this suggests that a patient with anxiety probably has a higher likelihood of being bipolar.

Legend

0: Medication not prescribed

1: Medication prescribed



(i) *(ii)* *(iii)*

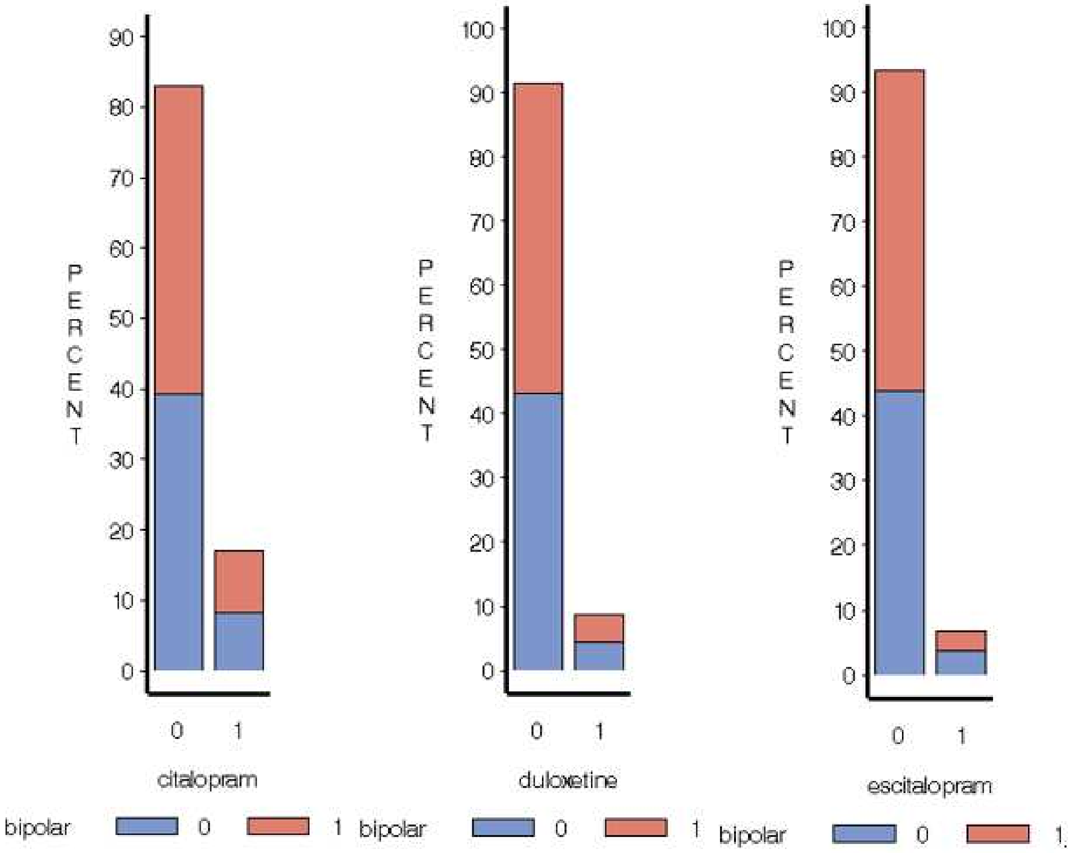
Figure 3.7 (From left) Distribution of bipolar disorder with (i) alprazolam, (ii) aripiprazole, and (iii) bupropion

Figure 3.7 (ii) shows the distribution of patients taking aripiprazole, used to treat schizophrenia and Tourette syndrome, with bipolar disorder. Here, even though the number of bipolar and non-patients is approximately equal amongst patients not taking this medication, the number of bipolar patients is higher amongst those who are taking it. Tourette syndrome is an inherited neuropsychiatric disorder while schizophrenia is a psychological disorder, and this graph suggests that there might be a comorbidity relationship between these two diseases and bipolar disorder.

Legend

0: Medication not prescribed

1: Medication prescribed



(i) *(ii)* *(iii)*

Figure 3.8 (From left) Distribution of bipolar disorder with (i) citalopram, (ii) duloxetine, and (iii) escitalopram

The distribution of bupropion, which is used to treat depression and to help quit smoking, is shown in Figure 3.7 (iii). This graph suggests that there is no correlation between bipolar disorder and this drug, as the number of bipolar and non-bipolar patients amongst the patients taking this medication, as well as not taking it, is equal. This can be due to the fact that bupropion is used in the treatment of depression, which, as we saw in Figure 3.1 (ii), is not a comorbidity that usually occurs with bipolar disorder.

Figure 3.8 (i) shows the distribution of the bipolar disorder in patients who have been prescribed citalopram, a drug used to treat depression. In this figure as well, as in Figure 3.7 (iii), we can see that the number of bipolar and non-bipolar patients is equal amongst the patients taking this drug, as well as not taking it. Because citalopram is also prescribed for depression, this further reinforces the assumption that it might be an indicator of the relationship between depression and bipolar disorder. A similar trend can be seen in Figure 3.8 (ii) and (iii), which shows the distribution of the bipolar disorder with the medications duloxetine and escitalopram respectively, both of which are used to treat depression and anxiety. The trend continues in Figure 3.9 (i), which shows the distribution of bipolar disorder with fluoxetine, which is used to treat depression, obsessive-compulsive disorder (OCD), bulimia nervosa, and panic disorder.

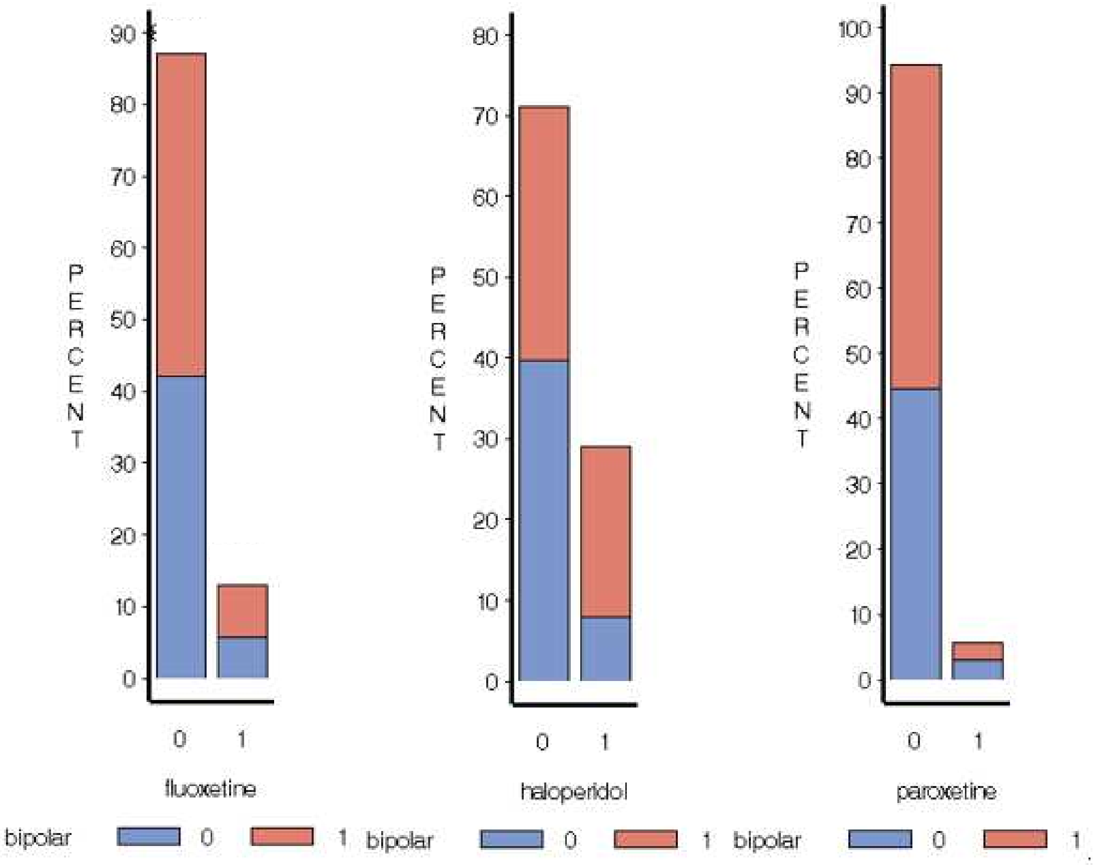
Figure 3.9 (ii) illustrates the distribution of bipolarity and the drug haloperidol, used to treat schizophrenia, Tourette syndrome, mania, nausea and vomiting, delirium, agitation, psychosis, and hallucinations. We can see in the graph that although the number of non-bipolar patients taking this medication is slightly higher than the bipolar patients, the number of bipolar patients is higher amongst the subset of patients taking it. This observation reinforces what we saw in Figure 3.5 (ii) with respect to aripiprazole, that there might be a comorbidity relationship between schizophrenia, Tourette syndrome, and bipolar disorder.

It can be observed that the drugs used in the treatment of depression have no correlation with bipolar disorder, highlighting further that these are two different diseases that might not have comorbidities.

Legend

0: Medication not prescribed

1: Medication prescribed



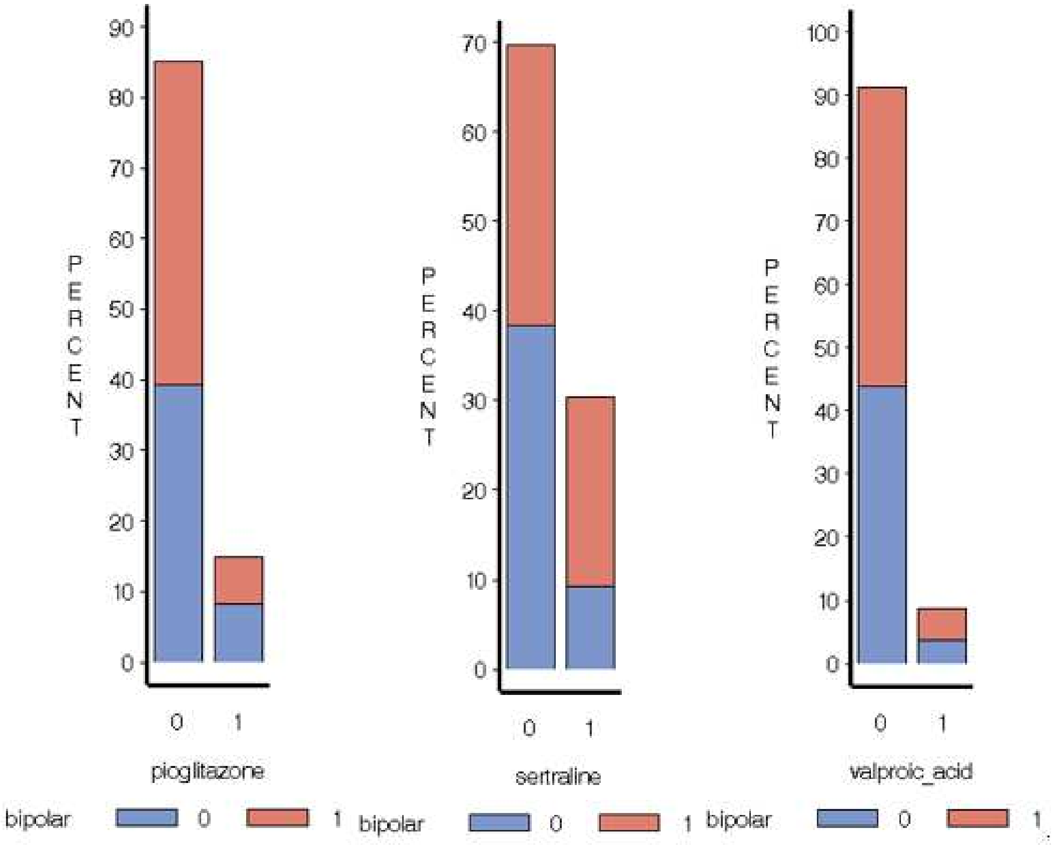
(i) *(ii)* *(iii)*

Figure 3.9 (From left) Distribution of bipolar disorder with (i) fluoxetine, (ii) haloperidol, and (iii) paroxetine

Legend

0: Medication not prescribed

1: Medication prescribed



(i) *(ii)* *(iii)*

Figure 3.10 (From left) Distribution of bipolar disorder with (i) pioglitazone, (ii) sertraline, and (iii) valporic_acid

Figure 3.9 (iii) shows the distribution of bipolar disorder with paroxetine, which is used to treat depression, anxiety, obsessive-compulsive disorder, and premenstrual dysphoric disorder. As has previously been seen in case of bupropion, citalopram, duloxetine, escitalopram, and fluoxetine, all of which are used to treat depression and anxiety, the number of bipolar and non-bipolar patients amongst the subgroups of patients taking or not taking this medication is equal, further reinforcing that depression and anxiety are not common comorbidities of bipolar disorder.

The distribution of bipolar disorder and pioglitazone is shown in Figure 3.10 (i). Pioglitazone is used in the treatment of diabetes mellitus, which is a metabolic disease where the patients suffers from high blood sugar levels for extended periods of time. Because of the nature of this disease, one would naturally assume that there would not exist any correlation between it and bipolar disorder. From Figure 3.10 (i), it can be seen that this assumption might be correct as the number of bipolar and non-bipolar patients is roughly equal amongst both the subgroups.

Figure 3.10 (ii) shows the distribution of sertraline, which is used to treat obsessive-compulsive disorder, PTSD, premenstrual dysphoric disorder, anxiety, and panic disorder. Here, we can see that while the number of bipolar and non-bipolar patients is almost equal for subjects not taking this medication, the number of bipolar patients is higher amongst those who are taking this medication. This corroborates our observations in Figure 3.1 (iv), which showed that PTSD might be a commonly occurring comorbidity with bipolar disorder.

Lastly, observe in Figure 3.10 (iii) that amongst the patients who are not prescribed valporic acid, the number of bipolar and non-bipolar subjects is approximately equal, suggesting the lack of a correlation between bipolar disorder and seizures and migraines, the diseases treated with this medication.

It can observed from the discussion in this section that the distributions for sertraline, haloperidol, alprazolam, and aripiprazole are the most skewed for bipolar and non-bipolar patients. This suggests that there might be a statistically significant correlation between the prescription of these drugs and a bipolar diagnosis. By effect, there might be a correlation between the diseases these drugs are used to treat and bipolar disorder. Whether there is a statistically significant association or not will be explored further in the following chapter through statistical analysis.

3.3 Summary

Table 3.1 below summarizes the first impression about the set of independent variables that may and may not have a correlation with the bipolar disorder based on the analysis in Section 3.2:

Correlation with *bipolar*	**No correlation with *bipolar***
ADHD	Panic_disorder
Depression	Gender
PTSD	Race
Marital_status	Census_region
Age_group	BPD_greater_than_90
BPD_less_than_60	BPS_greater_140
Respiratory_rate_abnormal	BPS_less_than_90
Alprazolam	Bupropion
Aripiprazole	Citalopram
Haloperidol	Duloxetine
Sertraline	Escitalopram
	Fluoxetine
	Paroxetine
	Pioglitazone
	Valporic_acid

Table 3.1 Summary of correlation based on an initial analysis

CHAPTER IV

METHODOLOGY AND RESULTS

4.1 Methodology

For analyzing the data used in this thesis with different statistical analysis tools, a dependent variable, *bipolar*, that is categorical in nature with two possible responses, (0 = no bipolar diagnosis and 1 = positive bipolar diagnosis), is used. The software used for this purpose is SAS® Enterprise Miner Workstation 14.1 for 64-bit Windows computer. The analysis is carried out in three steps:

1) In the first step, the data is divided into three sets – *training* (40%), *validation* (30%), and *test* (30%) – and the four models are built on the *training* dataset.
2) In the second step, multiple logistic regression, decision tree, and AutoNeural network models are built on the *training* dataset for the dependent variable *bipolar* as a function of all the independent variables described in the previous chapter. The details of each of these models are as follows:
   a) The logistic regression model is built using the stepwise selection method. In this method, independent variables are added to the model one by one and retained if they satisfy the user-specified significance level for the Wald chi-square tests that tests the hypothesis that at least one of the predictor's regression coefficient is not equal to zero (i.e., the predictor had a significant

effect on the target variable). If a variable added in a previous step because it was found to be significant is not significant in the current step, it is dropped from the regression model. The independent variable is retained if its Wald chi-square *p*-value is lower than the significance level. The value of significance level for this purpose is taken as 0.05.

b) The value of significance level for decision trees is set at 0.05. The selection criterion for the model is the misclassification rate. That is, the decision tree with the lowest misclassification rate is selected.

c) The AutoNeural network has a block layer architecture with four hidden layers, i.e., each of the four hidden layers in the network has the same number of neurons and they create a "block" pattern. Termination criterion for the model training is overfitting, i.e., the training of the network will stop when overfitting is detected by the software. Further, the target layer error function is multiple Bernoulli.

3) All three models are scored on the *validation* and *test* datasets and compared based on their *test* ROC curves and ROC index (i.e., area under the ROC curve) in the third and last step.

For a detailed explanation of the three statistical analysis tools used in this thesis, refer to Appendix B. The following sections will describe the results of the four models and compare them with each other.

4.2 Logistic Regression

As discussed in section 4.1, the first step in this study was to build a multiple logistic regression model on the *training* dataset with stepwise selection for the target dependent variable *bipolar* in order to determine the set of independent variables that constitute the best fitting model. The significance level for building this model was set at 0.05. The order in which the effects were added in the logistic regression model by SAS using the forward selection method is provided in Table 4.1.

Number In	**Effect Entered**	**Score Chi-Square**	***P* value**
1	Depression	98.3045	< 0.0001
2	Haloperidol	62.4981	< 0.0001
3	Sertraline	30.2860	< 0.0001
4	Aripiprazole	14.7292	0.0001
5	BPD_less_than_60	10.1702	0.0014
6	PTSD	9.6909	0.0019
7	Lithium	7.7270	0.0054
8	Respiratory_rate_abnormal	7.1352	0.0076
9	Marital_status	4.2051	0.0403

Table 4.1 Order of variable selection for the logistic regression model

Thus, the selected model with the best fit had *depression, haloperidol, sertraline, aripiprazole, BPD_less_than_60, PTSD, lithium, respiratory_rate_abnormal,* and *marital_status* as the effective predictor variables. The parameter estimates for these ten significant variables are as listed in Table 4.2 below:

Effect Number	Parameter	Estimate	Standard Error
1	Intercept	0.9122	0.3631
2	BPD_less_than_60	0.7560	0.2542
3	PTSD	-0.3647	0.1158
4	Aripiprazole	-0.4234	0.1311
5	Depression	0.6145	0.1038
6	Haloperidol	-0.5450	0.0840
7	Lithium	-0.5178	0.1949
8	Marital_status	0.1664	0.0813
9	Respiratory_rate_abnormal	0.2859	0.1122
10	Sertraline	-0.3823	0.0816

Table 4.2 Parameter estimates for the logistic regression model

Thus, the regression equation for *bipolar* can be written as follows:

$$logit(p) = ln(odds\ of\ a\ bipolar\ diagnosis) = \ ln\left(\frac{p}{1-p}\right)$$

$$= 0.9122 + 0.7560 * BPD_{less_{than_{60}}} - 0.3647 * PTSD$$

$$- 0.4234 * Aripiprazole + 0.6145 * Depression$$

$$- 0.5450 * Haloperidol - 0.5178 * Lithium$$

$$+ 0.1664 * Marital\_status + 0.2859 * Respiratory_{rate_{abnormal}}$$

$$- 0.3823 * Sertraline$$

Where, p = probability of a positive bipolar diagnosis

Thus, the probability of a person having bipolar disorder can be easily calculated from this equation. A medical professional using this will only have to enter the values of the independent variables in the equation and patients can be referred to a psychiatrist if their probability of having bipolar disorder is greater than a discrimination threshold, which is set at a default of 0.5 and varied by SAS Enterprise Miner® to find the near optimal value for which the performance of the model is the highest.

Figure 4.1 below shows the effects plot for the significant variables that graphically represents the importance of the independent variables and whether or not the point estimate for it is negative or positive.

The odds ratio estimates for the individual parameters in the logistic regression model that was selected by the software are as shown in Table 4.3.

Parameter		**Point Estimate**
BPD_less_than_60	0 vs 1	4.536
PTSD	0 vs 1	0.482
Aripiprazole	0 vs 1	0.429
Depression	0 vs 1	3.418
Haloperidol	0 vs 1	0.336
Lithium	0 vs 1	0.355
Marital_status	0 vs 1	1.395
Respiratory_rate_abnormal	0 vs 1	1.772
Sertraline	0 vs 1	0.466

Table 4.3 Odds ratio estimates for logistic regression model

Odds ratio is the ratio of the odds of a patient having bipolar disorder when the independent variable value is 0 to the odds when the independent variable value is 1. Mathematically this can be represented as follows:

$$Odds\ ratio = \frac{Odds\ (Patient\ is\ bipolar)_{Independent\ variable=0}}{Odds\ (Patient\ is\ bipolar)_{Independent\ variable=1}}$$

$$= \frac{p_{independent\ variable=0} \big/ 1 - p_{independent\ variable=0}}{p_{independent\ variable=1} \big/ 1 - p_{independent\ variable=1}}$$

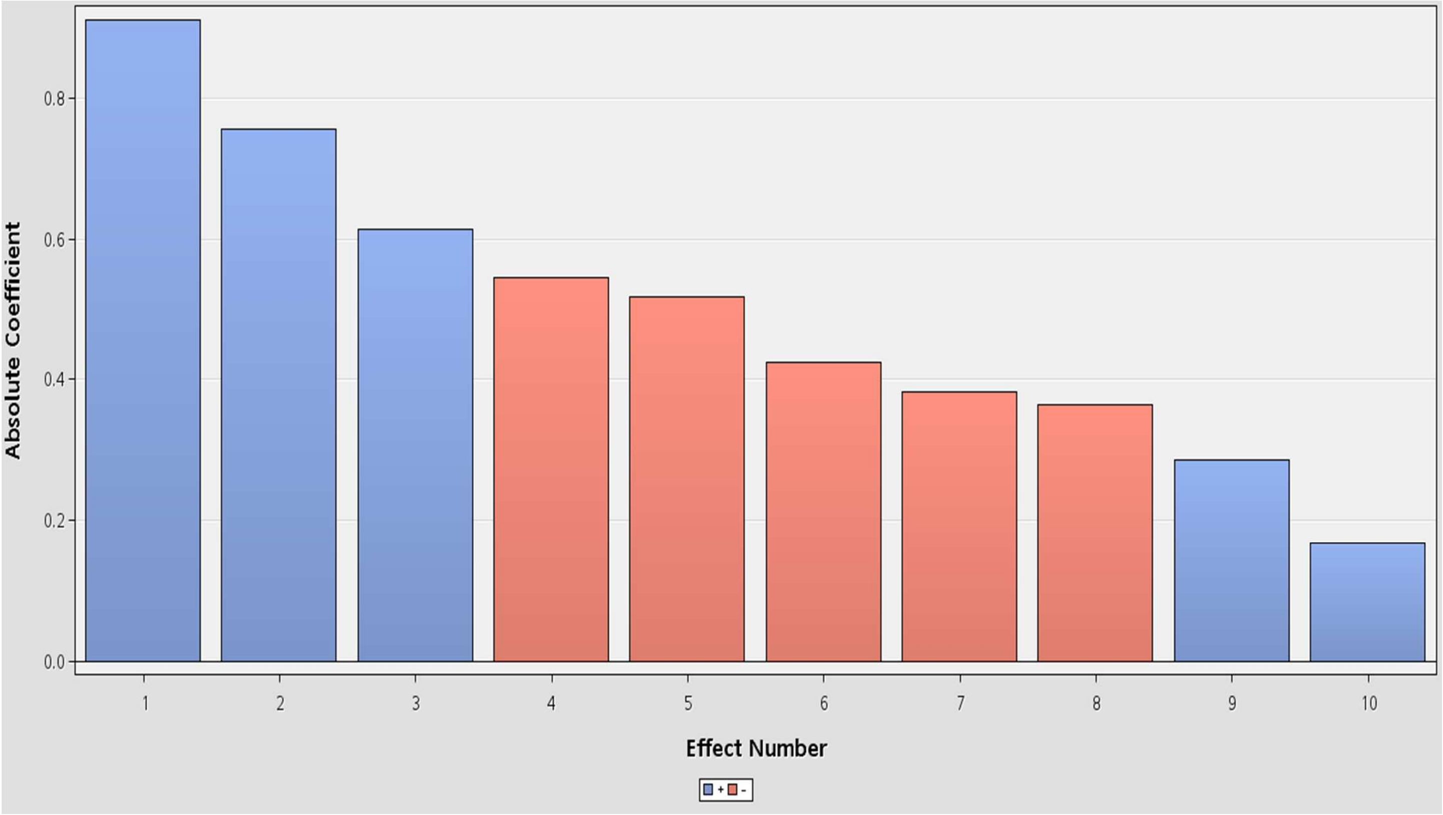


Figure 4.1 Effects plot for significant variables in the logistic regression model

We can see from Table 4.3 that for the independent variables *PTSD, aripiprazole, haloperidol, lithium,* and *sertraline,* the odds ratio value is lower than 1, while it is the opposite for *BPD_less_than_60, depression, marital_status,* and *respiratory_rate_abnormal.* Thus, from the odds ratio formula, it can be seen that the odds of patient having a bipolar diagnosis are higher when the person has PTSD, and has been prescribed aripiprazole, haloperidol, lithium, and sertraline. On the other hand, the odds of the patient being bipolar are lower when he or she has an abnormally low diastolic blood pressure, an abnormally high or low respiratory rate, has depression, and is single. In other words, the probability that the person will have bipolar disorder is higher when they also have PTSD and are taking aripiprazole, haloperidol, lithium, and sertraline, while it is lower when they have an abnormally low diastolic blood pressure, an abnormal respiratory rate, are single, and are depressed.

This model was scored on the *validation* dataset in the next step, using p = 0.5 as the threshold for decision making. That is, if the probability is lower than 0.5 the patient will be categorized as non-bipolar, while he or she will be categorized as bipolar if it is over 0.5. The event classification table for the *training* and *validation* datasets is as follows:

Dataset	**False Negative**	**True Negative**	**False Positive**	**True Positive**
Training	106	315	161	419
Validation	84	241	117	309

Table 4.4 Event classification table for logistic regression model

Thus, the sensitivity and specificity of the model for *training* and *validation* datasets is as follows:

$$Sensitivity_{Training} = \frac{True\ positive}{True\ positive + False\ negative} = \frac{419}{419 + 106} = 0.7981$$

$$Sensitivity_{Validation} = \frac{True\ positive}{True\ positive + False\ negative} = \frac{309}{309 + 84} = 0.7862$$

$$Specificity_{Training} = \frac{True\ negative}{True\ negative + False\ positive} = \frac{315}{315 + 161} = 0.6617$$

$$Specificity_{Validation} = \frac{True\ negative}{True\ negative + False\ positive} = \frac{241}{241 + 117} = 0.6732$$

Sensitivity is a measure of the test's ability to correctly detect patients who have bipolar disorder, while specificity is a measure of the ability to detect patients without a bipolar diagnosis. Thus, the model could correctly detect 79.81% bipolar patients in the training dataset and 76.82% in the validation dataset. Also, the model could detect 66.17% non-bipolar patients in the training dataset and 67.32% in the validation dataset. It can be seen that the model is better at detecting true positives than it is at detecting true negatives.

Table 4.5 below shows the fit statistics for the logistic regression model on the three datasets. It can be seen from this table that the model performance is consistent on all three datasets.

Fit Statistics	Train	Validation	Test
Average squared error	0.19	0.18	0.19
Average error function	0.56	0.56	0.56
Maximum absolute error	0.96	0.97	0.98
Mean square error	0.19	0.18	0.19
Root average sum of squares	0.44	0.43	0.43
Root mean squared error	0.44	0.43	0.43
Misclassification rate	0.27	0.27	0.27

Table 4.5 Fit statistics for logistic regression model

4.3 Decision trees

A decision tree model was built for the *training* dataset and scored on the *validation* datasets in the second step of statistical analysis. The resulting decision tree is shown in Figures 4.2, 4.3, and 4.3. The tree is broken into three parts as it does not fit on one page.

The importance of the individual variables in the decision tree model is as listed in Table 4.6 below. We can see from this table that *depression* is the most important variable in classifying the bipolar patients in the resulting decision tree model. In other words, the independent variable *depression* is the most effective classifier to divide the bipolar and non-bipolar patients into two different groups. This variable is followed by the other variables in the order listed in Table 4.6.

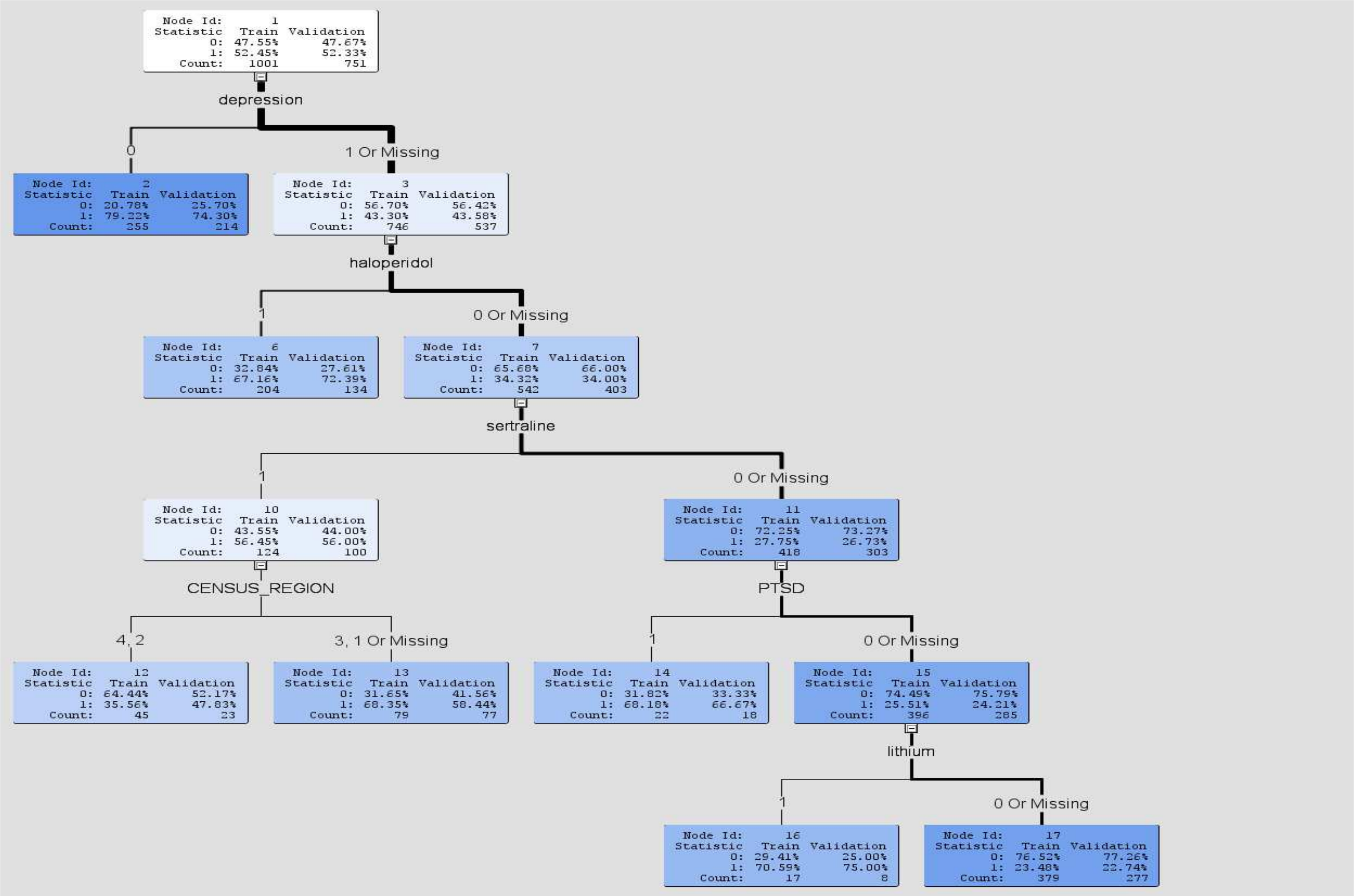


Figure 4.2 Decision tree

Variable	Training Importance	Validation Importance	Ratio of Validation to Training Importance
Depression	1.0000	0.9814	0.9814
Haloperidol	0.8074	1.0000	1.2385
Sertraline	0.5668	0.6659	1.1748
PTSD	0.3935	0.4579	1.1638
Lithium	0.3837	0.3848	1.0027
Census_region	0.3547	0.0000	0.0000

Table 4.6 Variable importance for decision tree model

The event classification table for the model on the *training* and *validation* datasets is as shown in Table 4.7 below:

Dataset	False Negative	True Negative	False Positive	True Positive
Training	105	319	157	420
Validation	74	226	132	319

Table 4.7 Event classification table for decision tree model

Thus, the sensitivity and specificity of the model for *training* and *validation* datasets can be calculated as follows:

$$Sensitivity_{Training} = \frac{True\ positive}{True\ positive + False\ negative} = \frac{420}{420 + 105} = 0.8000$$

$$Sensitivity_{Validation} = \frac{True\ positive}{True\ positive + False\ negative} = \frac{319}{319 + 74} = 0.8117$$

$$Specificity_{Training} = \frac{True\ negative}{True\ negative + False\ positive} = \frac{319}{319 + 157} = 0.6701$$

$$Specificity_{Validation} = \frac{True\ negative}{True\ negative + False\ positive} = \frac{226}{226 + 132} = 0.6313$$

Thus, to interpret this, the decision tree model could correctly identify 80% bipolar patients in the *training* dataset and 81.17% in the *validation* dataset. Also, the model could detect 67.01% non-bipolar patients in the training dataset and 63.13% in the validation dataset. Thus, we can conclude that even though the ability of the decision tree that was constructed to correctly detect bipolar patients is higher than that of logistic regression models, it is not significantly different than logistic regression when it comes to correctly identifying true negatives.

Table 4.8 below shows the fit statistics for the decision tree model on the three datasets. It can be seen from this table that the model performance is consistent on all three datasets.

Fit Statistics	**Train**	**Validation**	**Test**
Average squared error	0.19	0.20	0.19
Maximum absolute error	0.79	0.79	0.79
Root average squared error	0.44	0.44	0.44
Misclassification rate	0.26	0.27	0.27

Table 4.8 Fit statistics for decision tree model

4.4 AutoNeural Network

An AutoNeural network model was built for the *training* dataset and scored on the *validation* datasets in the third step. The event classification table for the model is as shown in Table 4.9 below:

Dataset	**False Negative**	**True Negative**	**False Positive**	**True Positive**
Training	89	352	124	436
Validation	50	138	220	343

Table 4.9 Event classification table for AutoNeural Network model

Thus, the sensitivity and specificity of the model for *training* and *validation* datasets is calculated as follows:

$$Sensitivity_{Training} = \frac{True\ positive}{True\ positive + False\ negative} = \frac{436}{436 + 89} = 0.8305$$

$$Sensitivity_{Validation} = \frac{True\ positive}{True\ positive + False\ negative} = \frac{343}{343 + 50} = 0.8728$$

$$Specificity_{Training} = \frac{True\ negative}{True\ negative + False\ positive} = \frac{352}{352 + 124} = 0.7395$$

$$Specificity_{Validation} = \frac{True\ negative}{True\ negative + False\ positive} = \frac{138}{138 + 220} = 0.3855$$

Thus, the AutoNeural network model could correctly detect 83.05% bipolar patients in the *training* dataset and 87.28% in the *validation* dataset. Also, the model could detect 73.95% non-bipolar patients in the training dataset and 38.55% in the validation dataset. We can see that the model is improving at correctly identifying bipolar patients as it learns, but its ability to detect non-bipolar patients reduces drastically at the same time. Thus, even though AutoNeural networks are better than both logistic regression and decision trees in terms of sensitivity, it fails to perform nearly as well in terms of specificity and also consistency. This can be explained by the limited number of data points used in the analysis. AutoNeural networks perform well when the dataset that they are built and scored on are significantly large.

Table 4.10 below shows the fit statistics for the AutoNeural network model on the three datasets. It can be seen from this table that the model performance is consistent on all three datasets, except for the mean squared error and root mean squared error, which are dropping dramatically for *validation* and *test* datasets. Also, the misclassification rate is more for *validation* and *test* datasets, which means that the number of wrong predictions are increasing with learning. This can also be explained by the small number of data points over which the model was built.

Fit Statistics	**Train**	**Validation**	**Test**
Average square error	0.14	0.22	0.23
Maximum absolute error	0.96	1.00	1.00
Root average squared error	0.38	0.47	0.48
Average error function	0.43	0.66	0.68
Misclassification rate	0.21	0.36	0.38

Table 4.10 Fit statistics for AutoNeural Network model

This inconsistent behavior in terms of the misclassifications can also be observed in Figures 4.5 and 4.6 below. Here, Figure 4.5 shows the change in the value of misclassification rate with the training step, while Figure 4.6 shows the actual number of wrong classifications with training step. It can be seen that even though both the plots are dropping with the training step, the performance of the model on *training* dataset is better than that on the *validation* dataset.

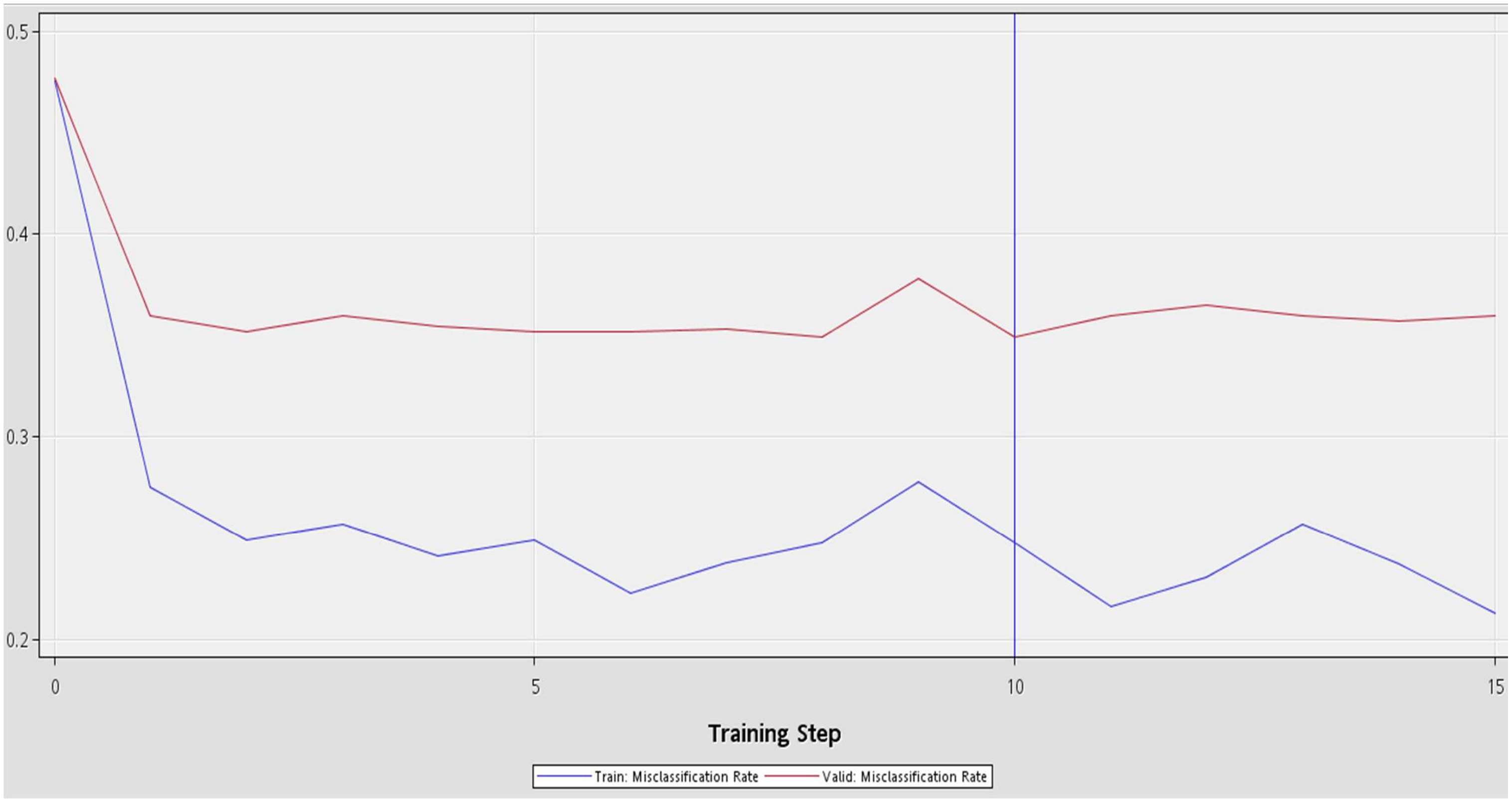


Figure 4.3 Misclassification rate for AutoNeural network model

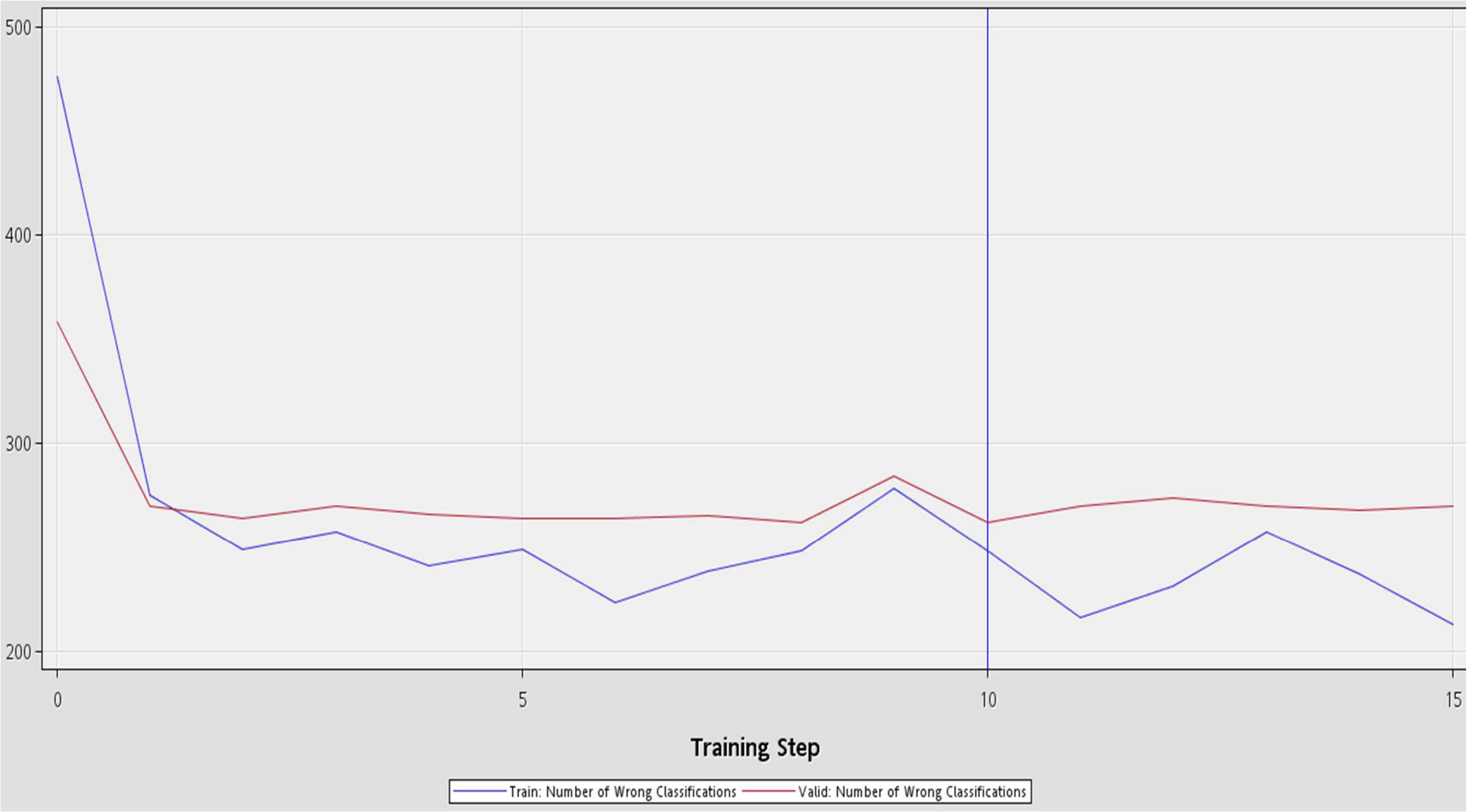


Figure 4.4 Number of wrong classifications for AutoNeural network model

CHAPTER V

CONCLUSION AND COMPARISON WITH PREVIOUS LITERATURE

5.1 Model Comparison and Conclusion

All three models described above are scored on the *validation* dataset and compared according to their receiver operating characteristic (ROC) curves and the ROC indices (i.e., the area under the ROC curve). ROC curve is a graphical plot that shows the performance of a statistical model as its discrimination threshold is varied. The curve is generated by plotting the sensitivity against false positive rate, i.e., $1 - specificity$, at different threshold settings. The ROC curves for all four models on *training, validation* and *test* datasets are as shown in Figure 5.1, and the ROC indices for all the models on the three datasets are given in Table 5.1. It can be seen from Table 5.1 that the performance of logistic regression and decision trees is consistent over all the three datasets, while that of AutoNeural network model appears to worsen as the model learns.

Dataset	**Logistic Regression**	**Decision Tree**	**AutoNeural Network**
Training	0.780	0.759	0.878
Validation	0.799	0.734	0.718
Test	0.791	0.759	0.689

Table 5.1 ROC indices for the statistical analysis models

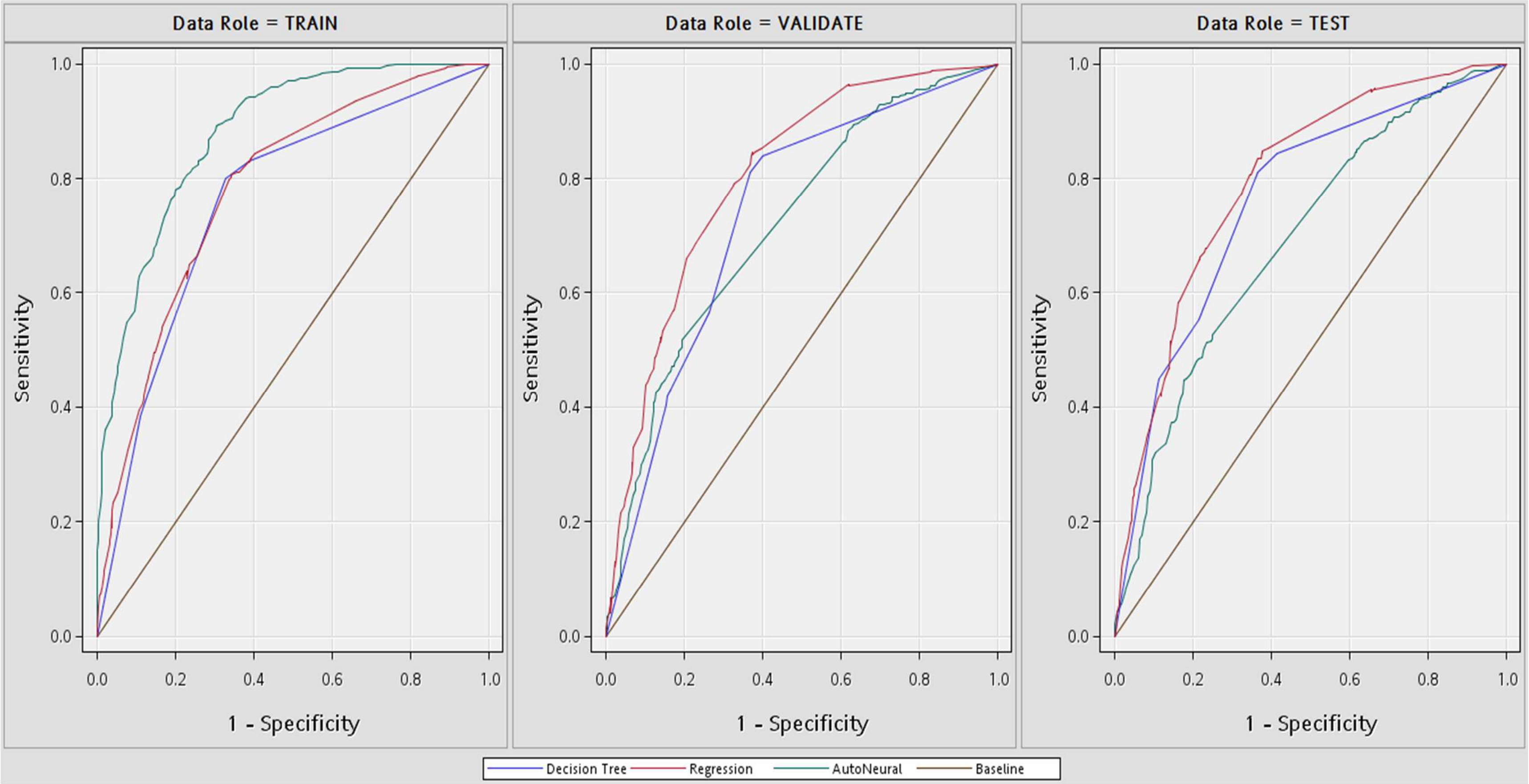


Figure 5.1 ROC curves for the statistical models on the three datasets

Also, the misclassification rates for the three models on *validation* and *test* datasets are as follows:

Dataset	Logistic Regression	Decision Tree	AutoNeural Network
Train	0.267	0.262	0.213
Validation	0.268	0.274	0.360
Test	0.274	0.274	0.375

Table 5.2 Misclassification rates for the statistical analysis models on validation and test datasets

Again, it can be seen from the values in Table 5.2 that the logistic regression model and the decision tree perform similarly in terms of wrong predictions, but the AutoNeural network model appears to worsen when compared to the other two models.

As the values of misclassification rate for logistic regression and decision tree are both approximately equal, it cannot be used as a selection criterion for the most effective model. Thus, the effective selection criterion is the ROC index for the *test* dataset. According to this and based on the dataset used in this thesis, the model that is considered the most effective in detecting the presence of bipolar disorder in a patient is the logistic regression model described in Section 4.2.

It can be concluded from the previous discussion that among the three techniques tested on the available dataset, logistic regression modelling is the most effective statistical analysis tool to detect whether or not a person has the bipolar disorder. The independent variables that have a significant impact on the patient getting a positive bipolar diagnosis are as follows:

1) Depression
2) Haloperidol
3) Sertraline
4) Aripiprazole
5) BPD_less_than_60
6) PTSD
7) Lithium
8) Respiratory_rate_abnormal
9) Marital_status

From the odds ratios listed in Table 4.3, we can conclude that compared to non-bipolar patients, the bipolar patients are more likely to also have PTSD, not have depression, have a normal respiratory rate, have diastolic blood pressure greater than 60 mmHg, be single, and to have been prescribed aripiprazole, haloperidol, lithium, and sertraline by doctors.

It can be concluded that if a patient is determined to have the bipolar disorder as per this statistical analysis, he or she may benefit from further screening by a psychologist for a medical diagnosis. Thus, this method can be used as a first step in diagnosing bipolar disorder in patients.

5.2 Comparison with Previous Literature

The study conducted in this thesis is similar to those described in previous journal publications by Inoue et al. (2015), Takeshima and Oka (2013), and Xiang (2013) that have been discussed in Section 2.3. Following is a summary of the comparisons made between these three publications and the results of this thesis:

1) According to Inoue et al.'s (2015) study, the factors found to be significantly correlated with bipolar disorder in patients are depression, antidepressant-related switch to hypomania, early age of onset of bipolar disorder, number of mood episodes in a year, family history of bipolar disorder, suicidal tendencies, interpersonal rejection sensitivity (hyper alertness and sensitivity to the reactions of other people in social situations), significant weight gain over a short period of time, and feelings of pathological guilt. Out of these, only pathological guilt has a negative correlation with bipolar disorder, i.e., bipolar patients are less likely to exhibit this kind of behavior. In our study, we could not test the significance of any of these factors because of the nature of the data available, except depression, which was found to have a negative correlation with bipolar diagnosis. That is, it was discovered that bipolar patients are less likely to have depression as a comorbidity. This is a significant deviation from the results of Inoue et al.'s (2015) study, but it can be due to the specific data that was used for analysis in this thesis. Also, the data used by Inoue et al. (2015) consists information about the patients that we did not have access to in our study, such as number of mood episodes and family history, among others. Further research needs to be done in order to determine the exact relationship between bipolar disorder and depression.
2) Takeshima and Oka's (2013) study reveals that the factors that have a significant correlation with a bipolar diagnosis are gender, family history of bipolar disorder, recurrent depressive episodes, cyclothymic temperament, depression, antidepressant-related switch to mania or hypomania, and early age at the onset of bipolar disorder on a person being diagnosed with bipolar disorder. All of these

features, except gender, were found by Takeshima and Oka (2013) to have a direct correlation with bipolar disorder, i.e., there is higher probability of the patient being bipolar when all these features take a positive value. In the case of gender, males were found to be more likely to be diagnosed with bipolar disorder. There are two major differences between the findings Takeshima and Oka's (2013) study and the results in this thesis:

a) Gender is found to be a significant variable by Takeshima and Oka (2013), while it was not an independent variable of significance in our logistic regression model.

b) According to Takeshima and Oka (2013), bipolar disorder patients are more likely to be diagnosed with depression as well, which is exactly the opposite of the findings in this thesis. Inoue et al. (2015) also suggest the comorbidity between the bipolar disorder and depression. Although our results show the exact opposite, we caution the reader that the dataset used in our analysis as well as the type of data used was different from those used by Takeshima and Oka (2013) and Inoue et al. (2015).

3) Xiang et al.'s (2013) study compared the features of bipolar disorder type I (BD – I) and type II (BD –II) to those of major depression. According to their findings, BD – I patients were more likely to be younger, males, suicidal, have more frequent depressive episodes with more atypical features (increased need for sleep, increased appetite, and weight gain), present psychotic features, have a strong family history of bipolarity, have an earlier age of onset. BD – II patients, in addition to showing all the same features as BD – I, were also more likely to be

married as compared to the patients of major depression. Even though the research methodology adopted by Xiang et al. (2013) is different from that adopted in this thesis, parallels can be drawn as both the studies use logistic regression to identify the variables with a significant correlation with bipolarity. Also, while Xiang et al. (2013) found that bipolar patients were more likely to be married, the results of our thesis suggest that they are more likely to be single.

As can be seen in the above paragraphs, the data used for analysis in this thesis was lacking a number of features that were considered by the three groups of researchers. On the other hand, the data was rich with other factors such as demographic information of patients (such as race and marital status, among other things), medication prescribed to the patients, pulse, respiratory rate, systolic and diastolic blood pressure, body mass index, body surface area, and several psychological other diseases such as diabetes and stroke.

CHAPTER VI

SCOPE, LIMITATIONS, AND FUTURE SCOPE

6.1 Scope

The statistical modelling tool that is found to be effective for detecting bipolar disorder in patients in the previous chapter, logistic regression, can be a useful tool in the hands of primary medical caregivers. It can be used to estimate the probability of a person having bipolar disorder as discussed in Section 4.2 and, if it is above a specific threshold, the corresponding patient can be recommended for additional testing and treatment. In other words, it can be used by medical professionals as a screening test before further detailed psychological interviews and tests are conducted to confirm the presence of the bipolar disorder.

6.2 Limitations

The main limitation of any statistical analysis tool, including logistic regression modelling, is that correlation does not necessarily imply causation. In other words, if there is a high correlation between the target variable and some of the independent variables, it does not necessarily mean that one causes the other. Thus, for example, if there is found to be a high correlation between a bipolar diagnosis and marital status of the patients, it would not be appropriate to assume that a person's being married or unmarried has any effect on their being a bipolar patient. Similarly, it would not be appropriate to assume that being prescribed aripiprazole, haloperidol, lithium, or sertraline causes the bipolar disorder. Instead, it can be interpreted that there is a high correlation between the bipolar disorder and the diseases these drugs are used to treat.

This can be explained by the comorbidities that commonly exist in a patient with bipolar disorder as discussed in Chapter 1.

Another limitation is that the application and efficiency of the three techniques used – logistic regression, decision trees, and AutoNeural networks – are heavily dependent on the quality and quantity of data available. If the input data is insufficient or has faulty and missing values, the output will also be poor as a result. This can be avoided by cleaning the *training* data to eliminate faulty values and data imputation in place of missing values.

6.3 Future Scope

In the future, datasets with information about the family history, age of onset, number of mood episodes in a year, etc., can be used to build statistical models for detecting the presence of bipolar disorder in patients. The number of data points used for analysis should also be increased in the future in order to improve the performance of AutoNeural networks. Also, because the bipolar disorder is a mental disorder, its effects on the cognitive abilities of a person can be studied using computational neuroscience techniques.

Logistic regression modelling, the statistical analysis tool that was found to be effective in detecting bipolar disorder in our study, can be developed into a decision support system that can be used by medical professionals with relative ease. The medical history charts that a person fills when they go to their doctor's office or to a hospital can be used as the data source for this system. The values of the independent variables in the logistic regression equation derived in Section 4.2 can be fed to the system, which will give the probability of the patient being bipolar as a result. Based on this probability, the patient can either be sent for further screening by a psychiatrist to medically diagnose the bipolar disorder if it is above a threshold (0.56 as calculated by SAS) or no further action taken if it is below the threshold, unless the patient presents symptoms in a specified time period following the analysis. A flow chart for the proposed system is shown in Figure 6.1:

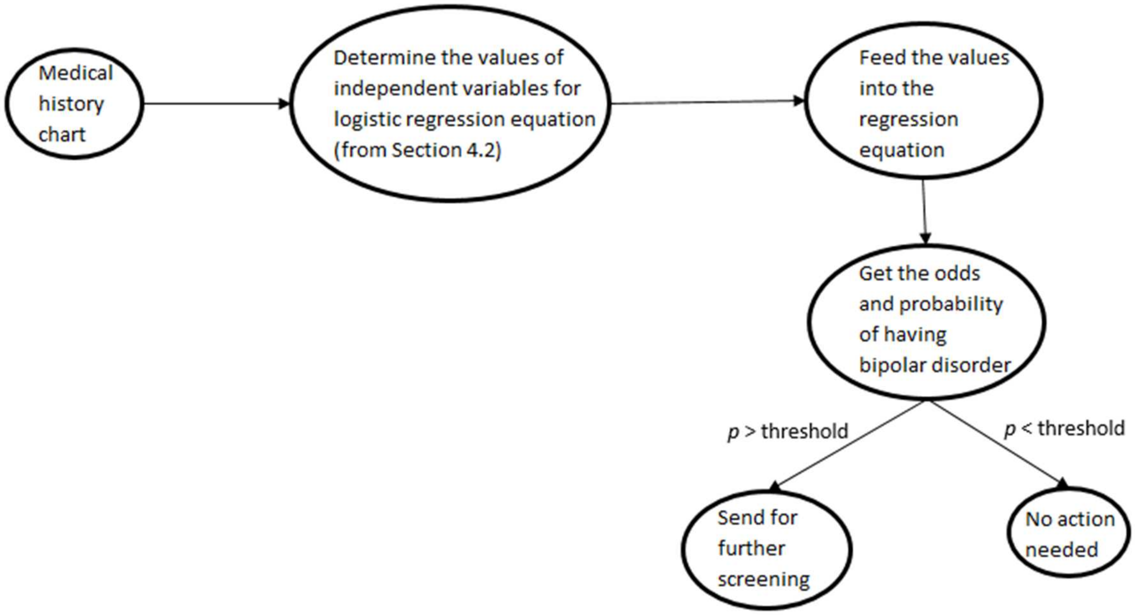

Medical history chart
Determine the values of independent variables for logistic regression equation (from Section 4.2)
Feed the values into the regression equation
Get the odds and probability of having bipolar disorder
p > threshold
p < threshold
Send for further screening
No action needed

Figure 6.1 Flow chart for the proposed decision support system

To summarize, this decision support system can be used as a screening tool for patients to identify the patients potentially suffering from the bipolar disorder, who can then be sent to a psychiatrist for further interviews, tests, and analysis in order to medically diagnose them as having bipolar disorder. On the other hand, the patients who the system identifies as not at risk of having the bipolar disorder can be checked for other psychological disorders that have symptoms similar to bipolar disorder, such as depression.

REFERENCES

1) Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., Snyder, P. J., “Voice Acoustical Measurement of the Severity of Major Depression,” *Brain and Cognition*, vol. 56, no. 1, 2004, pp. 30 – 35.

2) Cao, Liangyue (1997), "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D: Nonlinear Phenomena* 110.1: 43-50.

3) Charney, D. S., Reynolds, C. F., Lewis, L., Lebowitz, B. D., Sunderland, T., Alexopoulos, G. S., ... & Borson, S. (2003), “Depression and Bipolar Support Alliance consensus statement on the unmet needs in diagnosis and treatment of mood disorders in late life,” *Archives of General Psychiatry*, 60(7), 664-672.

4) Cohen, H., Kaplan, Z., Kotler, M., Mittelman, I., Osher, Y., & Bersudsky, Y. (2003), “Impaired heart rate variability in euthymic bipolar patients,” *Bipolar disorders*, 5(2), 138-143.

5) Das, A. K., Olfson, M., Gameroff, M. J., Pilowsky, D. J., Blanco, C., Feder, A., ... & Weissman, M. M. (2005), “Screening for bipolar disorder in a primary care practice,” *Journal of American Medical Association*, 293(8), 956-963.

6) de Jong, N. H., Wempe, T., "Praat Script to Detect Syllable Nuclei and Measure Speech Rate Automatically," *Behavior Research Methods*, vol. 41, no. 2, 2009, pp. 385 – 390.

7) Dean, B. B., Gerner, D., & Gerner, R. H. (2004), "A systematic review evaluating health-related quality of life, work impairment, and healthcare costs and utilization in bipolar disorder," *Current Medical Research and Opinion*, 20(2), 139-154.

8) Greco, A. Lanata, A., Valenza, G. Rota, G., Vanello, N., and Scilingo, E., "On the Deconvolution Analysis of Electrodermal Activity in Bipolar Patients," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Bio. Soc.*, 2012, pp. 6691 – 6694.

9) Hirschfeld, R. M., Cass, A. R., Holt, D. C., & Carlson, C. A., "Screening for bipolar disorder in patients treated for depression in a family medicine clinic," *The Journal of the American board of family practice*, vol. 4, article 18, July 2005, pp. 233 – 239.

10) Inoue, T., Inagaki, Y., Kimura, T., & Shirakawa, O. (2015). "Prevalence and predictors of bipolar disorders in patients with a major depressive episode: The Japanese epidemiological trial with latest measure of bipolar disorder (JET-LMBP)", *Journal of affective disorders*, 174, 535-541.

11) Koukopoulos, A., Reginaldi, D., Tondo, L., Visioli, C., Baldessarini, R. J., "Course Sequences in Bipolar Disorder: Depressions Preceding or Following Manias or Hypomanias," *Journal of Affective Disorders*, 151, May 2013, 105 – 110.

12) Krishnan, K. R. R. (2005), “Psychiatric and medical comorbidities of bipolar disorder,” *Psychosomatic Medicine*, 67(1), 1-8.

13) Matza, L. S., Rajagopalan, K. S., Thompson, C. L., & de Lissovoy, G. (2005), “Misdiagnosed patients with bipolar disorder: comorbidities, treatment patterns, and direct treatment costs,” *The Journal of Clinical Psychiatry*, 66(11), 1432-1440.

14) Mariani, S., Migliorini, M., Tacchino, G., Bertschy, G., Werner, S., Bianchi, A. M., “Clinical State Assessment in Bipolar Patients by Means of HRV Features Obtained with a Sensorized T-Shirt,” in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2012, pp. 2240 – 2243.

15) Migliorini, M., Mendez, M. O., Bianchi, A. M., “Study of Heart Rate Variability in Bipolar Disorder: Linear and Non-Linear Parameters during Sleep,” *Frontiers in Neuroengineering,* vol. 4, article 22, January 2012, pp. 1 – 7.

16) Moreno, C., Laje, G., Blanco, C., Jiang, H., Schmidt, A. B., & Olfson, M. (2007), “National trends in the outpatient diagnosis and treatment of bipolar disorder in youth,” *Archives of General Psychiatry*, 64(9), 1032-1039.

17) Perlis, R. H., Dennehy, E. B., Miklowitz, D. J., DelBello, M. P., Ostacher, M., Calabrese, J. R., ... & Nierenberg, A. A. (2009), “Retrospective age at onset of bipolar disorder and outcome during two ‐ year follow ‐ up: results from the STEP ‐ BD study,” *Bipolar Disorders*, 11(4), 391-400.

18) Quilty, L. C., Sellbom, M., Tackett, J. L., & Bagby, R. M. (2009), "Personality trait predictors of bipolar disorder symptoms," *Psychiatry research*, 169(2), 159-163.

19) Sachs, G. S., Baldassano, C. F., Truman, C. J., & Guille, C. (2000), "Comorbidity of attention deficit hyperactivity disorder with early-and late-onset bipolar disorder," *American Journal of Psychiatry*, 157(3), 466-468.

20) Sanchez-Moreno, J., Martinez-Aran, A., Tabares-Seisdedos, R., Torrent, C., Vieta, E., & Ayuso-Mateos, J. L. (2009), "Functioning and disability in bipolar disorder: an extensive review" *Psychotherapy and Psychosomatics*, 78(5), 285-297.

21) Stassen, H. H., Kuny, S., & Hell, D. (1998), "The speech analysis approach to determining onset of improvement under antidepressants," *European Neuropsychopharmacology*, 8(4), 303-310.

22) Slomka, J. M., Piette, J. D., Post, E. P., Krein, S. L., Lai, Z., Goodrich, D. E., & Kilbourne, A. M. (2012), "Mood disorder symptoms and elevated cardiovascular disease risk in patients with bipolar disorder," *Journal of Affective Disorders*, 138(3), 405-408.

23) Todder, D., Bersudsky, Y., Cohen, H., "Nonlinear Analysis of RR Interval in Euthymic Bipolar Disorder," *Autonomic Neuroscience*, 117.2 (2005): pp. 127 – 131.

24) Valenza, G., Nardelli, M., Lanata, A., Gentili, C., Bertschy, G., Paradiso, R., Scilingo, E. P., “Wearable Monitoring for Mood Recognition in Bipolar Disorder based on History-Dependent Long-Term Heart Rate Variability Analysis,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, September 2014, pp. 1625 – 1635.

25) Vanello, N., Guidi, A., Gentili, C., Werner, S., Bertschy, G., Valenza, G., Lanata, A., Scilingo, E., “Speech Analysis for Mood State Characterization in Bipolar Patients,” in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2012, pp. 2104 – 2107.

26) Voss, A., Baier, V., Schulz, S., Bar, K. J., “Linear and Nonlinear Methods of Analyses for Cardiovascular Variability in Bipolar Disorders,” *Bipolar Disorders*, vol. 8, issue 5p1, October 2006, pp. 441 – 452.

27) Takeshima, M., & Oka, T. (2013). “A comprehensive analysis of features that suggest bipolarity in patients with a major depressive episode: Which is the best combination to predict soft bipolarity diagnosis?” *Journal of affective disorders*, 147(1), 150-155.

28) Xiang, Y. T., Zhang, L., Wang, G., Hu, C., Ungvari, G. S., Dickerson, F. B., ... & Chiu, H. F. (2013), “Sociodemographic and clinical features of bipolar disorder patients misdiagnosed with major depressive disorder in China,” *Bipolar disorders*, 15(2), 199-205.

APPENDIX A: DATA DICTIONARY

Variable Name	Description	Possible Values
Encounter_ID	A sequential number created every time a patient visits a Cerner facility.	
Patient_ID	Unique identifier attached to each individual patient.	
Age_in_years	Age of the patient.	18-65
Census_region	Numeric value indicating which census region the patient is from.	1: Midwest 2: Northeast 3: South 4: West
Gender	Gender of the patient.	0: Male 1: Female
Urban_rural_status	Binary column indicating whether the patient is from a rural or urban background.	0: Rural 1: Urban
Race	Numeric value indicating the race of the patient.	1: African American 2: Asian 3: Caucasian 4: Hispanic 5: Native American
Marital_status	Binary column indicating whether the patient is single or not.	0: Single 1: Not single
PTSD	Binary column indicating whether the patient has PTSD or not.	0: Negative 1: Positive
ADHD	Binary column indicating whether the patient has ADHD or not.	0: Negative 1: Positive
Depression	Binary column indicating whether the patient has depression or not.	0: Negative 1: Positive
Panic_disorder	Binary column indicating whether the patient has panic disorder or not.	0: Negative 1: Positive

Bipolar	Binary column indicating whether the patient has bipolar disorder or not.	0: Negative 1: Positive
Alprazolam	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat anxiety and panic disorder.	0: Negative 1: Positive
Amphetamine_dextroamphetamine	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat ADHD and narcolepsy.	0: Negative 1: Positive
Aripiprazole	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat schizophrenia, bipolar disorder, and Tourette syndrome.	0: Negative 1: Positive
Bupropion	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression and to quit smoking.	0: Negative 1: Positive
Chlorpromazine	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat behavioral disorders and anxiety before a surgery.	0: Negative 1: Positive
Citalopram	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression.	0: Negative 1: Positive
Desipramine	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression.	0: Negative 1: Positive
Duloxetine	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression and anxiety.	0: Negative 1: Positive
Escitalopram	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression and anxiety.	0: Negative 1: Positive

Fluoxetine	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression, obsessive-compulsive disorder, bulimia nervosa, and panic disorder.	0: Negative 1: Positive
Haloperidol	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat schizophrenia, Tourette syndrome, mania, nausea and vomiting, delirium, agitation, psychosis, and hallucinations.	0: Negative 1: Positive
Lithium	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat bipolar disorder and schizoaffective disorder.	0: Negative 1: Positive
Methylphenidate	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat ADHD and narcolepsy.	0: Negative 1: Positive
Paroxetine	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression, anxiety disorder, obsessive-compulsive disorder, and premenstrual dysphoric disorder.	0: Negative 1: Positive
Pioglitazone	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat type II diabetes.	0: Negative 1: Positive
Sertraline	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat obsessive-compulsive disorder, PTSD, premenstrual dysphoric disorder, anxiety disorder, and panic disorder.	0: Negative 1: Positive
Trazodone	Binary column indicating whether the patient has been prescribed the	0: Negative 1: Positive

	medication or not. Used to treat depression.	
Valporic_acid	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat seizures, bipolar disorder, and migraines.	0: Negative 1: Positive
Venlafaxine	Binary column indicating whether the patient has been prescribed the medication or not. Used to treat depression, anxiety disorder, and panic disorder.	0: Negative 1: Positive
Alcohol_use	Alcohol consumption of the patient per day.	
BPD_greater_than_90	Binary column indicating whether the patient's diastolic blood pressure is above 90, i.e., is higher than normal.	0: Below 90 1: Above 90
BPS_greater_than_140	Binary column indicating whether the patient's systolic blood pressure is above 140, i.e., is higher than normal.	0: Below 140 1: Above 140
BPD_less_than_60	Binary column indicating whether the patient's diastolic blood pressure is below 60, i.e., is lower than normal.	0: Above 60 1: Below 60
BPS_less_than_90	Binary column indicating whether the patient's systolic blood pressure is below 90, i.e., is lower than normal.	0: Above 90 1: Below 90
Mean_arterial_pressure	Mean arterial pressure of the patient.	
BMI	Body mass index of the patient.	
BSA	Body surface area of the patient.	
Heart_rate_abnormal	Binary column indicating whether the patient's heart rate is abnormal.	0: Normal (between 60 and 100 beats a minute) 1: Abnormal (otherwise)
Pulse_abnormal	Binary column indicating whether the patient's pulse is abnormal.	0: Normal (between 60 and 100 beats a minute) 1: Abnormal (otherwise)
Respiratory_rate_abnorm	Binary column indicating whether the	0: Normal (between

al	patient’s respiratory rate is abnormal.	12 to 25 breaths a minute) 1: Abnormal (otherwise)
Smoking_packs_per_day	Number of cigarette packs the patient smokes in a single day.	
Tobacco_use_packs_day	Number of tobacco packs the patient consumes in a single day.	

APPENDICEX B: STATISTICAL METHODS

B.1 Logistic Regression Analysis

Regression analysis is a statistical method used to analyze a dataset with one or more independent variables so that one of the variables, called the dependent variable, can be predicted from the other(s). Examples include the cause-and-effect relationships between income and income tax, age and mortality rate, or education and income. As such, it is a very handy tool that allows users to predict a response variable (also known as the effect variable) based on the values of one or more independent variables. There are two types of regression analysis methods, and they are discussed in more detail in the following paragraphs.

B.1.1 Linear Regression

Linear regression is a method used to model the explanatory relationship between a nominal dependent variable (such as income tax in the example given above) and one or more independent variables. A functional form of a linear regression model with *n* independent variables is provided next:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \in$$

Where, y = continuous dependent variable

x_i = independent input variables

ε = error term

β_0 = intercept

β_0 = coefficients of the independent variables

In the above formulation, if the number of independent variables is one, the model is called as a *simple* linear regression model; otherwise, it is known as a *multiple* linear regression model.

To test accuracy of the regression model and to determine whether there is a significant relationship between the target variable and the independent variables, we test the hypothesis that the intercept and coefficients are equal to zero. If the *p*-value in the hypothesis test is greater than the significance level of α, the null hypothesis (that there is no significant relationship between the target and independent variables) can be rejected. If, on the other hand, the *p*-value is less than α, the null hypothesis cannot be rejected and the independent variables are said to have a significant relationship with the target variable.

B.1.2 Logistic regression

Logistic regression is a special type of regression analysis that is used when the dependent variable is binary, i.e., it can take on only one of two distinct values (e.g., positive or negative, 1 or 0, alive or dead, etc.). Similarly as in the case of linear regression discussed in 4.1.1, there can be one or more independent variables in logistic regression as well. A functional form of the generalized logistic regression model for a binary dependent variable is provided next:

$$logit(p_i) = ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ji}$$

Where, $y_i = \begin{cases} 1, \ target\ event\ for\ case\ i \\ 0, \ no\ target\ event\ for\ case\ i \end{cases}$

$$x_i = (x_{1i}, x_{2i}, \dots, x_{ji}) = vector\ of\ input\ variables$$

$$with\ j\ values\ for\ the\ i^{th}\ case$$

$$p_i = probability\ of\ the\ presence$$

$$of\ the\ target\ event\ for\ case\ i, i = 1, 2, \dots, n$$

Whether there is a significant relationship between the independent variables and the target variable in a logistic regression model is determined in the same fashion as in the case of linear regression analysis.

B.2 Decision Tree Learning

Decision tree is a tree-like model of decisions and their outcomes. It is used to predict the value of the target variable based on the values of input variables. An illustrative example of this is shown in Figure B.1 below:

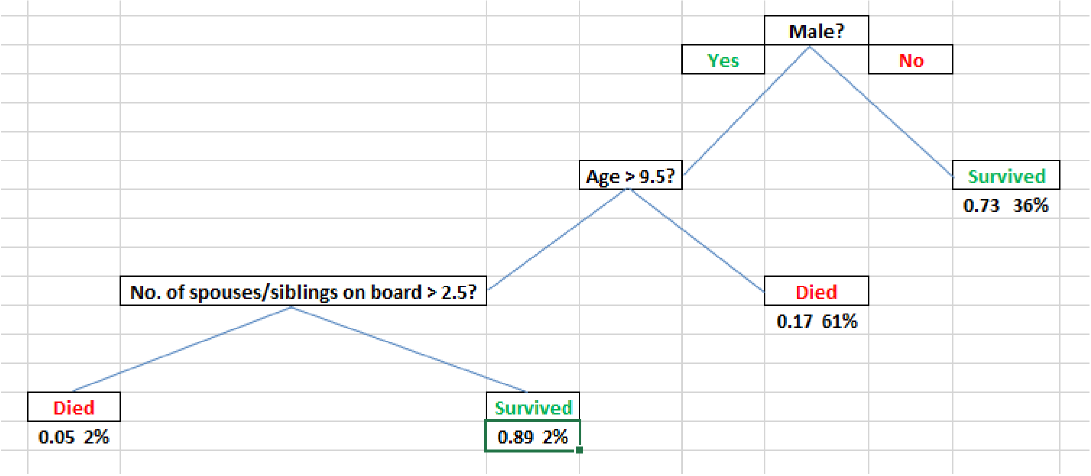


Figure B.1 Illustrative example of a decision tree

The decision tree in Figure B.1 models the survival of passengers aboard the *Titanic.* Each leaf node (end node in the decision tree) represents the value of the target variable (i.e., *died* or *survived*) if the input variable values are represented by the intermediate node values (also called decision nodes) of the path from root (topmost decision node) to leaf. A decision tree is always constructed top-down, starting with the best predictor as the root node, the second best predictor at the next step, and so on. In the above example, the numbers below each of the leaves represent the probability of surviving and the percentage of observations in the tree.

B.3 Artificial Neural Networks

Artificial neural networks is a machine learning tool based on the biological neural networks that are used to predict the approximate functions for the target variables that are dependent on a large number of input variables. These are constructed as different layers of *neurons* that communicate with each other. This can be explained better by looking at the most basic type of artificial neuron that takes in several binary inputs x_1, x_2, ..., x_n and produces a single binary output as shown in Figure B.2 below:

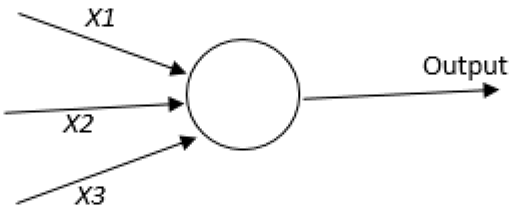


Figure B.2 Single artificial neuron

Suppose that w_1, w_2, and w_3 are real numbers that indicate the importance of the inputs x_1, x_2, and x_3 respectively in the example shown in Figure B.2. Then, the output of the neuron will be as follows:

$$Output = \begin{cases} 0 \ \ if \sum_j w\,.x + b \ \leq 0 \\ 1 \ \ if \sum_j w\,.x + b \ \geq 0 \end{cases}$$

Where, $w\,.x \equiv \sum w_j x_j$, dot product of the weight and input vectors

b = bias or negative *threshold* value of the neuron for making a decision

An artificial neural network consists of several such neurons that take in inputs and feed them to another layer of neurons that generate more complex outputs. These neurons can again feed their outputs to another layer of neurons or generate the final output by feeding to a single neuron. The more the number of layers, the more complex the artificial neural network. This is illustrated in Figure B.3:

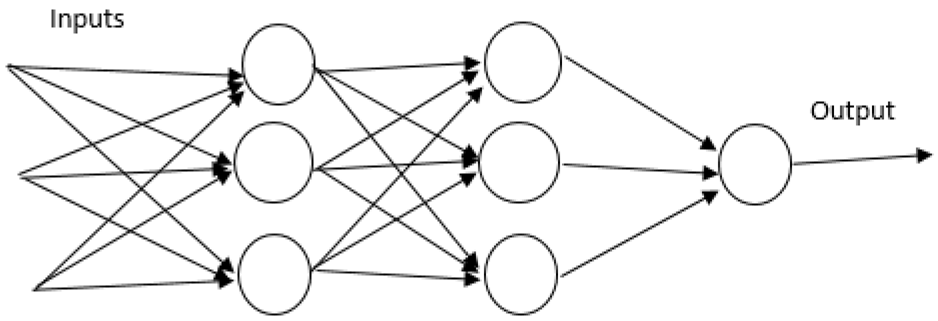


Figure B.3 Artificial neural network

VITA

UTTARA VINAY TIPNIS

Candidate for the Degree of

Master of Science

Thesis: STATISTICAL ANALYSIS OF DEMOGRAPHIC AND CLINICAL DATA OF PATIENTS TO DETERMINE THE MARKERS OF BIPOLAR DISORDER

Major Field: INDUSTRIAL ENGINEERING

Biographical:

Education:

Completed the requirements for the Master of Science in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in July, 2016.

Completed the requirements for the Bachelor of Engineering in Mechanical Engineering at Pune University, Pune, India in 2013.

Experience:

Graduate Research and Teaching Assistant (08/14 to Date)
Oklahoma State University

- Thesis: "Statistical Analysis of Demographic and Clinical Data of Patients to Determine the Markers of Bipolar Disorder" (as an RA)
- Teaching a lab section of ENGR 1322, a CAD design course (as a TA)

Professional Memberships:

INFORMS Student Chapter, Oklahoma State University
President (08/15 to 12/15)
President Elect (01/15 to 08/15)
Treasurer (08/14 to 12/14)