Project Report

On

PTSD Prediction and Analysis: A Machine Learning Approach

Submitted in partial fulfillment of the requirements

for the award of the degree of

BACHELOR OF TECHONOLOGY

IN

INFORMATION TECHNOLOGY

Malik Umair Nazir
IT/598/14

Muneeb Un Nabi
IT/582/14

Under the supervision of

Ms. Iqra Altaf Gillani **Dr. Shabir A. Sofi**

Department of Information Technology

National Institute of Technology Srinagar

J&K - 190006

Spring 2018

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLGY SRINAGAR, J&K

JUNE 2018

CERTIFICATE

This is to certify that the project titled **PTSD Prediction and Analysis: A Machine Learning Approach** has been completed by **Malik Umair Nazir (IT/598/14)** and **Muneeb Un Nabi (IT/582/14)** under my supervision in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Information Technology**. It is also certified that the project has not been submitted or produced for the award of any other degree.

Ms. Iqra Altaf Gillani
Project Mentor
Department of Information Technology
NIT, Srinagar

Dr. Shabir Ahmad Sofi
Project Mentor cum Coordinator
Department of Information Technology
NIT, Srinagar

STUDENTS DECLARATION

We, hereby declare that the work, which is being presented in the project entitled **PTSD Prediction and Analysis: A Machine Learning Approach** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology** in the session 2018, is an authentic record of our own work carried out under the supervision of **Ms. Iqra Altaf Gillani** and **Dr. Shabir A. Sofi**, Department of Information Technology, National Institute of Technology, Srinagar. The matter embodied in this project has not been submitted by us for the award of any other degree.

Dated:

Name : Malik Umair Nazir
Signature:

Name : Muneeb Un Nabi
Signature: .

ACKNOWLEDGEMENTS

It is our privilege to express our sincerest regards to our Head of the Department **Mrs. Arooj Nissar** for encouraging and allowing us to present the project on the topic **PTSD Prediction and Analysis: A Machine Learning Approach** at our department premises for the partial fulfillment of the requirements leading to the award of B-Tech degree.

We would like to express our special thanks of gratitude to our project mentor cum coordinator **Dr. Shabir A. Sofi** who provided us with timely guidance and contacts to reach out to the community and collect a dataset for the project.

We deeply express our sincere thanks to our project mentor, **Ms. Iqra Altaf Gillani**, for her valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this project. We are specially thankful to her for providing us with her fullest support to understand the conceptual underpinnings of the project and paved way for its implementation, which also helped us in doing a lot of Research and we came to know about so many new things. We are really thankful to you.

We would also like to thank specific people from the medical community especially **Dr. Mohd Maqbool** (HOD, Psychiatry Hospital), **Dr. Arshid** (Valley's reputed psychiatrist), **Dr. Rubia** (a research scholar) and staff members at SMHS Hospital and IMHNS Hospital who helped us a lot in data collection and made it possible to complete the project within the stipulated frame of time.

We take this opportunity to express gratitude to all of the Department faculty members for their help and support. We also thank our parents for the unceasing encouragement, support and attention. We are also very much grateful to each other (as project partners) for supporting each other and carrying out this work with zeal and zest.

Malik Umair Nazir IT/598/14

Muneeb Un Nabi IT/582/14

8th Semester

B.Tech, Information Technology

NIT Srinagar

Abstract

In our valley Kashmir, people from different regions are at increased risk of depressive symptoms and post traumatic stress disorder (PTSD) due to conflicting situations arising time and again, or extreme environmental conditions like natural disasters. Accurate diagnosis and determining the causes are very important to cure these kinds of psychological problems. Machine learning (ML) techniques are gaining popularity in neuroscience due to their high diagnostic capability and effective classification ability. In this project, advanced feature selection methods are used to discover the compact causal models of PTSD and at the same time predict the performance of classification algorithms like SVMs vis-a-vis selected features. Thus, not only the PTSD individuals are classified by ML techniques such as SVMs, but also the important indications of patients trauma are determined by the most popular causal discovery feature selection methods like HITON-PC. The effectiveness of the proposed system is examined on a real-world dataset collected by us as part of the clinician-based study we undertook to study the traumatic situations, causes and consequences from our valley, Kashmir.

Contents

List of Figures

List of Abbreviations

PTSD	Post Traumatic Stress Disorder
ASD	Acute Stress Disorder
ML	Machine Learning
DSM V	Diagnostic and Statistical Manual Version V
CAPS V	Clinician Administered PTSD Scale for DSM V
NB	Naive Bayes (Classifier)
SVM	Support Vector Machine
RAD	Reactive Attachment Disorders
MSF	Medicans Sans Frontiers
NCS	National Co-morbidity Survey
ECA	Epidemiological Catchment Area
WW2	World War Second
MB	Markov Boundary
PC	Parent Children
PTE	Potentially Traumatic Events
AUC	Area Under (Receiver Operating Characteristic) Curve

ROC	Receiver Operating Characteristic Curve
HITON-PC	HITON Parent Children
HITON-MB	HITON Markov Boundary
SMHS	Shri Maharaja Hari Singh (Hospital)
DS	Diagnostic Status
SS	Symptom Severity

Chapter 1

Introduction

A mental health disorder is an abnormal psychological condition that affects a person's thinking, mood or feeling in one way or the other. Such conditions affect the normal functioning of the person like social and occupational functioning and also all the cognitively demanding actions from simple decision making to complex problem solving. Each person will have different experiences, even people with the same diagnosis.The disorders may range from mild disorders like phobias to the most debilitating ones like depression. Many mental health disorders may lead to other chronic disorders like Post-traumatic Stress Disorder (PTSD) or suicidal ideation.A mental health disorder isn't the result of one event a person has faced during his lifetime. Researchers have been to the point that most of the psychological disorders are multi-causal in nature. Genetics, environment and lifestyle changes across societies determine profoundly whether a person will develop a mental health disorder or not. A stressful job or a unhealthy family condition makes a person more susceptible to mental issues, while as the traumatic situations like combat may give rise to severe mental health disorders like Post-traumatic Stress Disorder (PTSD). One of the debilitating disorders is PTSD, which is prevalent in almost every part of the world especially war and conflict-ridden zones. It is a psychiatric problem which arises in some people who face traumatic situations which in turn leaves indelible scars on their minds, leaving them mutilated and maimed and drastically bars their normal life process. Due to its multi-modal, multi-causal and heterogeneous nature, it is one of the complex most disorders in terms of diagnosis and analysis.

1.1 Importance and Objectives

1.1.1 Problem Definition

Determination of prediction models and context-driven causal discovery features as potential risk indicators of non-remitting PTSD from Kashmir using machine learning methods (Feature Selection and Supervised Classification Algorithms).

1.1.2 Motivation

Due on ongoing turmoil from last few decades, majority of people from Kashmir are suffering in one way or the other from mental health disorders, emanating in the form of depression, trauma related disorders like Acute Stress Disorder (ASD) and Post-traumatic Stress Disorder (PTSD), or substance abuse, to name a few. One of the debilitating disorders is PTSD, which has taken a toll of 15 - 19% of the population [1, 2]. The motivating factors which prompted us to work on this project are summarized as:

1. Diagnostic analysis of psychiatric problems like PTSD is based on compact set of high-level risk indicators that persist for around 1 month, thus has lesser implications for targeted prevention.

2. Statistical Analysis is not optimally suited to explore complex interactions between linear, nonlinear and non-normally distributed variables defining PTSD symptom trajectories thus newer statistical tools need to identified which mitigate those limitations.

3. Machine learning (ML) can handle large complex data with heterogeneous distributions, determine probabilistic relationships from complex conditional dependencies between variables, and test the reliability of the results through repeated cross validation.

4. ML-based analysis of PTSD has not been used previously in Kashmir to explore a compact set of risk factors, and considering the heterogeneity of the psychological problems like PTSD, a need arises for the determination of an exact set of risk factors in the context of Kashmir.

Newer advances in data analysis contributed by the field of Machine Learning greatly extend the researchers' ability to make meaningful discoveries also by:

1. Enabling accurate and reliable prediction using data with very large numbers of variables and small sample sizes.

2. Enabling causal inference within non-experimental data sets.

On a purely empirical and practical level, research using ML methods has met exceptional success in a wide range of scientific and technological fields, and it is beginning to penetrate the domain of clinical science, including the fields of psychiatry and pediatrics. ML has demonstrated utility in a variety of applications including the accurate classification in pediatric disorders such as epilepsy, asthma, heart disease, and head injury. Within psychiatry, ML has been successfully used in the predictive classification of autism, attention deficit hyperactivity disorder, and schizophrenia. ML has recently been used to predict PTSD in acutely traumatized adults [10] as well as children [11]. The possibility of using ML to identify causal processes initiated shortly after exposure to trauma has important implications for prevention. The detection of such causal processes may thus identify promising targets for preventative intervention and may shed light on the etiology of PTSD emergence in Kashmir.

1.1.3 Objectives

- To apply machine learning techniques in order to predict the morbidity of PTSD in a person after facing a trauma.

- To predict the most stable causal discovery risk indicators leading to the occurrence of PTSD in the context of Kashmir.

- Analyze and predict PTSD using Kashmir-based data set.

- To find causation effects of symptom severity and diagnostic status as target variables with PTSD.

- To build a smart app to act as an early on PTSD predictor.

1.2 Organization

This report is organized as follows. We survey the recent literature in Chapter 2 followed by an introductory discussion on various mental health disorders with special focus on PTSD in Chapter 3. We discuss the algorithms and tools used in Chapter 4. In Chapter 5 we discuss our project implementation in detail. We discuss our developed Android app in Chapter 6. Our results and associated discussion is presented in Chapter 7. Finally, we conclude and give directions for future work in Chapter 8.

Chapter 2

Literature Survey

2.1 PTSD Analysis: A case study from Kashmir

2.1.1 Introduction

Jammu and Kashmir has been witnessing a continuous mass trauma situation for more than 30 years, with thousands of people dead, maimed and mutilated, many missing or confined, thousands of children orphaned, and women widowed, a colossal damage to the property, and a damage to the cultural ethos including en-masses migration of a minority community. The amount of trauma continued and experienced,hence, remains anything but hard to imagine. The fury unleashed by natural disasters (snowstorm and earthquakes to recount a few) during the same time cannot be undermined either. Due to this continuous turmoil, traumatic events and political insurgency, Kashmiri population have been affected dramatically. Different studies regarding the mental health status of Kashmir have been carried out times and again.It had showed that different prevalent mental disorders like depression, anxiety disorders, and PTSD have been alarmingly increased in Kashmir.It had showed that in year 1990 about 1700 patients visited Kashmir's sole psychiatric services of the department, but in year 2002 the number had gone up to 48000. By December of 2004, 62000 patients had already visited the psychiatric services of the department. The studies based on these patients found women and children to be the worst hit . PTSD was a rather unknown diagnosis in pre-90s era, but a seizable number of outpatient cases were found suffering from it, besides depressive disorders. These patient population based figures are just the tip of the iceberg, whose chunk remains buried in the society, held by social ostracism associated with visiting mental health care persons or services[1].

2.1.2 Other studies done on PTSD in Kashmir

Kashmir Mental Health Survey 2015 conducted by Médecins Sans Frontières in collaboration with Institute of Mental Health and Neurosciences, Kashmir.This survey has been has been conducted in all the districts of Kashmir Valley, focusing mainly on most prevalent disorders like depression, anxiety disorders, and post traumatic stress disorder. Despite a limitation of use of only screening tools in the estimation of the prevalence of these disorders, the findings confirm a serious mental health situation, with highly prevalent common mental disorders and distress having continued to increase to reach epidemic levels among the traumatized population of Kashmir, with 37% of adult males and 50% of females suffering from probable depression; 21% of males and 36% of females from a probable anxiety related disorder and 18% men and 22% women suffering from probable PTSD [2].

A community study was undertaken by Margoob et al. [1], aiming at evaluation of the presence and magnitude of adult PTSD in a region from South Asia.The survey was done in all the 6 districts of Kashmir, selecting a total of 2391 adult subjects randomly. They were assessed using DSM IV (Diagnostic Statistical Manual IV) based MINI neuropsychiatric interview. Assessment yielded a current PTSD rate of 7.27% and lifetime PTSD rate of 15.19%. Importantly the rates in males and females were comparable.

2.2 PTSD in other parts of the World

There is a significant prevalence of PTSD in other developed and developing nations. Various general studies over population of developed countries reveal that 15-24% of people exposed to traumatic events have been reported to developed PTSD.As per National Co-morbidity Survey(NCS), 50% of adult community in US have been exposed to different traumatic events [15]. Community based survey showed PTSD to be the commonest anxiety disorder , with a lifetime prevalence of 8% in adults, and female: male prevalence ratio of 2:1.

The Epidemiological Catchment Area (ECA) survey, one of the largest in the field of psychiatry, was carried out in three different regions of the USA, with patients from different communities, reflecting prevalence and potency of different traumas. The St. Louis study done in two waves revealed lifetime prevalence of PTSD at 0.5% in males and 1.3% in females in general population, and rates being 15% in males and 10% in females, in populations subjected to trauma . The North Carolina site study revealed lifetime prevalence of 1.3 % . The Detroit/Michigan site study revealed lifetime trauma exposure of 39.1 %, with 23.6% of the exposed progressing to PTSD, yielding a lifetime prevalence of PTSD at 9.2% .Although the USA studies have reported a prevalence rate of 25% to 80% of trauma, among some of its population subgroups, only a fraction of them develop PTSD.

In another study on earthquake victims in two villages in China at different distances from the epicenter, using both DSM IV, it was found that the village with high level of initial exposure to earthquakes,and a high level of post quake support had less frequency of PTSD (19.8%, at 9 months), than the village with a lower level of initial exposure and less post quake support(30.3% at 9 months) [16].

Another important source of knowledge about PTSD prevalence and trends has been the war veteran group studies from the USA. In National Vietnam Readjustment study [17], 15% males active in war operations had a current episode of PTSD and 11% had a partial PTSD, in comparison to 3% prevalence in the same era Veterans who did not serve in South East Asian Region (SEAR), and 1% prevalence in civilian controls. The same study puts the lifetime prevalence of PTSD at 30.9% for males and 26% for females [18]. Another study noted a prevalence of 15% among Vietnam veterans, 20yrs after war[17]. A very high prevalence rate (50%-70%) of PTSD has been reported among the Prisoners of War [19] with 40 year prospective study on World War II (WW2) victims revealing rates of 47% and 50%, in males and females respectively[20, 21].Studies have also revealed a 20% to 40% PTSD rates following terror attacks [22, 23, 24].

The early identification of a traumatic person's level of risk towards PTSD opens the possibility of preventative intervention.Therefore, the ability to predict risk for PTSD from the time of the trauma is extremely important. Unfortunately, the extant research literature has been unsuccessful in reliably identifying a set of risk factors for PTSD common to all traumatized individuals or specific sets of risk factors that may allow the individualized treatment of a person based on their risk. This limited progress in the field points to the need to identify and apply new methods that might provide improved ways to conduct research towards the reliable and accurate identification of risk factors for PTSD.

2.3 Recent ML-based Research

The early identification of risk identifiers for PTSD is a major clinical and public health challenge. Previously identified risk identifiers are event characteristics [4], peri-traumatic responses [5], early symptoms, early physiological and neuroendocrine responses, gene expression profiles and recovery environment factors. Previous studies have identified risk indicators at the group level, thereby overlooking within-group heterogeneities and distinct individual paths to PTSD that emanate from the disorder's complex multi-causal etiology. As attested by its numerous risk factors, the etiology of PTSD is multi-causal, multi-modal, and complex. Based on the general linear model, statistical methods used were not optimally suited to explore the complex interactions between linear, non-linear or non-normally distributed risk indicators encountered during trauma and its early aftermath. Additionally, the relative contribution of any risk-indicator is necessarily context-dependent and thus does not directly translate across traumatic events and individuals exposed (e.g., female gender increases the likelihood of PTSD among survivors of physical assault, but not in accidents victims). Consequently proper risk assessment defies simple computation and requires knowledge-based, rule driven expert systems.

The latest insights into PTSD analysis using machine learning was done by Kartsoft et al [10] in a study entitled "Bridging a translational gap: using machine learning to improve the prediction of PTSD". The study used data collected for the Jerusalem Trauma Outreach and Prevention Study. Participants were adults (age: 18–70) consecutively admitted to ED following potentially traumatic events (PTEs). The study identified multiple sets of Markov Boundaries (MBs) with equal and exhaustive predictive power. The existence of such large number of MBs may reflect the presumed multi-causal and equifinal etiology of post-traumatic morbidity, which posits many interchangeable contributing factors and many causal pathways. It is also in line with prior evidence of multiplicity of distinct risk indicators of PTSD. The finding from the study extends previous work by translating the previously demonstrated multiplicity of risk indicators into versatile predictive model that can accommodate an array of traumatic situations where one or several known predictors is either unavailable or not contributing significantly. From a practical point of view, such multiplicity points to the potential usefulness of data-informed algorithmic prediction tools to future risk assessments. This work also extends the array of risk indicators identified by earlier studies: Former studies uncovered salient predictors within large groups, whereas this work demonstrated the ability of less consistently predictive, or less frequently recorded features (e.g., expressing a need for help, or ED length of stay) to carry important information. This underscores ML ability to not to reject features that are only weakly, or occasionally correlated with an outcome, and thereby fully extract the informational item of datasets. Within such multiplicity, however, this study identified multiple sets of predictive features that equally and exhaustively predicted non-remitting PTSD. Within such multiplicity, however, this study identified a few consistently predictive features (e.g., those included in over 75% of all MBs). Interestingly, these features comprised, side-by-side, prior variables (e.g., age), event and injury parameters, immediate bodily responses (e.g., ED pain), symptoms (nightmares, loss of concentration, total PTSD and depression symptoms), cliniciansFL observations (e.g., CGI) and more elaborated subjective responses (need for help, sense of worthlessness). Here the MB's predictive accuracy (AUC = .75) does not support a robust prediction from early information collected. This may illustrate the limited predictive power of data features available for this study, all collected within ten days of a traumatic event. Within such limitations, the results of this work still fare well on two accounts: They firstly show the already reasonable ability of simple, non-invasive, inexpensive observations to predict post traumatic morbidity. They additionally establish the usefulness of data features that are regularly collected in ED situations.

Chapter 3

Mental Health Disorders - An Introduction

3.1 Stress-related mental disorders

It has long been understood that exposure to a traumatic event, particularly combat, causes some individuals to display abnormal thoughts and behaviors that we today refer to as mental illness. The first medical diagnosis for a trauma-related disorder is attributed to Swiss military physicians who in 1678 identified a pattern of symptoms caused by exposure to combat which they called "nostalgia". This condition was characterized by melancholy, incessant thinking of home, disturbed sleep, weakness, loss of appetite, anxiety, cardiac palpitations, stupor, and fever. Accounts of trauma-related disorders have historically been linked to warfare.

During World War I "shell shock" was a significant medical and military problem. Symptoms included fatigue, tremor, confusion, nightmares, and impaired sight and hearing. Initially it was believed "shell shock" resulted from physical injury to the nerves due to exposure to heavy artillery bombardment. As medical personnel began recognizing that many soldiers suffered these symptoms without having been on the front lines a greater emphasis was placed on psychological factors as the cause. The onset of World War II once again resulted in a number of soldiers exposed to combat showing anxiety, intense autonomic arousal, flashbacks, and sensitivity to stimuli that reminded them of the original trauma. The term most commonly used for this condition at that time was "combat exhaustion". Unfortunately, while this condition was recognized by medical personnel as psychological in nature, those suffering with the condition were often shamed and reprimanded by their superiors because they were thought to be weak and cowardly.

The DSM-I published in 1952 included the diagnosis of gross stress reaction. The criteria for this disorder were intentionally broad and recognized that exposure to traumatic events other than combat (e.g., natural disaster) could also cause significant distress. Without explanation the DSM-II, published in 1968, omitted the diagnosis of gross stress reaction. The publication of the DSM-III in 1980 saw the inclusion of post traumatic stress disorder (PTSD) as an anxiety disorder for the first time. Subsequent editions of the diagnostics manual (DSM-III-R, DSM-IV, DSM-IV-TR) continued to include PTSD as an anxiety disorder. In the most recent edition of the manual, DSM-5, all trauma- and stressor-related disorders are for the first time grouped together in a single category. The primary trauma- and stressor-related disorders are reactive attachment disorder, disinhibited social engagement disorder, posttraumatic stress disorder, acute stress disorder, and the adjustment disorders.

3.1.1 Characteristic Symptoms

The trauma- and stressor-related disorders are serious psychological reactions that develop in some individuals following exposure to a traumatic or stressful event such as childhood neglect, childhood physical / sexual abuse, combat, physical assault, sexual assault, natural disaster, an accident or torture.

The characteristic symptom of psychological distress resulting from childhood emotional neglect is impairment in the child's ability to relate interpersonally to adults and peers. This symptom is unique to disorders resulting from a pattern of insufficient caregiving that limits an infant's opportunities to form stable attachments. The characteristic symptoms of all other trauma- and stressor-related disorders can be placed into four broad categories: intrusion symptom, avoidance symptoms, negative alterations in cognition and mood, and hyperarousal symptoms.
Intrusion symptoms include recurrent, involuntary and distressing memories, thoughts, and dreams of the traumatic event. The individual may also experience flashbacks, a dissociative experience in which they feel or act as if the traumatic event is reoccurring. Exposure to stimuli or cues that

resemble some aspect of the traumatic event may also result in marked distress for the individual.
Avoidance symptoms are efforts to avoid internal (memories, thoughts, feelings) and/or external (people, places, situations) reminders of the traumatic event. Preoccupation with avoiding trauma related feelings and stimuli can become a central focus of the individual's life.
Negative alterations in cognition and mood include problems remembering important aspects of the traumatic event, depression, fear, guilt, shame, and feelings of isolation from others.
Hyper-arousal symptoms are often reported to be some of the most troubling symptoms of the trauma- and stressor-related disorders. These symptoms include being jumpy and easily startled, irritability, angry outbursts, self-destructive behavior, problems concentrating, and difficulty sleeping.

3.1.2 Types of Stress-related Disorders

Reactive Attachment Disorder (RAD): Reactive Attachment Disorder usually presents before age five and is characterized by serious problems in emotional attachment to others. These children rarely seek comfort when distressed and are minimally emotionally responsive to others. As mentioned above RAD results from a pattern of insufficient caregiving or emotional neglect that limits an infant's opportunities to form stable attachments. Children who are adopted from foreign orphanages are commonly affected, particularly if they were removed from their birth parents during the first weeks of life.

Disinhibited Social Engagement Disorder: Disinhibited Social Engagement Disorder is characterized by a pattern of behavior that involves culturally inappropriate, overly familiar behavior with unfamiliar adults and strangers. Like RAD, this disorder results from a pattern of insufficient care giving or emotional neglect that limits an infant's opportunities to form stable attachments.

Post Traumatic Stress Disorder (PTSD): Post Traumatic Stress Disorder is characterized by significant psychological distress lasting more than a month following exposure to a traumatic or stressful event. Symptoms from all of the categories discussed above must be present.

Acute Stress Disorder: Acute Stress Disorder is similar to PTSD but the duration of the psychological distress last only three days to one month following exposure to a traumatic or stressful event.

Adjustment Disorder: Adjustment Disorders are characterized by the development of emotional or behavioral symptoms in response to an identifiable stress or (e.g., problems at work, going off to college). Adjustment disorder symptoms vary from person to person but must occur within three months of the stressful event and once the stressor has ended the symptoms do not persist more than six months.

3.2 What is PTSD?

PTSD (Post-Traumatic Stress Disorder) is a mental disorder that develops in some people after experiencing or witnessing a scary, dangerous or a life-threatening event, like combat, a natural disaster, a car accident, or sexual assault. It's normal to have upsetting memories, feel on edge, or have trouble sleeping after this type of event. At first, it may be hard to do normal daily activities, like go to work, go to school, or spend time with people you care about. But most people start to feel better after a few weeks or months. If it's been longer for at least one month and you're still having symptoms, you may have PTSD. For some people, PTSD symptoms may start later on, or they may come and go over time.

3.2.1 Factors affecting PTSD Development

PTSD can happen to anyone. It is not a sign of weakness. A number of factors can increase the chance that someone will have PTSD, many of which are not under that person's control. For example, having a very intense or long-lasting traumatic event or getting injured during the event can make it more likely that a person will develop PTSD. PTSD is also more common after certain types of trauma, like combat and sexual assault.

Personal factors, like previous traumatic exposure, age, and gender, can affect whether or not a person will develop PTSD. What happens after the traumatic event is also important. Stress can make PTSD more likely, while social support can make it less likely.

3.2.2 PTSD symptoms

PTSD symptoms usually start soon after the traumatic event, but they may not appear until months or years later. They also may come and go over many years. If the symptoms last longer than four weeks, cause you great distress, or interfere with your work or home life, you might have PTSD. According to CAPS-5 (Clinician Administered PTSD Scale for DSM-5 (The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition)) there are 7 types (groups) of symptom criteria of PTSD diagnosis, summing to a total of 30 items; a necessary combination of a certain set of which reveals the diagnostic status of the person who has faced a traumatic situation. Below is the list of CAPS-5 based set of criterion which have to be assessed by the clinician during the post-traumatic interview session [3].

A. Exposure to actual or threatened death, serious injury, or sexual violence in one (or more) of the following ways:

   1. Directly experiencing the traumatic event(s).
   2. Witnessing, in person, the event(s) as it occurred to others.
   3. Learning that the traumatic event(s) occurred to a close family member or close friend. In cases of actual or threatened death of a family member or friend, the event(s) must have been violent or accidental.
   4. Experiencing repeated or extreme exposure to aversive details of the traumatic event(s) (e.g., first responders collecting human remains; police officers repeatedly exposed to details of child abuse). Note: Criterion A4 does not apply to exposure through electronic media, television, movies, or pictures, unless this exposure is work related.

B. Intrusion symptoms (also called re-experiencing symptoms). Presence of one or more of the following intrusion symptoms associated with the traumatic event(s), beginning after the traumatic event(s) occurred.

   B1. Recurrent, involuntary, and intrusive distressing memories of the traumatic event(s). In Children older than 6 years, repetitive play may occur in which themes or aspects of the traumatic event(s) are expressed.

B2. Recurrent distressing dreams in which the content and/or affect of the dream are related to the event(s). In children, there may be frightening dreams without recognizable content.

B3. Dissociative reactions (e.g., flashbacks) in which the individual feels or acts as if the traumatic event(s) were recurring. (Such reactions may occur on a continuum, with the most extreme expression being a complete loss of awareness of present surroundings). In children, trauma-specific reenactment may occur in play.

B4. Intense of prolonged psychological distress at external to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).

C. Persistent avoidance of stimuli associated with the traumatic event(s), beginning after the traumatic event(s) occurred, as evidenced by one or both of the following:

C1. Avoidance of or efforts to avoid distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).

C2. Avoidance of or efforts to avoid external reminders (people, places, conversations, activities, objects, situations) that arouse distressing memories thoughts, or feelings about or closely associated with the traumatic events(s).

D. Negative alterations in cognitions and mood associated with the traumatic event(s) occurred, as evidenced by two (or more) of the following:

D1. Inability to remember an important aspect of the traumatic event(s) (typically due to dissociative amnesia and not to other factors such as head injury, alcohol, or drugs).

D2. Persistent and exaggerated negative beliefs or expectations about oneself, others, or the world (e.g. ”I am bad”, ”No one can be trusted”, ”The world is completely dangerous”, ”My whole nervous system is permanently ruined”).

D3. Persistent, distorted cognitions about the cause or consequences of the traumatic event(s) that lead the individual to blame himself/herself of others.

D4. Persistent negative emotional state (e.g. fear, horror, anger, guilt, or shame).

D5. Markedly diminished interest or participation in significant activities.

D6. Feelings of detachment or estrangement from others.

D7. Persistent inability to experience positive emotions (e.g. inability to experience happiness, satisfaction, or loving feelings).

E. Marked alterations in arousal and reactivity associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:

E1. Irritable behavior and angry outbursts (with little or no provocation) typically expressed as verbal or physical aggression toward people or objects.

E2. Reckless or self-destructive behavior.

E3. Hypervigilance.

E4. Exaggerated startle response.

E5. Problems with concentration.

E6. Sleep disturbance (e.g., difficulty falling or staying asleep or restless sleep).

F. Duration of the disturbance (Criteria B, C, D, and E) is more than 1 month.

G. The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

G1. Subjective distress.

G2. Impairment in social functioning.

G3. Impairment in occupational or other important area of functioning.

Chapter 4

Algorithms and Tools

4.1 What is Machine Learning?

4.1.1 Introduction

Machine Learning is the field of study that gives computers ability to learn from examples and experience, without being explicitly programmed. Instead of writing code, we can throw data as big as possible at a generic algorithm, and it builds logic based on the data given.

Tom Mitchell provides a more modern definition: *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*

Example: playing chess.

E = the experience of playing many games of chess.

T = the task of playing chess.

P = the probability that the program will win the next game.

4.1.2 Need for Machine Learning

Machine Learning is a field which is raised out of Artificial Intelligence(AI). Applying AI, we wanted to build better and intelligent machines. But except for few mere tasks such as finding the shortest path between point A and B, we were unable to program more complex and constantly evolving challenges.There was a realization that the only way to be able to achieve this task was to let machine learn from itself. This sounds similar to a child learning from its self. So machine learning was developed as a new capability for computers. And now machine learning is present in so many segments of technology, that we don't even realize it while using it.

4.2 Types Of Machine Learning

There are three kinds of Machine Learning Algorithms.

a. Supervised Learning

b. Unsupervised Learning

c. Reinforcement Learning

4.2.1 Supervised Learning

A majority of practical machine learning uses supervised learning. In supervised learning, the system tries to learn from the previous examples that are given. (On the other hand, in unsupervised learning, the system attempts to find the patterns directly from the example given.)

Speaking mathematically, supervised learning is where you have both input variables (x) and output variables(Y) and can use an algorithm to derive the mapping function from the input to the output.
The mapping function is expressed as

$$Y = f(X).$$

Supervised learning problems can be further divided into two parts, namely classification, and regression.

Classification is defined when the output variable is group or class for example 'spam' and 'no spam'.

A regression problem is when the output variable is a real value, such as 'Rupees' or 'height'.

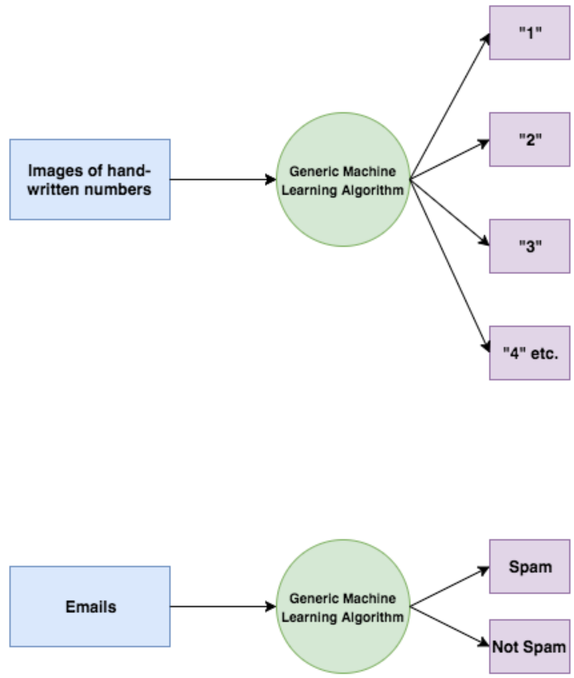


Figure 4.1: Supervised Learning

4.2.2 Unsupervised Learning

In unsupervised learning, the algorithms are left to themselves to discover interesting structures in the data.This is called unsupervised learning because unlike supervised learning above, there are no given correct answers and the machine itself finds the answers.

Mathematically, unsupervised learning is when you only have input data (X) and no corresponding output variables.

Unsupervised learning problems can be further divided into association and clustering problems.

Association

An association rule learning problem is where you want to discover rules that describe large portions of your data, such as “people that buy X also tend to buy Y”.

Clustering

A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

4.2.3 Reinforcement Learning

A computer program will interact with a dynamic environment in which it must perform a particular goal (such as playing a game with an opponent or driving a car). The program is provided feedback in terms of rewards and punishments as it navigates its problem space.

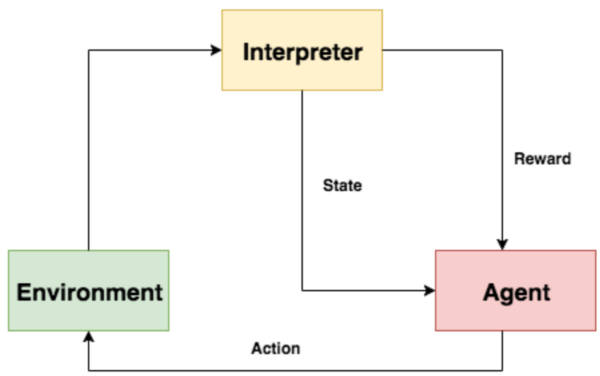


Figure 4.2: Reinforcement Learning

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it continuously trains itself using trial and error method.

4.3 Machine Learning Algorithms

4.3.1 Support Vector Machine (SVM)

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

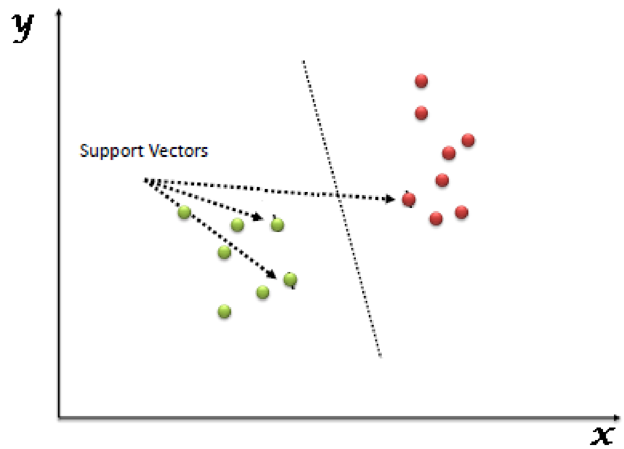


Figure 4.3: Support Vector Machine

in the example shown above, the line which splits the data into two differently classified groups is the black line, since the two closest points are the farthest apart from the line. This line is our classifier. Then, depending on where the testing data lands on either side of the line, that's what class we can classify the new data as.

4.3.2 Causal Discovery Feature Selection - HITON

Most variable selection methods are heuristic in nature and empirical evaluations have seldom exceeded domains with more than a hundred variables. Several researchers have suggested, intuitively, that the Markov Blanket of the target variable T, denoted as MB(T), is a key concept for solving the variable selection problem. MB(T) is defined as the set of variables conditioned on which all other variables are probabilistically independent of T. Thus, knowledge of the values of the Markov Blanket variables should render all other variables superfluous for classifying T.HITON is a a sound, sample-efficient, and highly scalable algorithm for variable selection for classification, and is based on inducing MB(T). In this algorithm, first HITON-MB identifies the parents and children of T by calling algorithm HITON-PC, then discovers the parents and children of the parents and children of T. This is a superset of the MB(T). HITON-PC admits one-by-one the variables in the current estimate of the parents and children set CurrentPC. If for any such variable a subset is discovered that renders it independent of T, then the variable cannot belong in the parents and children set and is removed and not considered again for inclusion. Thus in this way HITON-MB provably identifies the MB(T).

4.4 Matlab

MATLAB (matrix laboratory) is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization and programming.

It allows matrix manipulations; plotting of functions and data; implementation of algorithms; creation of user interfaces; interfacing with programs written in other languages, including C, C++, Java, and FORTRAN; analyze data; develop algorithms; and create models and applications.

It has numerous built-in commands and math functions that help you in mathematical calculations, generating plots, and performing numerical methods.

MATLAB is used in every facet of computational mathematics. From Linear Algebra to Curve Fitting and from Data analysis to 3D modeling almost all the mathematical and statistical facets are modelled in Matlab.

4.4.1 Features of Matlab

- It is a high-level language for numerical computation, visualization and application development.

- It also provides an interactive environment for iterative exploration, design and problem solving.

- It provides vast library of mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration and solving ordinary differential equations.
- It provides built-in graphics for visualizing data and tools for creating custom plots.
- MATLAB's programming interface gives development tools for improving code quality maintainability and maximizing performance.
- It provides tools for building applications with custom graphical interfaces.
- It provides functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET and Microsoft Excel.

4.5 Android Development

Android is a mobile operating system developed by Google, based on a modified version of the Linux kernel and other open source software and designed primarily for touchscreen mobile devices such as smartphones and tablets. In addition, Google has further developed Android TV for televisions, Android Auto for cars, and Wear OS for wrist watches, each with a specialized user interface. Variants of Android are also used on game consoles, digital cameras, PCs and other electronics.

Android is also associated with a suite of proprietary software developed by Google, including core apps for services such as Gmail and Google Search, as well as the application store and digital distribution platform Google Play, and associated development platform.

Chapter 5

Implementation

The methodology of this project involved multiple steps that are summarized as follows:

1. Dataset, which comprises of features that have or seem to have causal influence for PTSD, was collected by us at different psychiatry hospitals and trauma rehabilitation centers in Kashmir including SMHS, IMHNS and Dr Arshid mental health clinic.

2. Data was dummy coded into numerical equivalents, and was represented to accord with ML-based analysis. After this, the missing data entries were filled using Euclidean distance metric.

3. Repeated N-fold Cross Validation acted as the outer-shell algorithm for modal selection and error estimation, the internals of which comprised of feature selection and prediction model construction.The outer loop also acted as the bootstrapping procedure.

4. Feature selection was done here which revealed multiple sets of causal variables at each step of RNNCV.

5. Classification model was build using the features selected by HITON-PC.

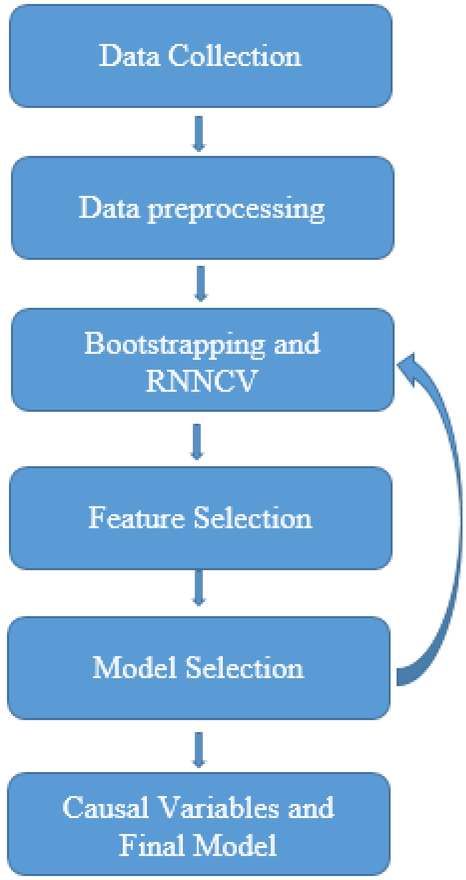


Figure 5.1: Block Diagram - Implementation Procedure

5.1 Data Collection

5.1.1 Dataset

The dataset comprises of 24 samples (less technically patients) as records of causal variables collected by us as part of the clinician-based study we undertook in the valley of Kashmir. The dataset consists of 34 variables constituting the peri-traumating responses arisen after trauma and features covering the previous life of the individual. The feature set is given in Appendix 1. These variables have been selected from studies based on Kashmir and outside. The features are not necessarily in a cause-effect relationship with PTSD but are based on physical significance of variables which have or seem to have causal influence for PTSD. These 'feature' variables belong to such domains as religious and education status, demographics, parent symptoms, stress, magnitude of injury, and peri-traumatic responses. Three types of target variables are specified, all measured more than 1 month after the occurrence of the event. One of the target variables is a CAPS-5 based diagnostic status measured after checking whether various CAPS-5 criterion are met by the patient after being interviewed. Second, the symptom severity score, calculated by summing the scores for each symptom in different CAPS-5 criterion groups and checking for PTSD or NOPTSD case by comparing with the Reaction Index score of 40 (which is the mean value of CAPS-5 severity score). Third, a target outcome calculated from both cases 1 and 2, is taken as the third target variable for analysis. The cutoff score 40 is based on a high level of symptoms of PTSD and is strongly related to a DSM-5 diagnosis of PTSD.

5.1.2 Data Preparation

A data sample consisting of feature and target variables was collected through the following procedure: The dataset variables were recorded at first based on the questionnaire we have built, followed by target variables which were calculated from CAPS-5. Prior to analysis, categorical variables were dummy coded and all continuous variables were normalized to have a range from 0–1 to reduce noise due to differences in scaling. Missing data was handled using the knnimpute command in MATLAB. To perform imputation in data sets with missing values, we applied a non-parametric nearest neighbor method. Specifically, this method imputes each missing value of a variable with the present value of the same variable in the most similar instance according to Euclidian distance metric.

5.2 Machine Learning

Machine Learning in this work involved multiple recursive steps that together informed the identification of potentially predictive features and the evaluation of the predictive accuracy of selected features using classification algorithms.

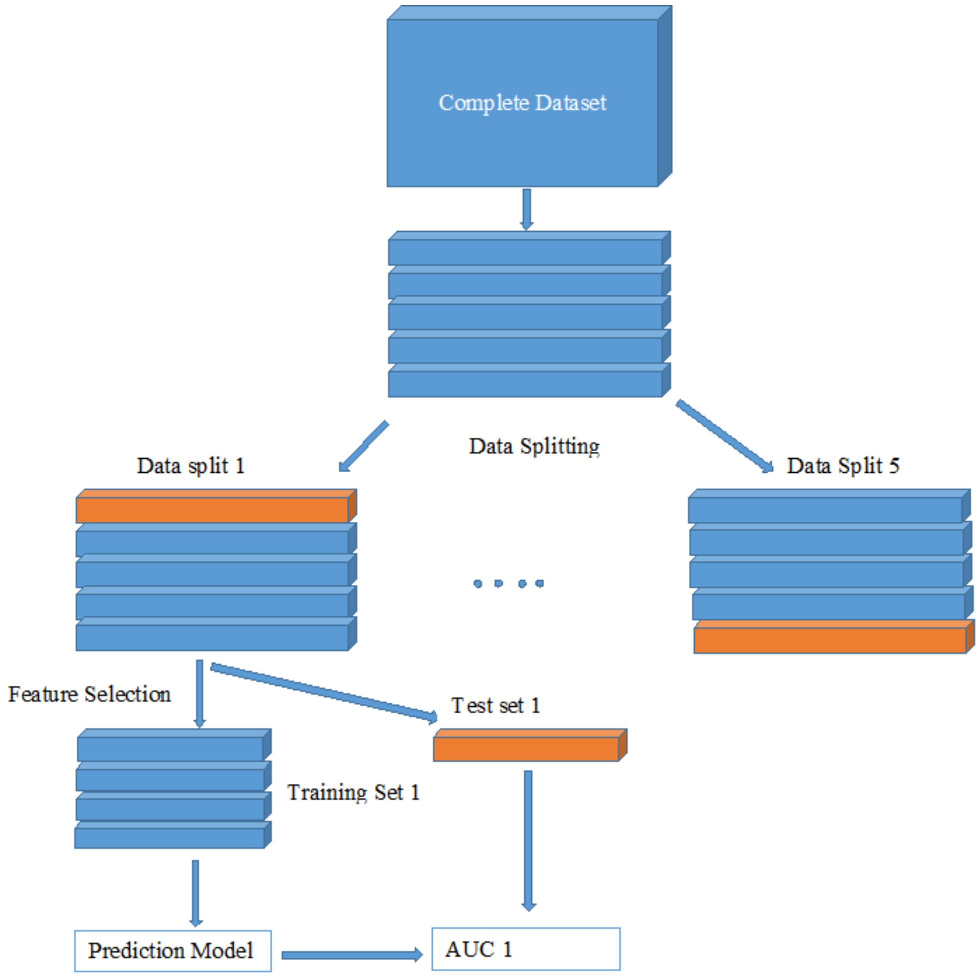


Figure 5.2: Complete model of the Algorithm

To examine different ways of optimizing an ML-based approach to the particular case of forecasting PTSD we repeated these steps using different predictive features and outcome definitions. The reliability of each choice was tested using cross-validation.

5.2.1 Feature Selection

We applied a Markov Boundary Induction algorithm (HITON-PC) to identify sets of features that provided the most direct predictive power of our target variables. The classification model is built with variables selected by the HITON-PC Feature Selection method rather than from all features in the data set. HITON-PC performs causal discovery through learning local causal networks and approximating Markov Boundaries. A Causal Network is a graph that represents the causal relations between a set of random variables. In this study, a causal network would represent the causal relations from sets of features in the data set and our response variable (the "target") – whether a person would have a CAPS-5 score of 40 or greater. This approach initially identifies variables that demonstrate a univariate association to the target variable and removes all others. It then tests each retained variable's association with the target variable while controlling for other retained variables and excludes those that become non-significant when controlling for other variables. The resulting list consists of predictors that are independently associated with the target. We used an algorithm from the Causal Explorer library.

5.2.2 Classification Algorithm

In the current analysis we utilize Linear Support Vector Machines (SVMs) with the C parameter = 1. SVMs identify a linear hyperplane in high dimensional space (each predictor variable is a dimension) that accurately separates the sample into classes based on the target.

5.3 Cross-Validation

To avoid overfitting solutions and ensure the generalizability of our findings, all feature selection and classification experiments were conducted in a 30 times repeated 5-folds cross-validation. In the 5-folds cross-validation, participants are randomly assigned into 5 non-overlapping subsets containing approximately the same number of cases and non-cases. The classification algorithm is trained in nine of these ten data subsets, and subsequently (and independently) tested in the remaining tenth subset. This procedure is repeated iteratively, resulting in all tenths of the data being used for both training and testing of the algorithm. The entire procedure is repeated 10 times, resulting in a total of 150 runs (30 repetitions of 5 trainings and testing). In the current study, a cross-validation algorithm was written in MATLAB R2015 to randomize cases into 90%/10% splits. The first 90% of the data was utilized in the feature selection algorithm and the best solution in that 90% was tested in the hold out 10%. This procedure was repeated 10 times. Features that were selected and confirmed across random splits of the sample were then introduced into the ML classification algorithms. Once again, data was randomly split using the same cross-validation procedure. For each classification algorithm, the solution was identified in a random 90% of the data and validated in the hold out 10%. This procedure was also repeated 10 times.

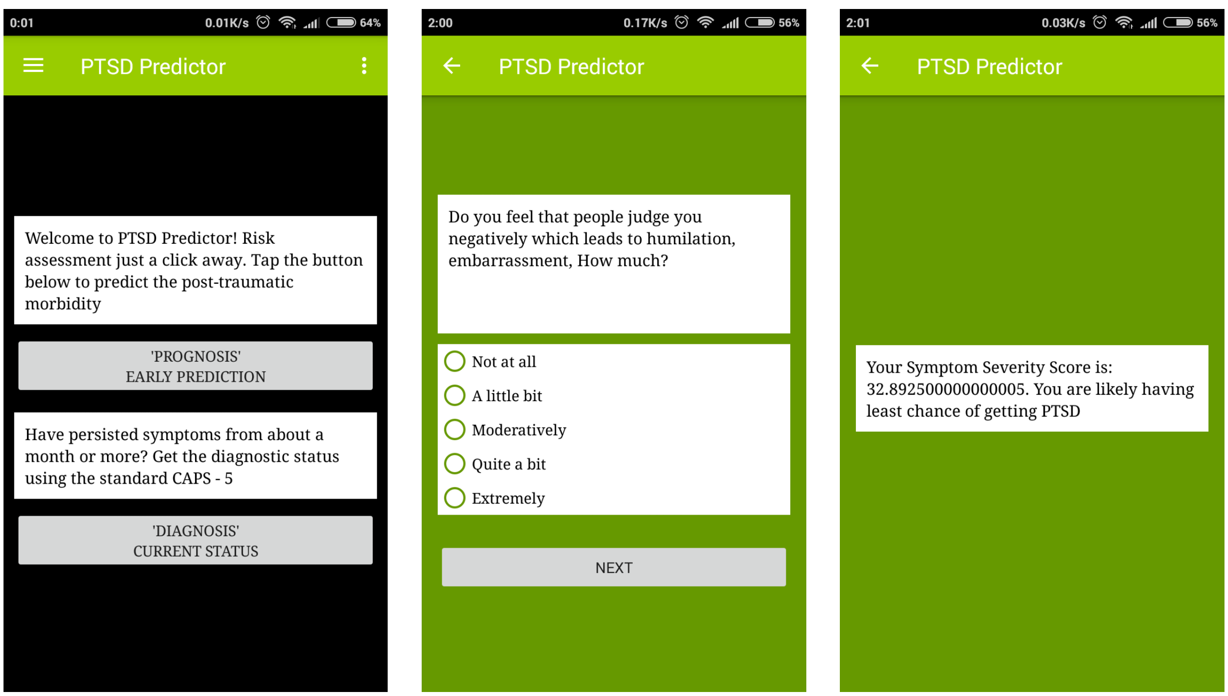
5.4 Accuracy Metrics

Estimates of predictive accuracy are expressed as Area Under the Receiver Operating Characteristics curve (AUC). The ROC curve is a plot of the sensitivity versus 1-specificity of a classification system, and infers the accuracy of that system, thereby creating a comparable metric across experiments. Following literature standards, we consider ROC curve AUC of .50–.60 as indicating prediction at chance; .60–.70 as indicating poor prediction; .70–.80 fair prediction; .80–.90 good prediction; .90–1.0 excellent prediction.

Chapter 6

PTSD Predictor - an Android app

We have developed an Android application which acts as an early on PTSD predictor. The app has been developed from the results obtained in the bootstrapping procedure i.e from the stable most causal features selected, therefore can be used anywhere, anytime to determine PTSD symptom severity of an individual who faces any traumatic situation.



(a) Android app - main page (b) Android app - a sample question (c) Android app - final result screen

Figure 6.1: App Screenshots

Chapter 7

Results

7.1 Performance of Prediction model

The performance of prediction model was measured using Area Under Receiver Operating Characteristic Curve (AUC). The classification model build from all the features had an AUC value of 0.84. This is an improvement to the previously determined accuracy of 0.79 [11]. The model build from the selected features revealed different AUCs for different target outcomes. The best model was selected when both symptom severity and diagnostic status were used for building the outcome variable. As seen from the figure, AUC in the best case came out to be 0.76.

The results reveal that the model with considerable predictive accuracy is the one build with target variable taken as a combination of both Diagnostic status and Symptom Severity. This in turn is due to fact that there lies a non-linear correlation between the two variables; That is the

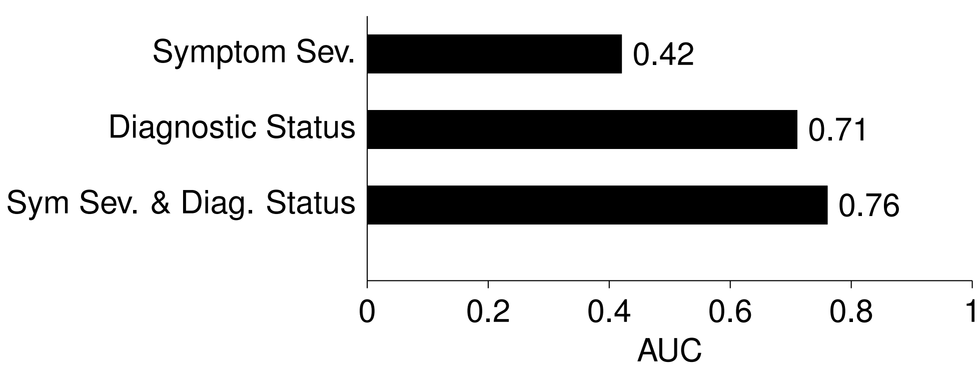


Figure 7.1: AUC against different target variables

Reaction Index score of 40 or more (using CAPS - 5 assessment) doesn't translate directly to the diagnosis of non-remitting PTSD diagnostic status (PTSD's presence). Further these values are stronger than the performance yielded by the same model in other studies.

7.2 Identification of most stable causal features

Considering all the cases, a maximum of 12 variables out of 35 were selected in 150 (5*30) bootstrapped samples. 8 features were selected in the best case (discussed above). The three cases are summarized as follows.

7.2.1 Case 1: Diagnostic status as target variable

Diagnostic status is a PTSD - NOPTSD signal determined using some standard questionnaire like UCLA RI, CAPS V, STQ etc. In this study we have used CAPS V as the standard measure for revealing the diagnostic status. The AUC in this case was 0.71 and a compact set of 12 features were selected as causal variables.

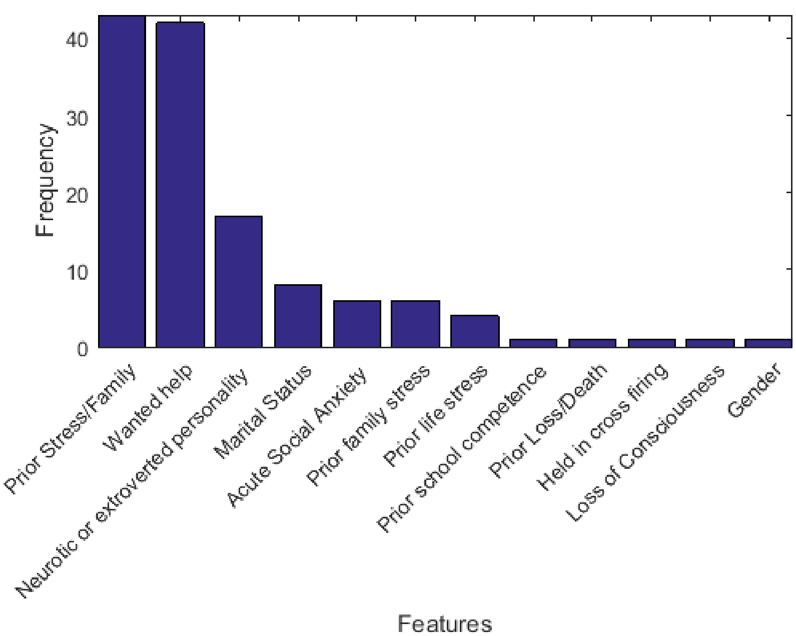


Figure 7.2: Frequency of selected features across 150 bootstrapped samples - Diagnostic Status as target variable

7.2.2 Case 2: Symptom Severity as target variable

Symptom Severity is total sum of intensity values of symptoms spanning different categories of CAPS - 5 assessment tool. An intensity value which is clinically significant is taken as input from each symptom towards the calculation of Reaction Index. The RI Score comes out to be 40 in this case, that is a score above 40 counts towards PTSD's presence and vice versa. Since there lies a gap between Symptom Severity and Diagnostic Status and because of the multi-causal nature of the problem, The AUC value comes out to be lowest in this case i.e. 0.45.

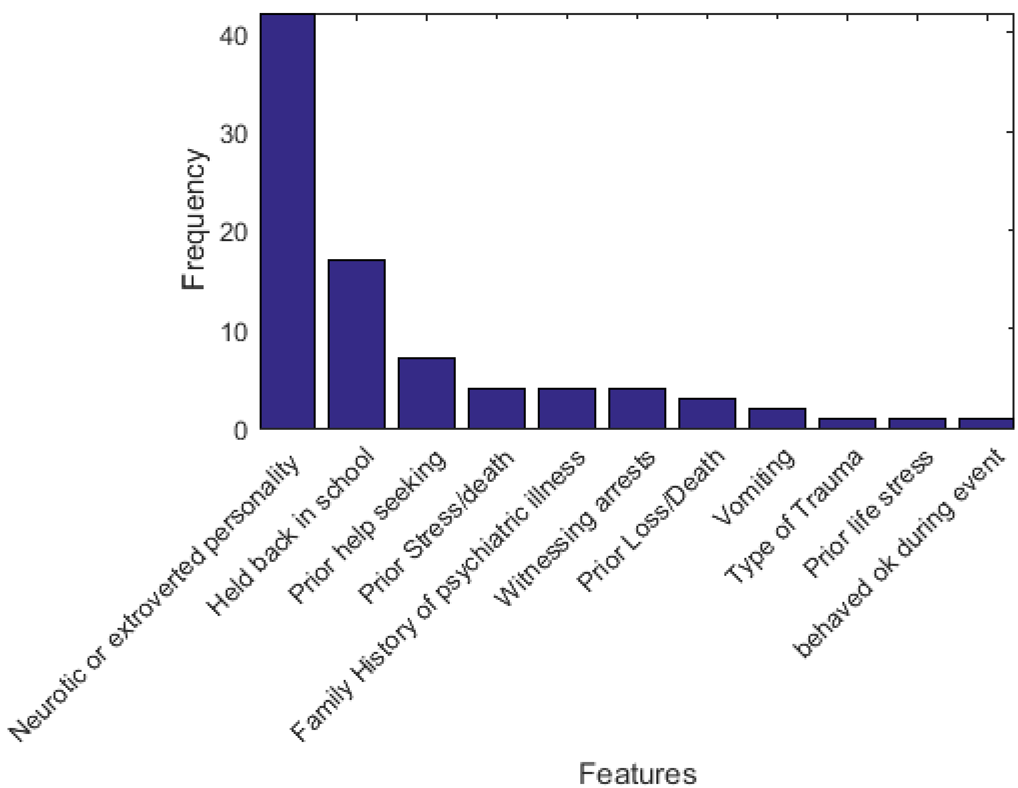


Figure 7.3: Frequency of selected features across 150 bootstrapped samples - Symptom Severity as target variable

7.2.3 Case 3: Both SS and DS as target variable

Target variable is the hybrid of Symptom Severity and Diagnostic Status. Both breadth and depth of severity is taken into consideration. Symptom Severity may have a well-nigh influence on the target variable by highlighting the more severe symptoms or symptom categories while as Diagnostic Status considers the simultaneous effect of multiple symptom categories. Thus better results expected here. The AUC score is 0.76 in this case thus meets the expected better performance than the other two.

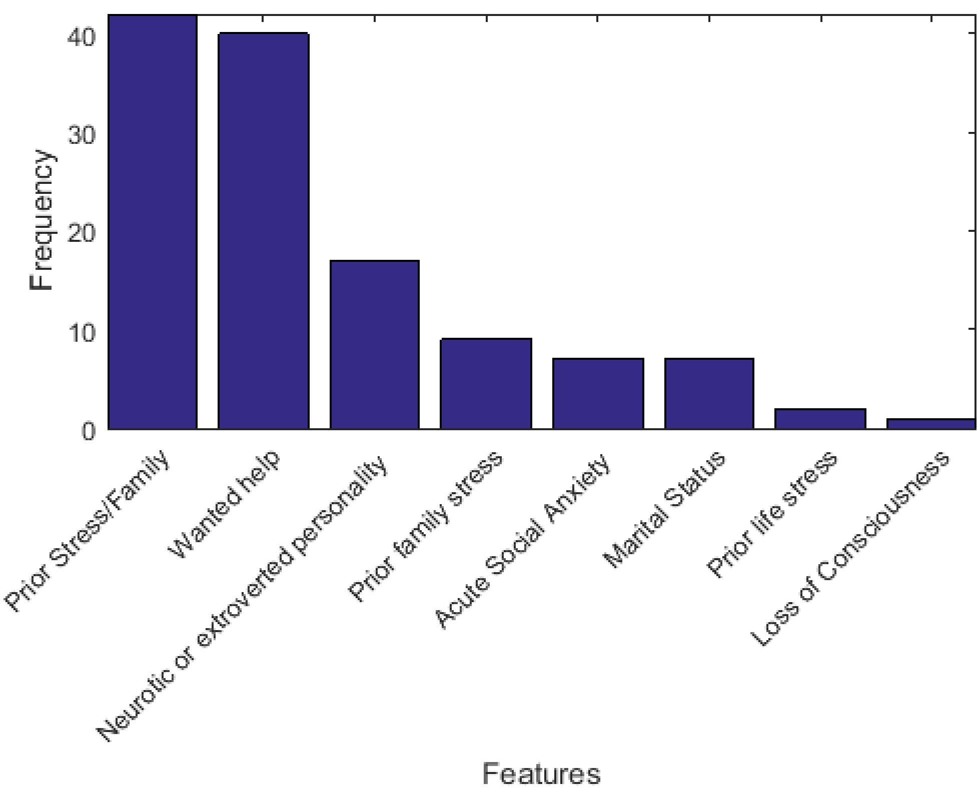


Figure 7.4: Frequency of selected features across 150 bootstrapped samples - Both Diagnostic status and Symptom Severity as target variable

7.3 Discussion

The AUC values prove that predictive models build from features selected by HITON-PC are comparatively similar in performance in relation to models constructed from all the variables. Also, the theoretical correctness of HITON-PC and the absence of ‘spouses’ proves that the variables selected are true causes of the target variable. The features thus selected can be taken for analysis and can be acted upon for preventative intervention, although a lot of research is further expected in this field to meet the predictive optimality of the discovered variables.

All the three cases revealed comparable results despite the predictive accuracy differences. In this section, we will try to discuss the most frequent variables which come under the three cases as discussed above.

7.3.1 Prior Stress/family

This variable came out to be the most frequent one in both Cases 1 and 3. Prior Stress/family is the level of stress a person had, prior to the happening of the event, regarding the family condition or the stress emanating from the stress of the family itself. This is consistent with the earlier studies [10, 11] although not as the most significant one. The tormented society which has arisen out of the long-lasting conflict in Kashmir has severely affected the mental, social and economical aspects of the society. It has lead the Kashmiri families mutilated and has caused a colossal damage to the the mental health of the society in general.

7.3.2 Wanted Help

This feature came out to be the second most frequent one in both Cases 1 and 3. ”Wanted Help” implies the need of help a person felt immediately after or within one week of the post-trauma period. This feature was not identified among the most stable causal variables in other studies. Its stronger influence in contrast to other variables may be attributed to the severity of trauma we have been facing since a long time ago i.e. most of the trauma patients in Kashmir are conflict-related. Other reasons may be associated with it in which we expect further understanding and research.

7.3.3 Neurotic or Extroverted Personality

The unreasonable worries a person had had prior to the occurrence of the event or the introvert behavior he had had prior to the happening of this trauma. The introvert behavior has previously been identified as a risk indicator for a large number of mental health disorders including depression and anxiety. Hence with the help of this study, the result extends to other disorders including Acute Stress Disorder (ASD) and Post Traumatic Stress Disorder (PTSD).

7.3.4 Held back in school

A person who couldn't perform well in the school is more likely to develop PTSD in contrast to the one who shows some academic brilliance. This is consistent with the previous studies [10, 11].

7.3.5 Prior Help Seeking

This variable is paradoxical for being the risk indicator as it was thought to be an intervention in case of PTSD. A person seeking help from others other than doctors, is thought to be suffering from a psychiatric problem, though the issue may be minute as negligible. Parents and close relatives need to take care of the person who seeks health advice from his relatives, friends, peers and hakims.

Chapter 8

Conclusion and Future work

The current study offers new heuristic for PTSD prediction and analysis and therefore paves ways for improved analytical studies regarding this complex disorder. The results obtained in this study are not far from theoretical expectations and therefore support the fact that PTSD can be predicted to a significant degree from the data available shortly after trauma. The study also posits new insights into PTSD prediction in particular and psychiatry in general by integrating causal discovery feature selection and classification models and therefore acts as a potential framework for future studies. The causal features selected offer a new promise for preventative intervention and therefore represent a compact set of variables to be taken care of. The study further suggests that ML-based techniques have significant potential for enhancing upon the methodologies previously or currently used in analyzing this mental health disorder.

The results obtained especially in case 1 represent a compact causal model fro PTSD analysis in Kashmir. Nonetheless, a lot more needs to be done to improve the accuracy of the said model or a better model can be devised with stronger predictive optimality. A recent study suggests that Generative Adversarial Network (GANs) - based predictive model outperforms the currently better performing ones like SVMs and therefore can be worked upon in future to replace the currently used ones. Data set can be enhanced horizontally (enhancing the sample size) as well as vertically (enhancing the feature size) to improve upon the accuracy of model as well as feature selection. Since this was a clinician-based survey, the results don't fully translate across this geographic region, because people approaching medical guidance and assistance represent only the tip of the iceberg, the rest of the junk lies within the depths of the society. Thus a community-based study can be done to converge fully to the exact set of features and predictive models.

Bibliography

[1] Mohd. A. Margoob, Sheikh Aijaz Ahmad. Community prevalence of adult post traumatic stress disorder in south asia: experience from Kashmir. *Psychol Med.* 2006;38(4):533–42.

[2] MSF, IMHNS. Kashmir Mental Health Survey. 2015;14:499–566.

[3] Weathers, F. W., Blake, D. D., Schnurr, P. P., Kaloupek, D. G., Marx, B. P., & Keane, T. M. (2015). The Clinician-Administered PTSD Scale for DSM-5 (CAPS-5) – Past Month / Worst Month [Measurement instrument]. Available from *http://www.ptsd.va.gov/*

[4] Ozer EJ, Best SR, Lipsey TL, Weiss DS. Predictors of posttraumatic stress disorder and symptoms in adults: A meta-analysis. *Psychol Bull.* 2003;129(1):52–73.

[5] Marmar CR, Weiss DS, Schlenger WE, et al. Peritraumatic dissociation and posttraumatic stress in male Vietnam theater veterans. *Am J Psychiatry.* 1994;151(6):902–7.

[6] Bryant RA, Harvey AG, Guthrie RM, Moulds ML. A prospective study of psychophysiological arousal, acute stress disorder, and posttraumatic stress disorder. *J Abnorm Psychol.* 2000;109(2):341–4.

[7] Shalev AY, Peri T, Brandes D, Freedman S, Orr SP, Pitman RK. Auditory startle response in trauma survivors with posttraumatic stress disorder: a prospective study. *Am J Psychiatry.* 2000;157(2):255–61.

[8] Segman RH, Shefi N, Goltser-Dubner T, Friedman N, Kaminski N, Shalev AY. Peripheral blood mononuclear cell gene expression profiles identify emergent post-traumatic stress disorder among trauma survivors. *Mol Psychiatry.* 2005;10(5):500. –513, 425.

[9] Koenen KC, Stellman JM, Stellman SD, Sommer Jr JF. Risk factors for course of posttraumatic stress disorder among Vietnam veterans: a 14-year follow-up of American Legionnaires. *J Consult Clin Psych.* 2003;71(6):980–6.

[10] Karstoft KI, Galatzer-Levy IR, Statnikov A, Li Z, Shalev AY. Bridging a translational gap: using machine learning to improve the prediction of PTSD. *BMC Psychiatry.* 2015;15(1):30.

[11] Glenn N. Saxe, Sisi Ma, Jiwen Ren and Constantin Aliferis. Machine learning methods to predict child posttraumatic stress: a proof of concept study. *BMC Psychiatry.* 2017:17(1):20.

[12] Steinberg AM, Brymer MJ, Decker KB, Pynoos RS. The University of California at Los Angeles post-traumatic stress disorder reaction index. *Curr Psychiatry Rep.* 2004;6(2):96–100.

[13] Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos X. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I:algorithms and empirical evaluation. *J Mach Learn Res.* 2010;11:171–234.

[14] Statnikov A, Tsamardinos I, Brown LE, Aliferis CF. Causal Explorer: A Matlab Library of Algorithms for Causal Discovery and Variable Selection for Classification. *Causation and Prediction Challenge Challenges in Machine Learning, Volume 2, 267.*

[15] Kessler R C, Sonnega A, Bromet E et.al. Post Traumatic Stress Disorder in the National Co morbidity Survey. *Arch Gen Psychiatry* 1995; 52:1048-1060

[16] Xingdong Wang, Langao M.D. Longitudinal study of earthquake related PTSD in a randomly selected community sample in North China. *Am Jour Psych*, 2000, 157:1260-1266

[17] Kulka R.A, Schlenger W.E, Fairbank J.A. Trauma and Vietnam War generation. *Report of the findings from the National Vietnam Veteran, Readjustment study.* NY Brunner/Mazel, 1990

[18] Weiss D.S, Marmar C.K, Schlenge W.E. The Prevalence of lifetime and Partial PTSD in Vietnam theatre veterans. *J. Traumatic stress* 1992; 5: 365-376.

[19] Basoglu M, Parker M, Parket O, et al. Psychological effects of torture: a comparison of tortured with nontortured political activists in Turkey. *Am J. Psychiatry*, 151: 76-81, 1994.

[20] Kluznik J.C, Speed N., Van Valkenburg C et.al. Forty year followup of United States prisoners of war. *Am J. Psychiatry* 1986; 143; 1443-1446.

[21] Goldstein G., Van Kammon W., Shell C. Survivor of improvement in pacific theatre during WW-II. *Am J.Psychiatry* 1987, 144; 1210-1213.

[22] benhaim L. Dab. Salmi LR. Study of civilian victims of terrorist attacks in France 1982-1987. *J Clin Egidemo* 145: 103-109. 1992.

[23] Curran PS. Bell P. Murray A. et al. Psychological consequences of Enniskillen bombing. *Br J Psychiatry* 156: 479-482.1990.

[24] Eiscisenbruch M. From Post Traumatic Stress Disorder of cultural bereavement: Diagnosis of South East Asian Refugees. *Soc. Sci Med.*1991; 33(6):573-680.

[25] Lisa J. Mull, Social Anxiety and Introversion in College Students. 7-24-2006 *http://commons.pacificu.edu/spp/37*

Appendices

A. Code Implementation

```
function [indexOfFeatures] = mainFile(feature_selection_name)
warning('off','all');
target_variable_index=35;
% Import data
data_file = 'newdset_ptsd.txt';
textFeatures = {'Gender', 'Age of Trauma','Marital Status',...
    'People in home','Socio-economic status',...
    'relatives in event', 'behaved ok during event',...
    'Loss of Consciousness', 'Vomiting', 'Injury Severity ',...
    'Head Injury', 'Whiplash Injury',' Wanted help',...
    ' Social support', 'Prior multiple Traumatic events faced',...
    'No. of times beaten up', 'Killing of the close relative',...
    ' Torture of a close relative', 'Held in cross firing',...
    'Witnessing arrests', ' Prior family stress',...
    'Family History of psychiatric illness',...
    'Prior life stress', 'Prior Stress/death',...
    'Prior Stress/Family', 'Prior Loss/Death',...
    'Prior help seeking', ' Acute Social Anxiety ',...
    'Neurotic or extroverted personality ',...
    'Acute Dissociation', 'Prior school competence', ...
    'Held back in school' ,'Type of Trauma',...
    ' Prior religious service'};
data_orig =importdata(data_file);
invData = data_orig';
imputedData = (knnimpute(invData))';
[r, c] = size(imputedData);
data = imputedData(:,1:c-2);
```

```
target = imputedData(:,c-1);
targetSev = imputedData(:,c);
%data = normalize(data);
normAndImputeddata = [data target];
normAndImputeddataSev = [data targetSev];
freq = zeros(1, c-1);
%freqnew = zeros(1, c-1);

temp = target;

     %target = newData(:, c);
        for k=1:r
          if targetSev(k) >= 40
            newtarget(k) = 1;
          else
            newtarget(k) = 0;
          end
        end

        targetSev = newtarget';

        for k=1:r
          if (targetSev(k) < 40) && (target(k) == 0)

            newtarget(k) = 0;
          else
            newtarget(k) = 1;
          end
        end
```

```
       targetTotal = newtarget';

       %newData = [newData(:, 1: c-1) target];
       normAndImputeddataSev = [data targetSev];
       normAndImputeddataTotal = [data targetTotal];



for j = 1: 30
 dataSplits = cvpartition(target ,'KFold', 3);
 for i=1:3


        %normAndImputeddata = [data temp];
        % Defining training and testing indices
        TEST_indx = dataSplits.test(i)
        TRAIN_indx=dataSplits.training(i);

        trainData = normAndImputeddata(TRAIN_indx,:);
        testData = normAndImputeddata(TEST_indx,:);
        trainDataSev = normAndImputeddataSev(TRAIN_indx,:);
        testDataSev = normAndImputeddataSev(TEST_indx,:);
        trainDataTotal = normAndImputeddataTotal(TRAIN_indx,:);
        testDataTotal = normAndImputeddataTotal(TEST_indx,:);

        %training data and test data arranged in order
        newData = [trainData;testData];
        newDataSev = [trainDataSev;testDataSev];
        newDataTotal = [trainDataTotal;testDataTotal];
```

```
% Redefining training and testing indices
TRAIN_indx=1:(size(trainData,1));
TEST_indx=(size(trainData, 1)+1):r;



% Performing feature selection

 % Input feature_selection_name 'HITON_PC'
 ...if wish to apply parent/children feature selection...
if strcmp(feature_selection_name,'HITON_PC')
    features = Causal_Explorer('HITON_PC',...
        trainDataTotal, target_variable_index, [], 'z', 0.05, 1)
    % Ipput feature_selection_name 'HITON_MB' if...
    ...wish to apply Markov Blanket feature selection
elseif strcmp(feature_selection_name,'HITON_MB')
    features =Causal_Explorer('HITON_MB', trainData,...
    target_variable_index, [], 'z', 0.05, 1);
end


 if ~(isempty(features))

     [label, score]=code_SVM(newDataTotal(:, features),...
     newDataTotal(:,target_variable_index), TRAIN_indx,...
 TEST_indx, 'linear');
     size(label)
     size(score)
     target(TEST_indx)
     [~,~,~,auc(i, j)] = perfcurve(newDataTotal(TEST_indx,...
```

```
            target_variable_index),score(:,2), 1);

            %Count freq of features in CV iterations
            noOfFeatures = size(features)
            for featureIndex = 1: noOfFeatures(2)

               freq(features(featureIndex)) = ...
               freq(features(featureIndex)) + 1;
               % fr = freq(1, features(featureIndex)) % For testing purpose

            end
            fprintf('\nAUC for %g cross-validation' ...
            'trial: %g)\n', j, mean(auc(:, j)));



        end



 end

mperf=mean(mean(auc));
fprintf('\nOverall performance for selected features: %g)\n',mperf);

%Neglect features with zero frequency
i = 1;
 for p = 1: c-1
     if freq(p) > 0
         vecOfFreq(i) = freq(p);
         indexOfFeatures(i) = p;
```

```
        netTextFeatures(i) = textFeatures(p);
        i = i + 1;

    end
end

%Plot table and bargraph of features selected

table(indexOfFeatures(:),vecOfFreq(:),...
'VariableNames',{'Feature','Frequency'})
for i = 1:(size(vecOfFreq, 2) - 1)
    for j = (i + 1):size(vecOfFreq, 2)

        if vecOfFreq(i) <= vecOfFreq(j)
            temp = vecOfFreq(i);
            vecOfFreq(i) = vecOfFreq(j);
            vecOfFreq(j) = temp;
            temp = indexOfFeatures(i);
            indexOfFeatures(i) = indexOfFeatures(j);
            indexOfFeatures(j) = temp;

            temp = netTextFeatures(i);
            netTextFeatures(i) = netTextFeatures(j);
            netTextFeatures(j) = temp;
        end

    end
end
table(indexOfFeatures(:),vecOfFreq(:)...
,'VariableNames',{'Feature','Frequency'})
```

```
 %vecOfFreq = sort(vecOfFreq);
 bar( vecOfFreq)
 set(gca,'XTick',1:numel(indexOfFeatures))
 set( gca,'xtickLabel',netTextFeatures, 'XTickLabelRotation',45)
 %set(gca, 'xtickrotation', 45)
 %title('Symptom Severity and Diagnostic Status as target variable')
 xlabel('Features')
 ylabel('Frequency')
 axis tight
end
```