ABSTRACT

Early developmental disorders are common in children between the ages of 3 through 17. These developmental disorders begin at early ages and affect the day-to-day activity of children. They also impact the growth and lifestyle of children. Most of the time these developmental disorders co-exist in children. The main focus of our research lies in the disorders 'Autism Spectrum Disorder', 'Attention-Deficit/Hyperactivity Disorder', 'Deletion syndrome (22q)' and their co-occurrences.

Most child psychologists and pediatricians diagnose these disorders in children through parent-oriented reviews. Our research uses three different parent-oriented reviews, which are 'Autism Diagnostic Interview', 'Behavioral Assessment Schedule for Children', and 'Vineland Adaptive Behavior Scales'. These reviews are questionnaires that parents answer under the inspection of certified professionals. While these examinations take up lots of time and yield results after at least 13 months of wait time, the process needs to speed up and hence, machine learning could play a vital role in this process.

Machine learning, when applied to reviews, can help understand the relevance and importance of them in diagnosing the disorders. Also, many techniques have been applied in our research to evaluate the co-occurrence of these disorders. The objective is to determine if machine learning could predict the occurrence of these disorders. Also, decide the significance of these reviews using feature selection algorithms of machine learning.

Computational Analysis of Developmental Disorders in Children

Siri Chandana Sambatur

Bachelor of Technology

VNR Vignana Jyothi Institute of Engineering and Technology

Telangana, (India) 2016

DISSERTATION

Submitted in partial fulfillment

of the requirements for the degree

Master of Science in Computer Science

May, 2018

Syracuse University

Syracuse, New York

Acknowledgments

I am grateful to my advisor, Dr. Reza Zafarani of the Department of EECS at Syracuse University. He guided me throughout my research and helped me overcome various obstacles that I faced. Dr. Zafarani directed me in the right path during my research and advised of various prospective areas to explore. Along with Dr. Zafarani, Dr. Russo helped me understand the psychological aspects of my research. I am thankful to them for the encouragement and constant support.

I would also like to thank the experts who were involved in the Defense Committee without whose participation, the Defense could not have been successfully conducted:

- Dr. Natalie Russo, Assistant Professor, Department of Psychology
- Dr. Senem Velipasalar Gursoy, Associate Professor, Department of Electrical Engineering and Computer Science
- Dr. Edmund Yu, Associate Professor, Department of Electrical Engineering and Computer Science

I would like to give credit to my parents Mamatha and Gowri Shankar Sambatur, my brother Abhishek Shankar for always motivating me. Also, I would like to thank my friends Sai Venkat Kotha, Ayesha Ahmad and Rashmi K. Shivanna, who supported and rooted for me during this journey. Finally, I am also grateful to my laboratory mates who assessed and advised on various aspects of my thesis. This accomplishment would not have been possible without all these people.

Sambatur, Siri Chandana

Syracuse, New York

May, 2018

Contents

List of Figures

List of Tables

List of Abbreviations

ABIDE Autism Brain Imaging Data Exchange

ADHD Attention Deficit/Hyperactivity Disorder

ADI Autism Diagnostic Interview

ADI-R Autism Diagnostic Interview Revised

ASD Autism Spectrum Disorder

BASC Behavioral Assessment Schedule for Children

IQ Intelligent Quotient

LASSO Least Absolute Shrinkage and Selection Operator

ML Machine Learning

PCA Principal Component Analysis

RFE Recursive Feature Elimination

RMSE Root Mean Square Error

ROC Receiver Operator Characteristic

sklearn scikit-learn package

SVM Support Vector Machine

VCFS 22q Deletion Syndrome

VINE Vineland Adaptive Behavior Scales

Chapter 1

Introduction

Most children suffer from different kinds of developmental disorders. Among the many disorders present, the most common ones are ASD, ADHD and VCFS. While VCFS is a genetic defect, the causes of ASD and ADHD are not known. There various studies in this field to identify the genetics behind these two disorders. Also, few children suffer from more than one disorder and sometimes one leads to another. Therefore, it is essential to understand as much about these developmental disorders as possible.

Clinicians take almost 13 months to diagnose children and even later in the case of comorbid disorders. Many researchers have developed techniques for early intervention of these disorders. As these disorders effect the trajectory of growth of children, early intervention is critical for children and families to led normal lives. A lot of researchers believe that machine learning could speed up this process and play a vital role in early diagnosis. So, the main aim of our research is to analyze these disorders, understand them better and try to build models for prediction of these disorders. Our key contributions are as follows-

- Identifying important features for developmental disorders ASD, VCFS and ADHD, along with the comorbid disorders.

- Assessing the impact of the three different parent-oriented reviews on these develop-

mental disorders and their comorbidity.

- Designing models to diagnose these different subgroups of developmental disorders present in the data
- Designing models to diagnose for the comorbid disorders 'ASD and VCFS' and 'ASD and ADHD'
- Designing models for each individual diagnosis that is ASD, VCFS and ADHD.
- Formulating hypothesis using the features relation found in our models to assess the developmental disorders

Initially, in our thesis, more information is provided on these developmental disorders and their comorbidity. The parent-oriented reviews which will be used for our data are explained. Also, in chapter 2, the various models that have been designed using machine learning to diagnose these disorders are explained. Before applying machine learning on our data, in chapter 3, the data is analyzed and preprocessed. Furthermore, analysis and observations are made on the significant features related to these developmental disorders. Later on in chapter 4, understanding of various subgroups in our data is done using both unsupervised and supervised techniques. Additional information on the comorbidity of the disorders is examined in chapter 5 and finally, in chapter 6, each of these individual developmental disorders are investigated.

Chapter 2

Related Work

Developmental Disorders comprise of a group of psychiatric conditions that originate in childhood and cause severe problems in certain areas. According to the Centers for Disease Control and Prevention (CDC), more than one out of every 100 school children in the United States has some form of mental retardation. The third most common disorder is Autism Spectrum Disorder. Developmental disorders are caused because of various reasons like genetics (Deletion syndrome) or due to complications in pregnancy or birth. However, most of the times the exact causes are not known. It can also be seen that developmental disorders are common occurrences in children when the mother goes through stress or consumes alcohol during pregnancy [1]. For proper livelihood of the child, early identification of these disorders is crucial as their causes could be reversed. Most children who were diagnosed with ASD usually reported initial signs of concern before the child was 2 years old in their records. However, most of these children were diagnosed only after age 4, almost 82% of the time by age 3 and 21% of the time at 8 years ages[2]. Early identification helps families to prepare strategies and other resources for successful families.

The occurrence of multiple disorders causes more complexity in identifying them as one of those disorders would be diagnosed and the other would take more time to identify. Children who are diagnosed with ASD could also be suffering from ADHD or vice-versa. The

relationship between these two disorders is still debated in the field of psychology. While children suffering from VCFS could also be affected with Schizophrenia. It was observed that 2 in every 178 patients suffering from Schizophrenia had VCFS [3]. Also, a study done with data collected from a congenital cardiac clinic reported that 22.6% of adults with VCFS had schizophrenia or schizoaffective disorder [4].

Machine learning could help build models to analyze these causes much faster and help in speeding up the diagnosis process. With this research, we hope to identify the causes for single or multiple disorders with our models with good accuracies and also differentiate subjects based on different features. While machine learning algorithms seem to be showing promising results, there could be some pitfalls. It is essential that when applying these techniques the researchers should be made aware of all aspects of the study as this is an interdisciplinary research and some of these issues like methodology and implementation are discussed by author Wall [5]. Further details about the relationship between these diseases are essential for our knowledge and hence, will be discussed before understanding how machine learning is playing a crucial role.

2.1 Developmental Disorders and Comorbidity

While most developmental disorders deal with mental retardation, Autism Spectrum Disorder(ASD) is characterized by social interaction difficulties and communication challenges. There are many types of autism which are caused by different genetic combinations and environmental influences. The Centers for Disease Control and Prevention (CDC) estimates that it is present in 1 out of 68 children in the United States. It seems to occur 4.5 times more frequently in boys than girls. The four sub-types of Autism are Autistic Disorder, Aspergers Syndrome, Childhood Disintegrative Disorder and Pervasive Developmental Disorder according to the fourth edition of Diagnostic and Statistical Manual of Mental Disorders(DSM)

which is published by American Psychiatric Association(APA). However, in the fifth edition published in 2013, all of these four sub-types were dissolved into a single disorder called Autism Spectrum Disorder and the American Psychiatric Association(APA) felt that it this would be helpful for accurately diagnosing individuals with autism [6].

Attention-Deficit/hyperactivity disorder(ADHD) is also a brain disorder with patterns of inattention and/or hyperactivity that affects the development of a child. Children who suffer from ADHD have difficulty in staying focused and paying attention, and difficulties with controlling behavior. Similar to ASD, ADHD also seems to have more prevalence among male children [7]. This disorder also continues to affect the children in their adulthood, almost one-third of children diagnosed with ADHD retain it even in adulthood [8]. According to the fourth edition of Diagnostic and Statistical Manual of Mental Disorders, published in 2000 by APA, there are three sub-types of ADHD. The three sub-types are combined type of ADHD, predominantly inattentive type ADHD and predominantly hyperactive-impulsive type ADHD. Both ASD and ADHD are disorders for which the causes are not clear and also researchers are not sure about the role played by environmental factors.

Deletion syndrome 22q11.2(VCFS) is also known as DiGeorgesyndrome as it was first described by Angelo DiGeorge in 1968 [9, 10]. It is caused by the deletion of 30 to 40 genes in the middle of chromosome 22 at a location known as 22q11.2. This is a genetic disorder and 10% of the cases are inherited by a parent. DiGeorge syndrome occurs in about 1 in 4,000 people. Children diagnosed with this disease have delayed growth and speech development, and also have learning disabilities. Individuals affected by this disorder have breathing problems, low levels of calcium in blood and kidney abnormalities. Unlike ASD and ADHD, VCFS can be diagnosed by performing genetic analysis to detect microdeletions. Fluorescencein situhybridization(FISH)is a method which helps diagnose VCFS, quantitative polymerase chain reaction(qPCR) is quicker than FISH. It has a turn around time of 3 to 14 days. Those affected by VCFS have the risk of developing other disorders like schizophrenia, depression,

anxiety, and bipolar disorder [11].

While each of these disorders affects a person in a different way they are also related to one another or more specifically one leads to another. When these disorders co-occur then they are called comorbid disorders. As mentioned earlier ASD and ADHD which co-occur are comorbid disorders while VCFS and schizophrenia are comorbid disorders. It is essential to understand their comorbidity to better analyze and diagnose these disorders.

2.1.1 Co-occurrence of VCFS and Schizophrenia

Schizophrenia is also a mental disorder where the diagnosed fail to understand reality and lack of normal social behavior. People with schizophrenia have unclear thoughts, anxiety, depression and reduced social engagement [12] [13]. The evidence for VCFS being the first identifiable genetic subtype of schizophrenia was given by Anne. S. Bassett [14]. In this research, it was found that individuals whose relatives have schizophrenia tend to have a lower probability of containing individuals with 22qDS. Most studies that were done with different samples of data show that the relative risk for schizophrenia in an individual initially diagnosed with 22qDS is about 20 to 25 times the lifetime general population risk of 1% [15] [16].

When a study was conducted on non-overlapping samples, in which 82 individuals had VCFS and schizophrenia showed some comparable clinical signs [14]. By comparing the IQ of the patients, some reports showed that the IQ was same that is ranged from 70-84 irrespective of whether the patient had schizophrenia or not [17]. On the other hand, another study found mental retardation in 69% patients, found lower IQ levels for patients with VCFS and schizophrenia than patients with no schizophrenia[18]. Some other findings that distinguish these patients were that they had smaller brain volume, midline defects such as cavum septum pellucidum, and white matter hyper intensities on MRI [19] [20].

Due to these studies being carried out even though most of the times it continues to be unrecognized, clinicians are educated of the possibilities during diagnosing individuals. This would help them look for signs of schizophrenia in patients with VCFS. Also, they could let patients known about possibilities of schizophrenia and other disorders too.

2.1.2 Co-occurrence of VCFS and ASD

According to research, 15-20% patients diagnosed with VCFS meet the behavioral criteria for a diagnosis of ASD. Due to the missing DNA of the 22 chromosome, there could be modifying affects in each persons set of genes. Some individuals could have social difficulties, developmental delays or learning disabilities. These are some of the symptoms of an autistic child. Hence, it is believed that understanding genetic causes for autism is important.

Various studies have been conducted to examine the extent to which VCFS individuals also have ASD. One of these studies uses children, adolescents and young adults whose males to females ratio is 2.3:1. Based on the information from the Autism Diagnostic Interview-Revised (ADI-R), the Autism Diagnostic Observational Schedule (ADOS), and a clinicians best-estimate diagnosis, they estimated that ASD is present in 15-50% of the cases [21].

By studying the phenotypes, differences with the children with VCFS and VCFS+ASD is being researched. From this analysis it was seen that 94% of the children with VCFS + ASD had a co-occurring psychiatric disorder, on the other hand, 60% of children with VCFS had a psychiatric disorder. Also, further studies of the brain showed that children with VCFS + ASD had larger right amygdala volumes and all other neuro-anatomic regions of interest were statistically similar between the two groups [22].

Apart from schizophrenia, ASD is also one of those psychiatric disorders that children with VCFS could possibly have. In our study, some analysis has been done to study the phenotypes of children with VCFS and VCFS+ASD using machine learning techniques. Techniques have

mainly been applied to find distinguishable features and help broaden the understanding with the co-occurrence of these two developmental disorders.

2.1.3 Co-occurrence of ASD and ADHD

ASD and ADHD are two developmental disorders that affect an increasing number of children. These two developmental disorders share some common symptoms like lack of concern, not being able to react to others emotions or feelings etc. The reason why it is difficult to distinguish symptoms of ASD from ADHD is that they occur at the same time. In one of the studies in pediatrics field, showed that 18% of the children diagnosed with ADHD showed signs of ASD [23]. So, when this disorder occurs in children, they tend to have learning complications and impaired social skills.

Various studies have been done in the field of psychology to understand the relation between these two developmental disorders. However, researchers are still not certain about why these two developmental disorders occur together frequently. Initially, doctors never diagnosed a child with these two disorders, as they believed that they could not coexist as per the APA. It was in 2013, the release of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) when the APA stated that ASD and ADHD could co-occur [24]. Researchers found that 20-50% of the individuals diagnosed with ADHD had shown symptoms of ASD as well and 30-80% of the individuals diagnosed with ASD meet the criteria for ADHD [25]. It is also believed that when these two disorders co-occur, they cause greater morbidity and create a more complicated clinical challenge.

As the research in this domain is more recent, there is two possible hypothesis of these two disorders co-occurring. The first hypothesis is that ASD and ADHD are distinct yet overlapping diseases which may share some common etiology. The second hypothesis is that ASD and ADHD co-occurrence stands alone as a distinct clinical disorder, with a

distinct etiology [26]. The author Nat Gene tried to find a genetic relationship between five different psychiatric disorders which included ADHD and ASD using genome-wide SNPs, and according to his research, there seems to be a non-significant genetic correlation between the two disorders (ASD and ADHD) [27].

A few researchers from the Netherlands have suggested that ASD and ADHD are different manifestations of a single medical condition with different subtypes. As per their research, ADHD can occur independently without signs of ASD, but ASD always occurs with symptoms of ADHD[28]. When examining the brain images of persons with ASD, ADHD and both, the analysis showed shared and distinct brain alterations. For patients with both the disorders, there was an overlap in the corpus callosum and cerebellum (lower volume in structural MRI and decreased FA in DTI), and superior longitudinal fasciculus. They also found that the corpus callosum and cerebellum are usually smaller than usual size [29].On the other hand, few researchers examined the brain images as per the second hypothesis and concluded that brain maturationin both conditions proceeds differently or is delayed for individuals with ASD and ADHD. They also believe that distinct patterns of thinning in definite brain images will help them define subtypes of the ASD-ADHD syndrome [30].

As there are varied researches related to the co-occurrence of these two disorders, another approach believes on focusing on the traits(behavior) rather than genetic or brain features diagnosis [31]. Similarly, in our research, we try to examine the relations with the phenotypes of ASD, ADHD and both. So, different machine learning techniques have been used to draw conclusions about their co-occurrence.

2.2 Parent-oriented Reviews for developmental disorders

These are reviews in the form of questions which parents answer. For developmental disorders, parents are given a set of different questions to access the conditions of the children and

diagnose various disorders. Parents play a vital role when diagnosing children for developmental disorders, as they are the ones who seem to observe most of the signs and symptoms like when the child starts to speak when a child distracted etc. There are many parent-oriented review, but this section, three parent-oriented reviews will be explained in detail. The three parent-oriented reviews are Autism Diagnostic Interview (ADI), Behavioral Assessment Schedule for Children (BASC) and Vineland Adaptive Behavior Scales (VINE).

2.2.1 Autism Diagnostic Interview (ADI)

This is a semi-structured parent review to check for ASD related behaviors in a child. This test is performed to mainly analyze four domains in children. These four domains are Reciprocal Social Interaction, Language/Communication, Restricted, Repetitive, and Stereotyped Behaviors and Abnormality Present in early development (before age 3). This review usually takes two hours and the parents are asked 93 different questions which are related to the above four domains. The scoring for these questions is 0- Behavior of the type specified in the coding is not present, 1- Behavior of the type specified is present in an abnormal form, but not sufficiently severe or frequent to meet the criteria for a 2, 2- "Definite abnormal behavior and 3- "Extreme severity of the specified behavior. A child is diagnosed with ASD when the scores exceed the minimum cutoff scores [32]. These minimum cutoffs have been identified after many years of research.

Extensive research on the ADI shows that it is an extremely helpful mode of assessment and useful for treatment as well as education planning. The ADI was able to diagnose children with a chronological age of at least five years and a mental age of at least two years. Later on, in 1994, ADI was revised (ADI-R) to focus on the first three domains and the Abnormality Present in early development was removed as these features very less relevant compared to other domains[33]. The main advantage of ADI-R is that it means the DSM-IV criteria and is

focused on children in the 3-5 years range and a mental age of 18 months. It also has adequate sensitivity and specificity when administered by highly-trained personnel. The extensive use of ADI-R in the international research community seems to provide strong evidence of the reliability and validity of its categorical results in diagnosing ASD. It is proven to be affective in distinguishing ASD from other developmental disorders and identifying new subgroups[34].

Even though research shows that ADI-R is a good method to diagnose children with ASD at an early age(3 years), there also are researchers whose research show examples of children meet the criteria of ADI-R but dont have ASD. One such example is given where three children met research criteria for an ASD by meeting or exceeding the cut-off scores in the communication and social interaction domains, however, their social and communication behaviors were not similar to those of an autistic child. Also, when a clinical psychologist reviewed the children, they were not diagnosed with ASD [35]. So, it can be seen that there are chances of having false positives results with these tests.

2.2.2 Behavioral Assessment Schedule for Children(BASC)

Behavioral Assessment Schedule for Children is mainly designed to evaluate children with psychological problems. In 2004, BASC was revised to BASC-2 which helped to evaluate behavioral and personality aspects which included positive(Adaptive) along with negative(Clinical) dimensions [24]. To improve the flexibility of administering and enhancing progressive monitoring of children with emotional disabilities, BASC 2 was further revised to BASC 3 in 2015.

The main purpose of this review is assessing emotional/behavioral disorders in children and adolescents. The set of rating sales of BASC are comprehensive that help in evaluating child behaviors from a different perspective like a parent, teacher and so on. The different scales and forms of BASC are Teacher Rating Scales (TRS), Parent Rating Scales (PRS), Self-Report

of Personality (SRP), Student Observation System (SOS), and Structured Developmental History (SDH) [24]. Most of our research deals with the PRS and TRS which has the highest correlations with Hyperactivity, Aggression, Atypicality, Withdrawal, and Attention Problems [24].

The SRP, the TRS, and the PRS are scored with T-scores that depend on a national norm group, by gender in the norm group, or in comparison to clinical population. However, the SDH and the SOS do not have specific norms [24]. The children having T-scores above 40 are considered 'Average', T scores between 30-39 are considered 'Borderline' or 'At Risk' and T scores below 30 are considered 'clinically significant'.

In one of the studies, there were signs of elevated 'Anxiety', 'Atypicality' or 'Social Withdrawal' scores for children with VCFS indicating risk for schizophrenia[36]. Most of the time BASC has been applied to diagnose ASD and ADHD over VCFS. When diagnosing attention deficit disorders in children BASC, it has proven to perform more accurately then Child Behavior Checklist (CBCL). Also, it was able to explain salient behavioral dimensionsrelated to various ADHD subtypes [37]. BASC has also shown that children diagnosed with ADHD are rated lower on adaptive skills when compared to children with no diagnosis [38]. However, when trying to diagnose children with ASD, atypical behavior, attention and adaptive functions were complicated. It was also observed that the parent-rated social withdrawal was higher for children with ASD [39].

2.2.3 Vineland Adaptive Behavior Scales (VINE)

Vineland Adaptive Behavior Scales which one of the many assessment tools available for students with special needs. It was developed by three social research scientists Sara Sparrow, David Balla, and Domenic Cicchett [40]. It is used to measure adaptive behaviors, including

the ability to cope with environmental changes, to learn new everyday skills. The revised version of the Vineland Adaptive Behavior Scales was made in 2005, to better measure adaptive skills in very young children and to capture qualitative differences in communication and social interaction for individuals on the autism spectrum.

Researchers have found that when diagnosing children with ADHD who had average full-scale IQs, they had Vineland standard scores in the borderline to low-average range. It was also observed that ADHD children with tertiary attention problem had significant social adaptive dysfunction on the Vineland [41]. On the other hand, when analyzing the adaptive behavior of children with ASD using VINE, low scores were found in social skills while high scores were found in motor skills [42]. Also, when distinguishing the ASD+ADHD children from the autistic children, the prior had lower scores on the VINE and the Pediatric Quality of Life Inventory. It was found that autistic children had greater impairment in adaptive functioning and clinically significant in children who suffered from both ASD and ADHD. However, children who suffered from only ADHD had fewer symptoms [43]. Furthermore, when children with VCFS were studied it could be seen that boys diagnosed with VCFS scored lower than girls diagnosed with VCFS on communication, daily living skills and socialization measures of VINE. The study also found that there was a negative correlation between age and cognitive function with girls, that is the scores did not keep up with expected improvement with age, while it was not the case for boys [44].

All the above example show that these parent-oriented reviews are good indicators of the three developmental disorder ASD, ADHD, and VCFS. They are reliable and used by most psychologists or clinicians to diagnose children. For our research, the variables/features are these three parent-oriented reviews results for subjects which will be used for our analysis.

2.3 Machine Learning applied to assess developmental disorders

Machine learning is a field of computer science wherein the computers have the ability to learn from data rather than being explicitly programmed [45]. The term machine learning was given by an American pioneer Arthur Samuel in 1959 [46]. It was not until the 1990s that machine learning started to flourish as a separate field. While psychology and machine learning seem to be two independent studies, the field which is a combination of both these fields is Cognitive Science. Apart from these two fields, there are researchers from other fields like biology, neurosciences, sociology and so on who contribute towards this field. There are different research methods used in cognitive sciences which are derived from different fields. However, behavioral experiments and brain imaging are two methods related to developmental disorders and will be discussed in greater details.

2.3.1 Behavioral Experiments

These are experiments to measure behavioral responses to different stimuli and understand about how those stimuli proceed. The measures used are of three types- behavioral traces, behavioral observations, and behavioral choice [47]. Behavioral traces are pieces of evidence that indicate behavior occurred, but the actor causing the behavior is not present. On the other hand, behavioral observations involve the direct witnessing of the actor engaging in the behavior and behavioral choices are when a person selects between two or more options.

The reaction time of the person can indicate differences between cognitive process. This could be used to analyze different things in individuals and draw various conclusions. It was proven that reaction times are highly correlated to intelligence, which means that highly intelligent people tend to process speed faster. So, it could be seen that reaction times could be indicators of psychological distress [48]. Similarly, in this study, the authors try to

differentiate ASD patients(18) and ADHD patients(30) from control groups(13) [49]. The basis for their differentiation is that patients with developmental disorders like ASD and ADHD fail to understand certain emotions. Also, it was observed that children with ASD have shorter reaction times when compared to those that do not have ASD [50]. So, the data for this study was collected based on the childrens ability to understand emotions. The features are related to the response and response latency of the children. The children were grouped into seven different groups and the based on the responses features were created. They applied ReliefF feature selection algorithm and the machine learning algorithms applied are Decision Tree, Random Forest, Support Vector Machine, K-nearest neighbor and Ada Boost. Their main aim was to study the emotional status of the patients and based on this identify their diagnosis. The overall study shows that ASD children could be differentiated from ADHD and control group with 80% accuracy [49].

2.3.2 Brain Imaging

The analysis of activities in the brain while performing various tasks is called brain imaging. By linking the brain function and behavior, we can understand how information is processed. As discussed earlier, brain images of children with developmental disorders are studied, these help us during diagnosis.

The functional magnetic resonance imaging(fMRI) measures brain activity by detecting changes associated with blood flow [51]. As some researchers have used deep learning techniques to study the fMRIs of children suffering from autism and tried to find patterns that could differentiate them from control groups. There are two widely known datasets Autism Brain Imaging Data Exchange(ABIDE) and ADHD-200 which have fMRI brain images of children who are diagnosed with ASD and ADHD respectively. In 2016, researchers applied a learning technique called '(f)MRI HOG-feature-based patient classification(MHPC)'

on these datasets that uses the Histogram of oriented gradients (HOG) features to predict ADHD and ASD from their respective datasets. This algorithm was able to achieve a hold-out accuracy of 69.6% for distinguishing ADHD and hold-out accuracy of 65.0% for distinguishing ASD from the control groups [52].

Recently in 2017, researchers improved the state-of-art model by achieving 70% accuracy in identifying ASD patients from control groups. The data used for this research is also the brain imaging data from a world-wide multi-site database known as Autism Brain Imaging Data Exchange(ABIDE). A connectivity matrix was built using correlation which is calculated for the average of the time series of the regions of interest. Then different classification algorithms like Support Vector Machines, Random Forest, and Deep Neural Nets was applied for the purpose of prediction. The DNN achieved a mean accuracy of 70% with 74% sensitivity and 63% specificity, while the SVM and RF model achieved a mean accuracy of 65% and 63% respectively. Apart from prediction, the researchers applied machine learning techniques to identify different areas of the brain that are negatively correlated and positively correlated to ASD. Even though their research is not unto to the biomarker standards, they believe that further research can be very helpful for developmental disorders diagnosis [53].

Support vector machine was also applied to imaging data of patients with VCFS. When applying diffusion imaging methods, white matter micro-structural abnormalities identified have affected a variety of neuro-anatomical tracts in 22q11.2DS. So, applying SVM on these diffused images can help optimize the selection process to discriminate VCFS patients from others. The mean accuracy obtained on the validation set was 84.8% and also the researchers were able to identify important diffusion features in the imaging data [54].

So, these examples show that machine learning techniques have aided psychological methods. They have helped provide more insights and even though the results do not seem to result in clear real-world usage, they definitely show signs for more valuable research in this field.

2.3.3 Applying Machine Learning on Screening processes

Most of these parented-oriented reviews are time-consuming and there are lots of different reviews that exist which aid in diagnosing children with developmental disorders. For diagnosing children with ASD, there are reviews like ADI, ADI-R, ADOS and so on. Also similarly for diagnosing children with ADHD, there reviews like BASC, ADI, VINE, CBCL and so on. The method of diagnosing is based on the clinician or psychologist who is assessing the children. While each of these reviews have their own advantages and disadvantages, the common problem with them is that they consume a lot of time. So, researchers believe that applying machine learning techniques like feature selections can help speed up the diagnosis process and also aid in better understanding of these reviews.

Feature selection is also known as variable selection is the process of selecting a subset of relevant features for our model construction [55]. Researchers apply feature selection algorithm to Social Responsiveness Scale(SRS) score sheets of individuals who either had ASD or ADHD. There are 65 questions in the SRS and using forward feature selection algorithm called minimal-redundancy-maximal-relevance (mRMR) criterion, they selected top 6 features. This feature selection algorithm tends to select features with a high correlation with the class (output) and a low correlation between themselves. Then they applied four different machine learning algorithms from the scikit-learn package which are SVC, LDA, Categorical Lasso and Logistic Regression. The observed that the comparable accuracies to be 0.962 - 0.965 and plotted the ROC. This shows that these features are doing a great job of distinguishing ASD patients from ADHD patients [56].

Other researchers applied the ADTree machine learning algorithm on the Autism Diagnostic Observation Schedule-Generic(ADOS) to evaluate ASD. This behavioral evaluation consists of four different modules and each module takes around 30 and 60 minutes to deliver. When applying this technique to the module 1 of ADOS, it showed that 8 of the 29 questions are

relevant. After training the classifier on these eight items, it achieved 99.7% sensitivity and 94% specificity [5]. So, there model is a tree-based approximation that classifies subjects into the ASD and non-ASD groups.

From the above examples it can be seen that even though machine learning has been applied to diagnose developmental disorders recently, it has shown promising results. Hence, for our research, different machine learning techniques have been applied on combinations of data, to contribute to this research to some extent. Some of these machine learning techniques will help us build models like the supervised and unsupervised learning techniques, while other machine learning algorithms like feature selection will help is analysis of the data. Therefore, different kinds of analysis has resulted in conclusions that can contribute to understanding these developmental disorders from a different perspective.

Chapter 3

Data Exploration, Data Preprocessing and Feature Selection

The performance of machine learning algorithms depends on the type of data. When the data is analyzed and transformed appropriately then meaningful results can be obtained from our Machine learning algorithms. So, the main aim of this chapter is to explore, analyze and preprocess the data that will be used for our models. In the first section, the different datasets used for our analysis are explained and different aspects of the datasets has been discussed. There are there different datasets that have been used for our research, the main one is provided by three different psychological labs in SU. Then, in the second section, various preprocessing steps to transform the actual dataset have been discussed and in the final section, discussion about the results obtained from different feature selection algorithms for various parts of our research.

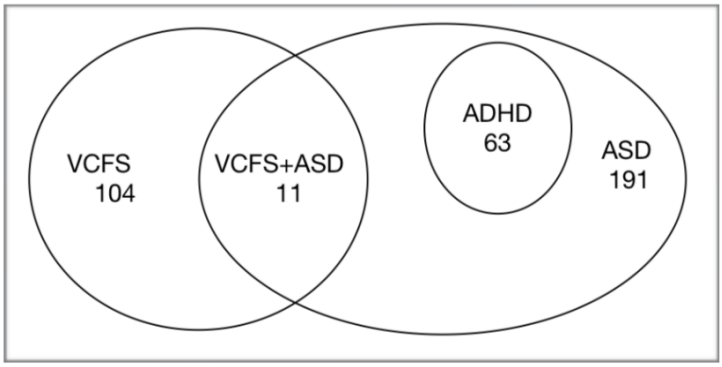

Figure 3.1: Venn Diagram of the division of diagnosis

3.1 Data Exploration

The data that has been used for our research has been provided by the labs 'Russo', 'Antshel' and 'Kates'. It has taken them over 10 years to collect this data and consists of information related to 369 children. Each of these children have different developmental disorders and the Venn diagram in figure 3.1 shows how the disorders are distributed. The age of children in this data is from 3 years to 18 years and the average age is 10 years. Also, the age at which the test has been taken by each subject is given. However, 70% of the subjects don't have the age at which they have taken the test. There are a total of 98 features in the given data. The data consists of features where the children have taken three different parent-oriented reviews.

However, not all children have taken all three parent-oriented review, figure 3.2 gives details of how many children among the 369 have taken each of the tests. It can be observed that there only 48% of subjects who have taken the VINE while majority have taken the ADI and the BASC.

The verbal IQ, non-verbal IQ and full scale IQ of the subjects is measured, and the range of IQ is from 40 to 160 with a mean IQ of 100. Out of the 369 children, there are 252 male children which constitutes 68% of the data and there are 117 female children which constitutes

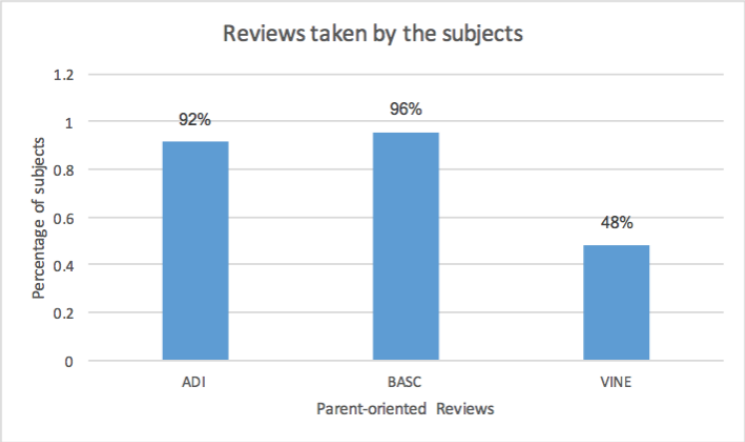

Figure 3.2: Reviews taken by subjects in percentage

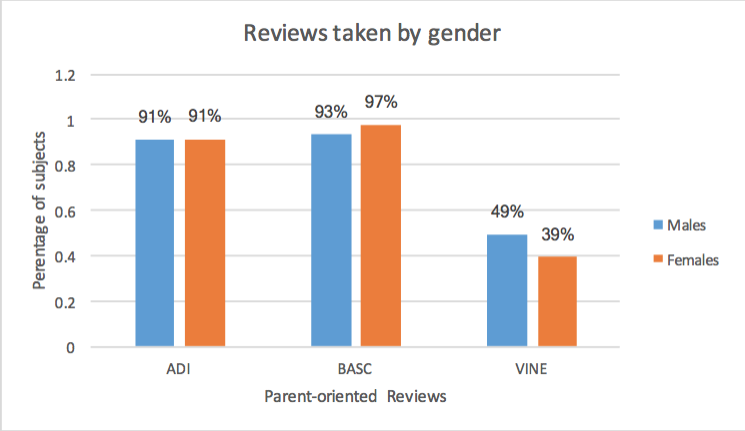

Figure 3.3: Parent-oriented reviews taken by gender

the rest 31% of the data. The figure 3.3 gives information on how many children have taken the reviews based on gender. If our models were to be trained on this data, then it could be observed that models would be biased towards males over females as they are more dominant the dataset. On the male subjects, the classifier had an accuracy of 96.03%, ROC Area value of 0.997 and F-measure of 0.959. On the other hand, the classifier gave an accuracy of 88.03%, ROC Area value of 0.972 and F-measure of 0.851. Therefore, most of our models predict the diagnosis labels in the subgroup data for male subjects with a better accuracy of 7% when compared to female subjects.

Also, further analysis of the actual data shows that for certain subjects, certain feature values

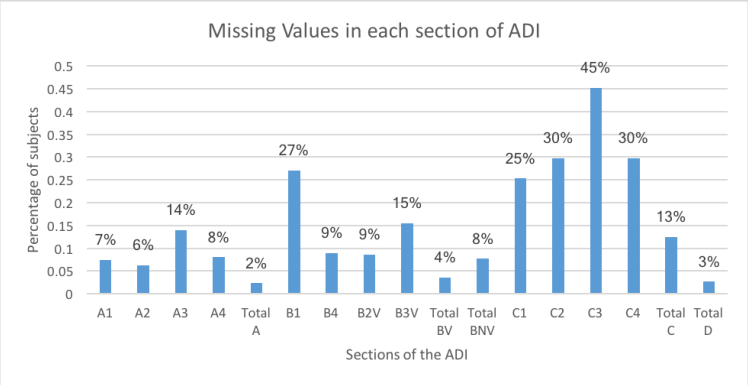

Figure 3.4: Analyzing missing values of features in ADI

are missing. When the person has taken a parent-oriented review, then certain questions in the review were missing. So, features of each parent-oriented having missing values have been analyzed. It could be observed that, from subjects who have taken the VINE parent-oriented review, none of the feature values were missing. The features of ADI parent-oriented review and BASC parent-oriented review which had missing values are shown in figures3.4 and 3.5 respectively. When considering the missing features for ADI, the missing sections have been considered as the parents have failed to fill a particular section of the review. Even though the missing values of individual sections is as high as 45%, it can be seen that the missing values of each of the four domains is very low(A-2%, BV-4%, BNV-8%, C-13%, D-3%). This shows that particular sub sections of the review are not filled, but overall sections in the review are filled. So, in the case of ADI, the missing features when handled accurately can avoid model over-fitting due to outliers. When we analyze the features of BASC, it was observed that only 4 out of 16 features had missing values. The feature 'Adaptability' had 20% of its features missing and the other features had very low percentage of missing values.

In our research, when analyzing about the comorbid disorders ASD and ADHD, the 'ABIDE' and 'ADHD-200' datasets have also been used along with our actual dataset. These datasets have phenotypes and f(MRI) images of the brain and for our research the phenotypes aspect of the dataset has been taken. When analyzing how IQ is effecting the comorbidity of ASD and ADHD, the subjects from these datasets who have been diagnosed with ASD and ADHD

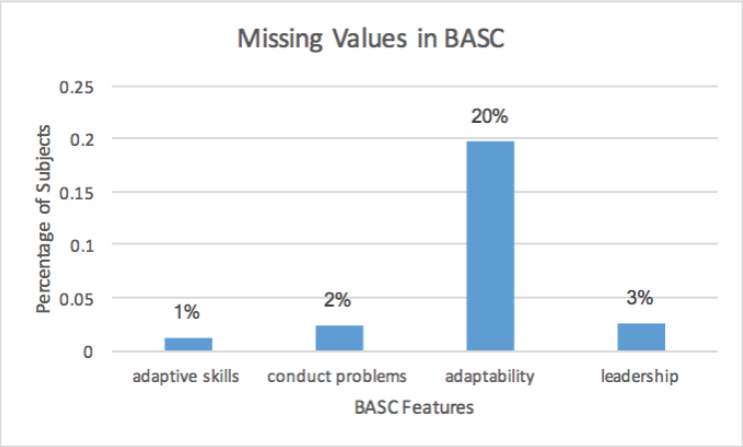

Figure 3.5: Analyzing missing values of features in BASC

were added to our dataset. The main reason behind this is that our dataset doesnt contain any children with only ADHD. To keep our data balanced subjects were selected randomly from 539 ASD individuals and 285 ADHD individuals.

3.2 Data Preprocessing

This is one of the most important steps of the data mining process. The presence of noisy, irrelevant and unreliable data makes it difficult for knowledge discovery during the training phase of machine learning. Some of the preprocessing steps applied to our data to make it suitable for our training phase are as follows:

- Data Transformation- This is the process of converting the given data into a desired format. The most common transformation technique applied to our dataset is the conversion of categorical features. Most of the data transformation steps that have been applied are given below and these steps are necessary for using the machine learning algorithms from sklearn package:

  1. The feature 'Sex' is converted to 1 (indicates male) and 0 (indicates female).
  2. All other feature columns whose attribute values are YES/NO will be converted

to 1 (for yes) and 0 (for no) .

3. The target column 'Diagnosis' is mapped to labels as follows:
   (a) ASD is 0
   (b) ADHD+ASD is 1
   (c) VCFS+ASD is 2
   (d) VCFS is 3

   Also, when individual developmental disorders are studied or when we study comorbidity, the target column will be changed according. For example when analyzing ASD, all the children diagnosed with ASD will be mapped to 1 and others will be mapped to 0. Similar mapping will be done during different parts of our research.

Some parts of our research use machine learning algorithms from WEKA, for this purpose, the transformation technique used is to convert the data types of ADI features from numeric to nominal. Therefore, by applying these transformation techniques the data types will be correctly interpreted by the machine learning algorithms that will be used for our analysis.

□ Data Reduction- This is the process of reducing the given data to include only the necessary columns. Our data consists of many unique identifiers and unnecessary columns. By doing this, over-fitting of the data during the training phrase can be avoided. Over-fitting of our model means that our model is too closely fit to the data and it makes the model more complex than necessary. The following are the steps taken to reduce our feature set:

1. The columns- 'Subject ID', 'Lab', 'BASC', 'VINE', 'ADI' and 'ADIAUTISM' are removed from the dataset and are not used by the model for prediction. The feature 'ADIAUTISM' is a direct indicator of whether a child has ASD or not, hence it needs to be removed from our data. Now, the total number of features/variables

on which the model will be trained is 92.

2. Further more, the columns which have totals for the ADI test have also been removed from the dataset, as these columns contain redundant feature information and are just simply sum of previous features.
3. Also, the age attribute has been condensed from four, that is three columns 'ADI Age', 'BASC Age' and 'VINE Age' remain. For the children who dont have these age, their actual age from the 'Age(Years)' is considered to be the age when the reviews were taken by the children.

After applying the data reduction techniques on our data, the final number of features on which the training will be done is 73 including the target column 'Diagnosis'. In WEKA, this step can be carried out using the Remove attribute feature. Hence, with the help of this step, our models will be able to generalize over our data.

□ Data Cleaning- It is a process of detecting, correcting or removing inaccurate instances from our data. The most important task in our data cleaning step is handling missing values. As seen in the data exploration section, most of our features have missing values. While handling missing values in our data, the structure of our data should not be effected, that is they should not become noise or outliers of our data. Hence, it is important that each features of our feature set are handled differently so that the structure of our actual data is maintained. The following are the steps as to how the missing values are handled:

1. Handling missing values for IQ test features- The PRI, VCI and FSIQ columns missing values are replaced with the mean value which is 100. This has not affected the mean value of the dataset, so it is a good idea to replace missing values with 100.
2. Handling missing values for ADI parent-oriented review features- The scores for ADI features/variables ranges from 0 to 2. In this case, the missing values are

replaced with a 0 because this indicates not present. When the value is missing it is safe to assume that that the symptom is not present in the children than making any other assumptions and leading to false outcomes.

3. Handling missing values for BASC parent-oriented review features- The scores have a mean of 50 and the standard deviation of 10. When the values are 70 it means it is critical. So, for this test the values are going to be replaced with the mean that is 50. For any value below 50, the score is not to be considered, hence, the missing values in the dataset are replaced with 50.
4. Handling missing values for VINE parent-oriented review features- The scores are standard and the missing values are going to be replaced with 0. This is because most subjects don't have the VINE gets scores and the subjects that have taken the VINE tests have the results without any missing values.

After the preprocessing of the data, the model is trained on the modified features. The next sections discuss the training steps and the results obtained after training the model. Also, before training the model out of the 73 features, the features best used to train the model are selected so that the model can be trained correctly with these important features.

3.3 Feature Selection

Feature selection also known as variable selection is the process of selecting a subset of features from our given data. The different feature selection algorithms used in our research are LASSO regression, ReliefF and Recursive Feature Elimination . The LASSO regression method that selects subset of variable when there is highly correlated predictors in the data, as in our dataset. On the other hand, ReliefF is a noise tolerant and robust algorithm which is independent of variable dependencies. While RFE uses an elimination process to select a subset of features recursively from our data. One or all of these feature selection algorithm

have been applied on different aspects of our research and the results obtained are discussed in the following sections.

3.3.1 Subgroup Diagnosis

As part of our research, initially, predictions are done on the actual data sample. Feature selection algorithms are used on the preprocessed data which has 73 features and the best 5 features are selected. The root mean squared error is used to analyze the performance of our feature selection models. Also, the best features are selected by doing 10-fold cross validation, that is at each fold the best features are taken and then finally best 5 features are found using a weighted average. ReliefF algorithm was not applied to our data to select best features as there are multi class labels in our data and ReliefF works well for binary classifiers. The RMSE for LASSO and RFE is 0.6837 and 0.66575. Both of these models have similar RMSE and hence, it can be said that these models are have similar performance measures when selecting the best features. However, LASSO has selected features from BASC and VINE, whereas RFE has selected features from ADI and are shown in table 3.1. So, there are no convergent features selected by these two algorithms.

LASSO	**Recursive Feature Elimination**
1. Performance IQ 2. Activities of Daily Living 3. Vineland daily living 4. Adaptability 5. Vineland Composite	1. Criteria for repetitive behaviors and stereotyped patterns 2. Criteria for qualitative impairments in reciprocal social interaction 3. Quality of social overtures 4. Offers comfort 5. Inappropriate questions or statements

Table 3.1: Best 5 variables by different feature selection algorithm for Subgroup Diagnosis

3.3.2 Comorbid Developmental Disorders

There are two groups of comorbid disorders that are analyzed in our research. The first one is ASD and ADHD comorbidity and the other group is ASD and VCFS comorbidity. The actual data is divided to separate ASD and ASD+ADHD to analyze the comorbid disorders ASD and ADHD. This data consists of 254 subjects who have been diagnosed with ASD, out of which 63 subjects have ADHD as well. This data consists of 190 males and 64 females where 53 males and 10 females have ADHD along with ASD. For analyzing ASD and VCFS comorbidity, the data used consists of 115 subjects diagnosed with VCFS and 11 have ASD as well. Even though there are 49 females and 55 males, this data is more balanced than other data that we have used so far. The best features for both these comorbidities is analyzed in this section.

ReliefF	**LASSO**	**Recursive Feature Elimination**
1. Vineland socialization 2. Attention problems 3. Performance IQ 4. Vineland Communication 5. Vineland Composite	1. Vineland socialization 2. Quality of social overtures 3. Offering to share 4. Inappropriate questions or statements 5. Attention problems	1. Quality of social overtures 2. Inappropriate questions or statements 3. Abnormality of development evident at or before 36 months 4. Criteria for Qualitative impairments in reciprocal social interaction 5. Conventional/Instrumental Gestures

Table 3.2: Best 5 variables by different feature selection algorithm for ASD and ADHD comorbidity

For ASD and ADHD comorbid disorders, the best 5 features from the three feature selection algorithms are given in table 3.2. The RMSE for LASSO, ReliefF and RFE is 0.289, 0.493

and 0.224 respectively. So, it shows that RFE and LASSO are performing better than ReliefF feature selection algorithm as they have lower RSME values. The LASSO feature selection algorithm has selected features from all three parent-oriented reviews and the RFE feature selection algorithm as selected all variables from ADI review. While the ReliefF has selected variables from BASC and VINE. It can be seen that there are convergent variables in the LASSO feature selection algorithm.

The best features for ASD and VCFS comorbidity selected when applying the three feature selection algorithms is given in table 3.3. The RMSE for LASSO, ReleifF and RFE is 0.0093, 0.9484 and 0.07 respectively. By the observed RMSE values, it can be observed that LASSO and RFE high performance and ReliefF has poor performance. another observation made after looking at the features is that there no convergent features from different parent-oriented reviews. In fact, the LASSO selects the features from BASC, ReliefF from VINE and RFE from ADI. The only convergent feature is the 'Performance IQ' selected by both LASSO and ReliefF.

LASSO	**ReliefF**	**Recursive Feature Elimination**
1. Anxiety 2. Hyperactivity 3. Performance IQ 4. Conduct Problems 5. Depression	1. Verbal IQ 2. Vineland Composite 3. Performance IQ 4. Full Scale IQ 5. Vineland Daily Living	1. Criteria for Repetitive behaviors and stereotyped patterns 2. Repetitive use of objects or interest in parts of objects 3. Verbal Rituals 4. Criteria for Communication 5. Hand and Finger mannerisms

Table 3.3: Best 5 variables by different feature selection algorithm for ASD and VCFS comorbidity

3.3.3 Individual Developmental Disorders Diagnosis

Our data consists of information related to three different developmental disorders that is ASD, VCFS and ADHD. For each of these disorders, the features are analyzed and the best features are found. The target label 'Diagnosis' is converted to 'ASD Diagnosis', 'ADHD Diagnosis' and 'VCFS diagnosis' for each of the disorders. Then, the 73 features preprocessed data is given as input to our feature selection algorithms and the best features for each developmental disorder are observed.

LASSO	**ReliefF**	**Recursive Feature Elimination**
1. Behavioral Symptoms 2. Performance IQ 3. Full Scale IQ 4. Adaptive Skills 5. Externalizing problems	1. Activities of Daily Living 2. Vineland Communication 3. Vineland Composite 4. Vineland Daily Living 5. Functional Communication	1. Criteria for Repetitive behaviors and stereotyped patterns 2. Use of other's body to communicate 3. Criteria for Qualitative impairments in reciprocal social interaction 4. Hand and Finger mannerisms 5. Repetitive use of objects or interest in parts of objects

Table 3.4: Best 5 variables by different feature selection algorithm for ASD

The best features for distinguishing children with ASD from those who don't have it are given in table 3.4. The RMSE values for LASSO, RFE and ReliefF are 0.181, 0.218 and 0.523. This shows that LASSO and RFE are performing better than ReliefF due to the low RSME scores. LASSO selects the best features from the BASC parent-oriented review, ReliefF selects from both the BASC and the VINE parent-oriented review and RFE selects from ADI. Also, it can be observed that there is no overlap of features by all three feature selection algorithms.

LASSO	**ReliefF**	**Recursive Feature Elimination**
1. Performance IQ 2. Full Scale IQ 3. Behavioral Symptoms 4. Adaptive Skills 5. Direct Gaze	1. Performance IQ 2. Vineland socialization 3. Vineland communication 4. Vineland daily living 5. Vineland composite	1. Inappropriate questions or statements 2. Quality of social overtures 3. Conventional/Instrumental Gestures 4. Criteria for Qualitative impairments in reciprocal social interaction 5. Offers comfort

Table 3.5: Best 5 variables by different feature selection algorithm for ADHD

The children with ADHD and those who don't have can be distinguished using the best features given in table 3.5. The RMSE values for LASSO, RFE and ReliefF are 0.190, 0.201 and 0.407. The LASSO selects features from BASC, while the ReleifF selects from VINE and RFE selects from ADI. Even in this case there are no convergent parent-oriented reviews features except the Performance IQ feature. As the RMSE values are low for LASSO and RFE, the models are performing better than ReliefF.

For the developmental disorder VCFS, the RMSE values for LASSO, RFE and ReliefF are 0.189, 0.308 and 0.555. The best features selected by each of the three feature selection algorithms are given in table 3.6. LASSO select features from all of the three parent-oriented reviews, while ReliefF selects from BASC and VINE. On the other hand, RFE selects from ADI. In this case, the convergent features are from ADI and VINE which are selected by LASSO.

When all of the feature selection algorithms are compared, it can be seen that RFE usually tends to select algorithms from the ADI parent-oriented review, while ReleifF selects mostly from VINE parent-oriented review and sometimes from BASC. LASSO is the only feature selection algorithm that has selected features from all three parent-oriented reviews. LASSO

LASSO	**ReliefF**	**Recursive Feature Elimination**
1. Hyperactivity 2. Criteria for Qualitative impairments in reciprocal social interaction 3. Pronominal reversal 4. Vineland Communication 5. Vineland Socialization	1. Activities of Daily Living 2. Functional Communication 3. Vineland Communication 4. Vineland daily living 5. Vineland composite	1. Criteria for Repetitive behaviors and stereotyped patterns 2. Use of other's body to communicate 3. Pronominal reversal 4. Neologisms/idiosyncratic language 5. Stereotyped Utterances & delayed echolalia

Table 3.6: Best 5 variables by different feature selection algorithm for VCFS

and RFE are both performing better than ReliefF. However, LASSO feature selection algorithm can be considered to be doing a better job as it has selected features from all three reviews. Also, LASSO assigns coefficients to each of the features and these coefficients can be negative or positive. It eliminates all the unimportant features by assigning them weights 0. Therefore, for most parts of our research only LASSO features have been used for our analysis.

Chapter 4

Investigating the Subgroup Diagnosis

The dataset collected in different labs consists of different groups of diagnosis. Each of these diagnosis is dependent on the features of the reviews. In this chapter, analysis of these groups has been done using different Machine Learning techniques. For the purpose of understanding the diagnosis, initially unsupervised learning technique (k-means clustering) is applied. Then, feature selection algorithms are applied to understand the feature importance of our entire data available. Finally, supervised learning techniques have been applied to predict each of these different groups diagnosis. Also, the features of the data are divided into four sections (IQ, ADI, BASC, VINE) and supervised learning techniques have been applied to predict the diagnostic label using each section of features.

4.1 Unsupervised Learning Techniques to Categorize Subjects

The dataset is clustered using K-Means Clustering algorithm to analyze the clusters in the dataset. The dataset consists of 73 features/variables after the preprocessing of our data. For applying K-Means clustering, the dimension of the data has to be reduced further and for this purpose Principal Component Analysis has been used. The PCA transforms the

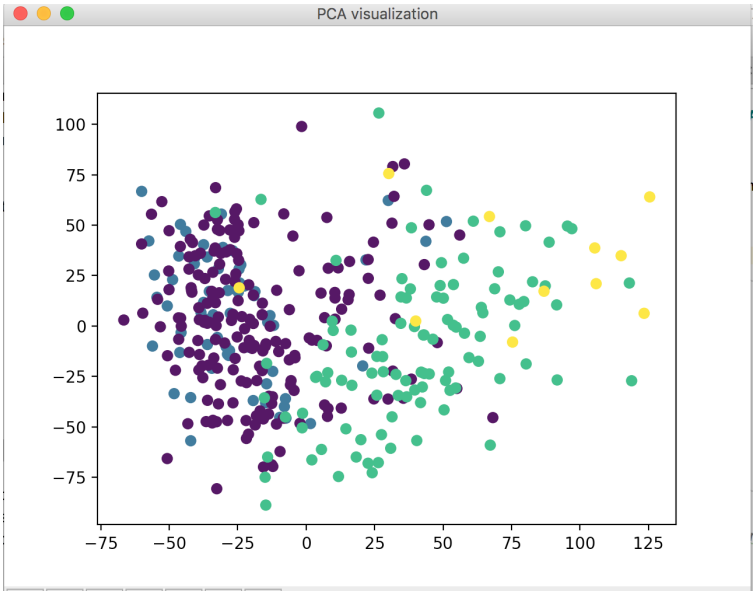

Figure 4.1: Analyzing missing values of features in BASC

set of features of the dataset orthogonally. Also, PCA handles the missing values and the outliers in our data, so for this section of our analysis missing values have not be handled as part of preprocessing. For the given dataset, the PCA algorithm in sklearn package in python is used. For visualizing our dataset, the PCA algorithm is applied to find the first two principle components of our dataset, so that our data can be visualized in 2D. After performing PCA, the figure 4.1 gives the visualization of our dataset. From this figure, it can be seen that there are four different/unique groups. However, there is an overlap between 2 groups as children diagnosed with 'ASD+ADHD' and falling into the group of children diagnosed with 'ASD'. Similarly, children diagnosed with 'VCFS+ASD' fall into the group of children diagnosed with 'VCFS'. Due to this overlap there are only two distinguishable majority groups in our subgroup data. Now, on this reduced data from PCA has been given as input to K-Means clustering algorithm without the diagnostic labels. The K-Means clustering algorithm available in sklearn package has been used. As proper analysis of the data needs to be done, the k value was varied from 2 to 5. Since, there are 4 different diagnostic groups in our dataset, the data has been cluster unto k value equal to 5. The main reason behind this is to check if there are any pure clusters. For each of these k values, the visualization is given

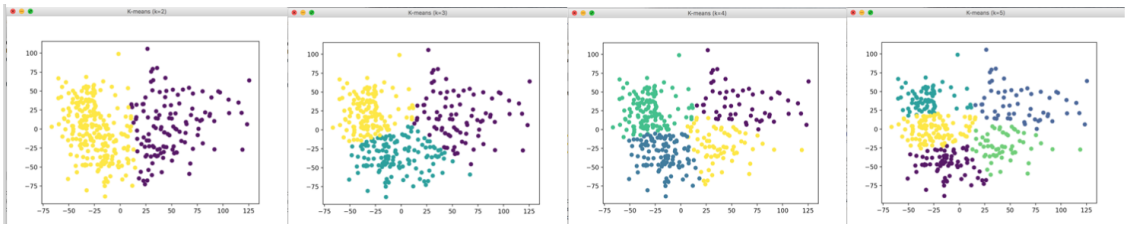

Figure 4.2: After applying k-means clustering on the PCA reduced dataset

in the figure 4.2.

In the figure 4.2, when k is equal to 2, cluster 1 (yellow color) consists of all VCFS diagnosed children (including children with VCFS+ASD) and cluster 2(purple color) has ASD and ASD+ADHD diagnostic children. However, there are 16 children in cluster 1 who have ASD only and 1 child in cluster 1 having ASD+ADHD. Then, the cluster size was increased to 3 and it was observed that cluster 1 consisted of all the VCFS+ASD children. Further analysis showed that children with ASD and ASD+ADHD were spread over all the three clusters, whereas children with VCFS only were spread over first two clusters. When the cluster size was increased further to 4 and 5, it was observed that the clusters where more of a mixture, there was no purity in the clusters. The main reason for this could be that there are lot of features overlap for each of the diagnostics.

As mentioned above, PCA handles missing values, if the preprocessed data were given as input to the K-Means then the accuracy of the K-Means clustering algorithm dropped to 83.5%, that means that there was more randomness in the data and higher chances of not grouping them accurately. However, the average accuracy of clustering the data was 95.3%. After, K-means clustering it could be said that children diagnosed with VCFS are closely clustered that is they belonged to one cluster most of the times and more specifically children with VCFS+ASD were most likely to be in the same cluster than the other diagnostic subgroups.

4.2 Supervised Learning Techniques to Predict Diagnostic Groups

The subgroup dataset after preprocessing is trained using different supervised learning techniques. As the dataset is small and simple, there is no need for very complex classification models. The data is loaded into WEKA and different classification algorithms have been applied on this data. The algorithms are compared based on different metrics as seen in table 4.1. The different metrics on which the models are evaluated are Accuracy, Precision, Recall, ROC Area and F-measure. Cross validation technique has been applied on the data and the dataset has been divided into 10 folds wherein training is done on 9 folds and testing on one. The different classification algorithms that have been used to train our models are Naive Bayes, Logistic Regression, Multi-class classifier and Random Forest. Since, our dataset consists of feature relations and multiple class labels, the above algorithms were chosen.

ML Algorithm	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Naive Bayes	0.934	0.924	0.928	0.981	92.4119%
Logistic Regression	0.926	0.924	0.924	0.990	92.4119%
Multi Class Classifier	0.902	0.902	0.902	0.980	90.2439%
Random Forest	0.960	0.959	0.957	0.994	94.85%

Table 4.1: Supervised learning techniques to predict subgroup diagnosis

Among the different supervised learning techniques applied on the dataset, Random Forest has a better accuracy and ROC Area value. So, it can be said that Random Forest is doing a good job of classifying our dataset. Also, each of the class labels are measured to check their precision and recall. For the group of children diagnosed with ASD+ADHD, precision and recall of Random Forest is 96.6% and 90.5% respectively. Similarly the precision and recall for the children diagnosed with ASD is 93.9% and 96.3% respectively. For the children diagnosed with VCFS, precision and recall are 95.4% and 100% respectively. However, precision and recall for the children with VCFS+ASD is 100% and 45% respectively and this could be because their are a low percentage of children with VCFS+ASD when compared to the other

diagnostic groups. Now, rather than giving the entire feature set, only the best features selected from the three different feature selection algorithms as in chapter 3 are given to machine learning algorithms to predict our subgroup diagnosis. The results of each of the feature selection algorithms are given below in table 4.2 by LASSO and table 4.3 by RFE.

ML Algorithm	Accuracy	Precision	Recall	ROC Area
Naive Bayes	66%	0.690	0.656	0.851
Logistic Regression	73%	0.605	0.732	0.827
Multi Class Classifier	72.6%	0.584	0.7267	0.829%
Random Forest	70.4%	0.680	0.705	0.859%

Table 4.2: Supervised learning techniques to predict subgroup diagnosis by LASSO

ML Algorithm	Accuracy	Precision	Recall	ROC Area
Naive Bayes	83.1%	0.836	0.832	0.924
Logistic Regression	84.2%	0.842	0.843	0.917
Multi Class Classifier	84.2%	0.841	0.843	0.916
Random Forest	84.5%	0.840	0.846	0.912

Table 4.3: Supervised learning techniques to predict subgroup diagnosis by RFE

From the tables 4.2 and 4.3, the features selected from the RFE feature selection are preforming better than out LASSO feature selection algorithm. However, when using only the best 5 features, the accuracy of the random forest drops from 96% to 84% but the ROC Area value is still high in the case of Random forest. Also, with the help of LASSO features we selected all the relevant features with non-zero weights. Out of the 73 features after preprocessing, the relevant features were 45 and when Random forest model was trained on these 45 features, an accuracy of 92% was achieved.

Since, Random Forest is doing a good job at predicting the subgroup features, another types of tree algorithm J48 also known as ID3 from WEKA is used to represent our model in the form of the tree. The figure 4.3 shows the results obtained from the J48 algorithm.It had an accuracy, precision and recall of 87%. Also, precision and recall for the subgroups

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

Criteria for Repetitive behaviors and stereotyped patterns = Y
|   Vineland Composite <= 59
|   |   Criteria for Communication = Y
|   |   |   BASC AGE <= 13.67: VCFS+ASD (9.0)
|   |   |   BASC AGE > 13.67: ASD (4.0/1.0)
|   |   Criteria for Communication = N: VCFS (4.0)
|   |   Criteria for Communication = 0: VCFS+ASD (0.0)
|   Vineland Composite > 59
|   |   quality of social overtures = 0
|   |   |   Performance IQ <= 90: VCFS (2.0/1.0)
|   |   |   Performance IQ > 90: ASD (7.0)
|   |   quality of social overtures = 1: ASD (140.0/6.0)
|   |   quality of social overtures = 2
|   |   |   inappropriate questions or statements = 0: ASD (7.0)
|   |   |   inappropriate questions or statements = 1
|   |   |   |   Behavioral symptoms <= 78: ASD (15.0/1.0)
|   |   |   |   Behavioral symptoms > 78: ADHD+ASD (2.0/1.0)
|   |   |   inappropriate questions or statements = 2
|   |   |   |   Offering to share = 0: ADHD+ASD (0.0)
|   |   |   |   Offering to share = 1: ASD (3.0)
|   |   |   |   Offering to share = 2
|   |   |   |   |   ADIGPw_PF = 0: ASD (1.0)
|   |   |   |   |   ADIGPw_PF = 1: ASD (6.0/1.0)
|   |   |   |   |   ADIGPw_PF = 2
|   |   |   |   |   |   Response to Approaches of other children = 0: ADHD+ASD (0.0)
|   |   |   |   |   |   Response to Approaches of other children = 1: ASD (3.05/1.05)
|   |   |   |   |   |   Response to Approaches of other children = 2: ADHD+ASD (57.95/3.0)
Criteria for Repetitive behaviors and stereotyped patterns = N
|   Performance IQ <= 102
|   |   Verbal Rituals = 0: VCFS (70.0/1.0)
|   |   Verbal Rituals = 1: ASD (3.0/1.0)
|   |   Verbal Rituals = 2: VCFS (0.0)
|   Performance IQ > 102: ASD (5.0)
Criteria for Repetitive behaviors and stereotyped patterns = 0
|   ADI AGE <= 11.333: ASD (2.0)
|   ADI AGE > 11.333
|   |   ADI AGE <= 11.622242: VCFS (26.0)
|   |   ADI AGE > 11.622242: ASD (2.0)

Number of Leaves  :     24

Size of the tree :      39
```

Figure 4.3: J48 pruned tree

ASD, ASD+ADHD, VCFS ranges from 81% to 90%, however, the precision and recall for VCFS+ASD group is 60%.

4.2.1 Predicting Subgroup Diagnosis using the IQ feature set

In this section of our research, IQ feature set consists of 3 features and using only these features, the subgroup diagnosis of the data is predicted. So, only these 3 features are given to our machine learning algorithms and the results are observed in table 4.4.

ML Algorithm	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.604	0.623	0.612	0.802	66.32%
Naive Bayes	0.401	0.485	0.438	0.612	48.50%
Logistic Regression	0.546	0.580	0.517	0.716	57.99%

Table 4.4: Supervised learning techniques to predict subgroup diagnosis by IQ feature set

When all the different machine learning algorithms are compared, then Random forest is performing better than other machine learning algorithms. However, the Random forest has

very low accuracy of 66%. It has dropped from 92% to 66% which is very low and not an accurate model for the purpose of prediction. In this case, it could be mainly because the IQ range is similar for all the diagnostic groups.

4.2.2 Predicting Diagnosis based on ADI feature set

The ADI review has 45 features/variables in the dataset that analyze the behavior of children. These attributes are used to predict the diagnosis of the child. Initially, many supervised techniques have been applied on the Subgroup data and the various metrics of evaluation are listed in the table 4.5.

Classifier	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.966	0.965	0.964	0.976	96.47%
Naive Bayes	0.876	0.870	0.869	0.959	86.99%
Logistic Regression	0.868	0.864	0.866	0.936	86.44%

Table 4.5: Supervised learning techniques to predict subgroup diagnosis by ADI feature set

When using the reduced feature set from LASSO, which consisted of 45 features, these features were a combination of all the three parent-oriented reviews and the accuracy of our Random Forest model was 92%. However, when we use only the 45 ADI features , the accuracy of our Random Forest model is 96% with high precision and recall. Since, random forest model is doing great when predicting the subgroup diagnosis with ADI feature set, another similar algorithm J48 has been applied to our data. this will help our model to be concisely visualized in the form of a tree. The results of J48 in the form of a tree are given in figure 4.4. The accuracy of our model is 87%.

```
J48 pruned tree
------------------

Criteria for Repetitive behaviors and stereotyped patterns = Y
|   quality of social overtures = 0
|   |   Criteria for Communication = Y: ASD (7.0/1.0)
|   |   Criteria for Communication = N: VCFS (5.0/1.0)
|   |   Criteria for Communication = 0: ASD (0.0)
|   quality of social overtures = 1
|   |   Criteria for Qualitative impairments in reciprocal social interaction = Y: ASD (143.0/7.0)
|   |   Criteria for Qualitative impairments in reciprocal social interaction = 0: ASD (0.0)
|   |   Criteria for Qualitative impairments in reciprocal social interaction = N: VCFS (2.0)
|   quality of social overtures = 2
|   |   inappropriate questions or statements = 0
|   |   |   Reciprocal Conversation = 0: ASD (1.0)
|   |   |   Reciprocal Conversation = 1: VCFS+ASD (4.0)
|   |   |   Reciprocal Conversation = 2: ASD (6.0)
|   |   inappropriate questions or statements = 1: ASD (17.0/3.0)
|   |   inappropriate questions or statements = 2
|   |   |   Offering to share = 0: VCFS+ASD (2.0)
|   |   |   Offering to share = 1: ASD (3.0)
|   |   |   Offering to share = 2
|   |   |   |   Group play with Peers OR Friendships = 0: ASD (1.0)
|   |   |   |   Group play with Peers OR Friendships = 1: ASD (6.0/1.0)
|   |   |   |   Group play with Peers OR Friendships = 2
|   |   |   |   |   Offers comfort = 0: ADHD+ASD (0.0)
|   |   |   |   |   Offers comfort = 1
|   |   |   |   |   |   Range of Facial Expressions = 0: ADHD+ASD (0.0)
|   |   |   |   |   |   Range of Facial Expressions = 1: ADHD+ASD (2.0)
|   |   |   |   |   |   Range of Facial Expressions = 2: VCFS+ASD (2.0)
|   |   |   |   |   |   Range of Facial Expressions = 3: ADHD+ASD (0.0)
|   |   |   |   |   Offers comfort = 2
|   |   |   |   |   |   Response to Approaches of other children = 0: ADHD+ASD (0.0)
|   |   |   |   |   |   Response to Approaches of other children = 1: ASD (4.07/1.07)
|   |   |   |   |   |   Response to Approaches of other children = 2: ADHD+ASD (55.93/3.0)
Criteria for Repetitive behaviors and stereotyped patterns = N
|   Hand and Finger mannerisms = 0
|   |   Criteria for Communication = Y
|   |   |   Inappropriate facial expressions = 0: VCFS (10.0/1.0)
|   |   |   Inappropriate facial expressions = 1: ASD (4.0/1.0)
|   |   |   Inappropriate facial expressions = 2: VCFS (0.0)
|   |   Criteria for Communication = N: VCFS (59.0)
|   |   Criteria for Communication = 0: VCFS (0.0)
|   Hand and Finger mannerisms = 1: ASD (5.0/1.0)
|   Hand and Finger mannerisms = 2: VCFS (0.0)
Criteria for Repetitive behaviors and stereotyped patterns = 0: VCFS (30.0/4.0)

Number of Leaves  :     30

Size of the tree :      44
```

Figure 4.4: J48 pruned tree with ADI feature set

4.2.3 Predicting Diagnosis based on BASC feature set

BASC is another parent-oriented review where there are 19 different features to analyze the behaviors of the children. These features of BASC are used to predict the diagnosis of the child. Initially, many supervised techniques have been applied on the Subgroup data diagnosis and the various metrics of evaluation are listed in the table 4.6.

ML Algorithm	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.713	0.751	0.712	0.879	75.06%
Naive Bayes	0.694	0.659	0.671	0.841	65.85%
Logistic Regression	0.693	0.713	0.694	0.861	71.27%

Table 4.6: Supervised learning techniques for subgroup diagnosis by BASC feature set

Random Forest is performing better than all other machine learning algorithms with an accuracy of 75.06%, but this accuracy is not high for predicting subgroup diagnosis as

compared to previous models of predictions.

4.2.4 Predicting Diagnosis based on VINE feature set

VINE is also another parent questionnaire, when compared to other two reviews, VINE has fewer features for analyzing the behaviors. Also, the children who have taken the VINE tests are less compared to other two tests. These 4 features of VINE are used to predict the diagnosis of the child. Initially, many supervised techniques have been applied on the Subgroup data and the various metrics of evaluation are listed in the table 4.7. In the case of VINE parent-oriented review, the best model is by Logistic Regression with an accuracy of 71% and the random forest only gives an accuracy of 68%.

ML Algorithm	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.560	0.680	0.612	0.785	68.02%
Naive Bayes	0.550	0.553	0.517	0.600	55.28%
Logistic Regression	0.656	0.710	0.680	0.787	71%

Table 4.7: Supervised learning techniques for subgroup diagnosis by VINE feature set

4.3 Observations

The machine learning algorithms applied to distinguish the subgroups are doing good job and the results are great. Among all the machine learning algorithms that have been applied on our actual data, Random Forest is doing well at predicting our subgroup diagnosis labels. This shows that the model is successfully at finding the required relations between the feature set to the diagnostic labels. Using LASSO feature selection algorithm, the features were reduced from 72 to 45 and the performance of the Random Forest model is still very good. This shows that not all the 72 features are required to predict the diagnosis of the children

with different combinations of these developmental disorders.

Apart from this when the feature set was divide into four different sub features set, it was observed that using the IQ features, the BASC parent-oriented review features and VINE parent-oriented review features can not be used individually to define models that can predict the subgroup diagnosis accurately. On the other hand, ADI parent-oriented reviews have resulted in good models when predicting the subgroup diagnosis features. The Random Forest model designed with ADI features is better than the Random Forest model built with 45 features from LASSO. Also, the J48 pruned tree with the entire feature set has similar model performance as the J48 pruned tree with only ADI parent-oriented reviews, that is both models have an accuracy of 87%. When the ADI features are analyzed in more depth, it could be seen that some of the important features as selected by the J48 algorithm among the 45 ADI features are as follows:

1. Criteria for Repetitive behaviors and stereotyped patterns
2. Criteria for Qualitative impairments in reciprocal social interaction
3. Inappropriate questions and statements
4. Offers Comfort
5. Quality of social overtures
6. Offering Comfort
7. Range of Facial Expressions
8. Hand and finger mannerisms
9. Response to approaches of other children
10. Reciprocal Conversation

Most of the features that have been selected are key symptoms of the developmental diagnosis that our research is trying to predict in the children. This shows that our models are able to find valuable information like this from our features. As models built with ADI features are doing well and comparable with the entire feature set, it can be said that ADI

features are sufficient. Therefore, ADI parent-oriented review is better than BASC and VINE parent-oriented review for building models to predict the subgroup diagnosis.

Chapter 5

Analyzing Comorbidity of Developmental Disorders

The actual data is divided into two types, the first deals with children who are diagnosed with ASD and both ASD and ADHD and the second deals with children who are diagnosed with VCFS, ASD and both these disorders. In the first half go this chapter using the first data, the comorbid disorders ASD and ADHD are analyzed and then in the second half of this chapter comorbid disorders VCFS and ASD are analyzed. Different supervised and unsupervised machine learning techniques will be applied in both the sections to build models and results obtained will be analyzed. Also, analysis will be done on the features selected in chapter 3 by building models using these features and comparing them with other models.

5.1 ASD and ADHD Comorbidity

The data being used to analyze ASD and ADHD comorbidity consists of 254 subjects who have been diagnosed with ASD, out of which 63 subjects have ADHD as well. This data consists of 190 males and 64 females where 53 males and 10 females have ADHD along with

ASD. Initially, K-Means algorithm with PCA has been applied to check if these two groups could be distinguished by our K-means algorithm. However, our means algorithm was not able to distinguish these two groups of children in both two and three dimensions. There was an overlap in the feature values and our unsupervised learning technique couldnt result in pure clusters even as we increased the size of k to 9. So, supervised learning techniques were applied on the data and this is discussed in the next sub section.

5.1.1 Supervised Learning Techniques to predict ADHD in Autistic children

The data used for this analysis has already been preprocessed in the chapter 3, now this data is loaded into WEKA and different machine learning classifiers in WEKA are used to build our models. The different Machine Learning classifiers used for our analysis are Naive Bayes, Logistic Regression, Random Forest, Support Vector Machines and K-Nearest Neighbors. These algorithms have been applied to our data to build models as they are suitable for data and these algorithms work well with the type of data that we have. The results obtained from our machine learning algorithms on our data is given below in table 5.1.

ML Algorithm	**Accuracy**	**Precision**	**Recall**	**ROC Area**
Naive Bayes	92.126%	0.922	0.921	0.956
Logistic Regression	83.46%	0.838	0.835	0.864
Random Forest	96.063%	0.960	0.961	0.965
Support Vector Machine	92.9134%	0.929	0.929	0.900
K Nearest Neighbors (n=3)	90.1575%	0.913	0.902	0.937

Table 5.1: Results of Supervised Learning Techniques to predict ADHD in Autistic children

From the results obtained by different machine learning algorithms it can be seen that Random forest is performing well. The Random Forest algorithm is able to classify 244 subjects correctly. For more information on how our model is performing, we analyzed how Random Forest algorithm performs with both the groups and the model had a precision

of 97% and recall of 98% for children with ASD only and precision of 93.4% and 90% for children with ASD and ADHD. It could be seen that the model is performing well with both the groups. So, Random Forest algorithm yields the best results for our model and could be considered to be the best model. For further analysis of features, the features can successfully distinguish our two groups, the features selected in chapter 3 are used to build are models and then the results by LASSO, ReliefF and RFE feature selection algorithms are shown in table 5.2,5.3 and 5.4 respectively.

ML Algorithm	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**ROC Area**
Naive Bayes	91.73	91.7	91.7	0.930
Logistic Regression	91.7	91.6	91.7	0.959
Random Forest	95.669	95.7	95.7	0.978
Support Vector Machine	92.126	92.5	92.1	0.910
K Nearest Neighbors(n=3)	93.3071	93.6	93.3	0.946

Table 5.2: Results of Supervised Learning Techniques to predict ADHD in Autistic children by LASSO

Random Forest is performing well for diagnosing the comorbidity and K-nearest Neighbor(n=3) is also doing a good job. So, the features selected by LASSO could be used to distinguish if an autistic child has ADHD or not. On the other hand, the features selected by the ReliefF are not doing a good job at analyzing the comorbidity. The best model is Support Vector Machines with a low accuracy of 75%.

ML Algorithm	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**ROC Area**
Naive Bayes	64.17	76.1	64.2	0.721
Logistic Regression	72.83	71.3	72.3	0.705
Random Forest	72.83	71.3	72.3	0.705
Support Vector Machine	75.19	56.5	75.2	0.5
K Nearest Neighbors(n=3)	74.409	72.5	74.4	0.662

Table 5.3: Results of Supervised Learning Techniques to predict ADHD in Autistic children by ReliefF

```
J48 pruned tree
------------------

Quality of Social Overtures <= 1: 1 (156.0/4.0)
Quality of Social Overtures > 1
|   Inappropriate questions or statements <= 1: 1 (26.0/2.0)
|   Inappropriate questions or statements > 1
|   |   Conventional/Instrumental Gestures <= 1: 0 (65.0/8.0)
|   |   Conventional/Instrumental Gestures > 1: 1 (7.0)

Number of Leaves  :     4

Size of the tree :      7
```

Figure 5.1: J48 pruned tree with ADI parent-oriented review

The features selected by RFE are able to distinguish the two subgroups of children extremely well. From table 5.4, it can be seen that all the models are consistent when diagnosing the children. Hence, the ADI parent-oriented features are the best features to diagnose these two subgroups of children.

ML Algorithm	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**ROC Area**
Naive Bayes	694.488	94.6	94.5	0.937
Logistic Regression	94.4	94.6	94.5	0.953
Random Forest	94.488	94.6	94.5	0.942
Support Vector Machine	94.488	94.6	94.5	0.931
K Nearest Neighbors(n=3)	92.9134	93.2	92.9	0.929

Table 5.4: Results of Supervised Learning Techniques to predict ADHD in Autistic children by RFE

J48 algorithm is used to best summarize our model using the ADI features and the pruned tree in shown in figure 5.1 has an accuracy of 94.4% and precision and recall range from 88% to 96%. The J48 results in figure 5.2 are obtained using the BASC and the VINE parent-oriented review. Th accuracy of the resulting model is 80%. When comparing both these trees, it can be seen that the J48 pruned tree resulting from ADI parent-oriented reviews is better than the one from BASC and VINE. Therefore, ADI parent-oriented review is better than BASC or VINE parent-oriented review at distinguishing if an autistic child has ADHD or not.

```
J48 pruned tree
------------------

Attention Problems <= 69: 1 (174.0/24.0)
Attention Problems > 69
|   Attention Problems <= 80
|   |   Attention Problems <= 75: 1 (48.0/19.0)
|   |   Attention Problems > 75
|   |   |   Vineland Communication <= 79: 1 (3.0/1.0)
|   |   |   Vineland Communication > 79: 0 (19.0/1.0)
|   Attention Problems > 80: 1 (10.0/1.0)

Number of Leaves  :     5

Size of the tree :      9
```

Figure 5.2: J48 pruned tree with BASC and VINE parent-oriented review

5.1.2 Individual Feature Sets Analysis

As the date consists of four different types of features, Random Forest model is trained on each of these feature sets. The four different feature sets are IQ, ADI BASC and VINE feature set. The results of these Random Forest models are obtained in table 5.5. When analyzing the IQ feature set, additional data from the ABIDE and ADHD-200 dataset was taken. The subjects from these datasets were taken by random sampling strategy. So, when the Random Forest model was trained on these there groups of data(ASD, ADHD, ASD+ADHD), the performance of the model dropped. It had an accuracy of 60% which is lesser compared to the IQ feature set analysis of groups(ASD, ASD+ADHD). Therefore, from the models obtained, IQ of the child is not a clear indicator of the diagnosis of the child, that is the ASD and ADHD comorbidity cannot be analyzed with IQ.

Feature Set	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**ROC Area**
IQ	65.7%	0.628	0.657	0.588
ADI	96%	0.960	0.961	0.966
BASC	75.98%	0.726	0.760	0.727
VINE	73.6%	0.562	0.736	0.557

Table 5.5: Random Forest algorithm with different feature sets.

By analyzing the models, it can be seen that ADI parent-oriented review is performing better than the other three feature sets. Similar results were obtained when analyzing the feature selection models. Therefore, for distinguishing the two groups ADI parent-oriented review plays a vital and only this review could be used to diagnose the children as well.

5.1.3 Ensemble methods to predict ADHD in Autistic children

Ada Boosting and Logit Boosting are two different types of boosting techniques that have been used on our data. Ada Boost uses weights to give more importance to certain variables over others. It is an iterative method that classifies data in the best way possible at each step. Logit Boosting also works similar to Ada Boost however, the cost function through which weights are given to variables varies in both these algorithms. For logit boosting the cost function is of that applied to logistic regression and minimizes the logistic loss in (1) and the training error function minimized by Ada boosting algorithm at each iteration t is given in (2).

$$\sum_i \log(1 + e^{-y_i f(x_i)}) \tag{1}$$

$$\sum_i E[F_{t-1}(x_i) + \alpha_t h(x_i)] \tag{2}$$

The Ada Boosting technique took 5 iterations and the accuracy of the model is 90.5% while the ROC Area is 0.951. Also, the precision and recall for ASD+ADHD group is 76% and 92% and for the ASD group is 97% and 90%. The features selected by the Ada Boosting algorithm at each iteration are given below:

1. Quality of social overtures- According to Ada Boosting if the value of this feature is equal to 2, then with 60% confidence the class is ASD+ADHD. However, when the feature value is not equal to 2, then with 97% confidence it is ASD. This was selected

in the first iteration with weight of 1.59.

2. Attention Problems- When the subjects feature value is greater than 69.5 than it belongs to ASD class with 52% confidence, but it has higher confidence of 92% when it is less than or equal to 69.5. This was selected in iteration 2 and 4 with weights of 1.47 and 1.24 respectively.
3. Offering to share- When the subjects feature value is not equal to 1 than it belongs to class ASD+ADHD with 63% confidence. On the other hand, when feature value is equal to 1, then it belongs to ASD class with 97% accuracy. This was selected in iteration 3 with weight of 0.9.
4. Inappropriate questions or statements- When the subjects feature value is not equal to 2 than it belongs to class ASD with 87% confidence. However, when feature value is equal to 2, then it belongs to ASD+ADHD class with 66% accuracy. This was selected in iteration 3 5 with weight of 1.18.

Similarly, when Logit Boosting technique has been applied to the data, the accuracy of the model is 92% and the ROC Area is 0.949. The precision and recall for the ASD class is 94% each. On the other hand, the precision and recall for the ASD+ADHD class is 84% each. The features selected by Boosting technique at each iteration is given below:

1. Quality of social overtures
2. Inappropriate questions or statements
3. Attention Problems
4. Response to Approaches of other children
5. Conventional/Instrumental Gestures

5.1.4 Observations

During the comparison of all the three feature selection algorithms, it can be seen that the variable Vineland Socialization of the VINE is most important over others as two of our feature selection algorithms have selected it. Similarly, Attention Problems variable from BASC and Quality of social overtures and Inappropriate questions or statements variables from ADI. These variables that our algorithm has selected are actually the most important signs of ASD and ADHD in individuals when compared to other signs.

The results obtained after applying boosting techniques show that rules can be made for these features in the form of decision stumps, which will help in better diagnosis. Also, Logit Boosting performed better with our data over Ada Boosting. The most important features that have been selected by majority of our machine learning algorithms are given below:

1. Quality of social overtures
2. Inappropriate questions or statements
3. Attention Problems
4. Conventional/Instrumental Gestures

Our data doesnt contain control groups, so the false diagnosis of these disorders was not a problem. But, even in the existing data false negatives of ADHD are minimal, this shows that machine learning is doing a good job at predicting if autistic children are showing signs of ADHD or not. It shows positive signs of being able to distinguish ASD patients from the ASD+ADHD patients with limited variables. These techniques could be applied to larger datasets and our models would be more generalized. Hence, even though these methods are not ready to be used in real-time, the results of our study indicate that in the near future, machine learning will show promising results in the clinical diagnosis of developmental disorders specifically those that do not have genetic relations like ASD and ADHD.

5.2 VCFS and ASD Comorbidity

When analyzing the comorbid disorders ASD and VCFS, the data has been separated for this purpose. It consists of subjects who are diagnosed with ASD, VCFS and both these disorders together. On this modifies data, different machine learning algorithms have been applied. Also, the features selected by different feature selection algorithms in chapter 3 are also used to build models. The observations made in this analysis are discussed in the next section.

5.2.1 Applying Supervised Learning Techniques

The data which consists of the three different class labels that is used for this analysis has already been preprocessed in the chapter 3, now this data is loaded into WEKA and different machine learning classifiers in WEKA are used to build our models. These algorithms have been applied to our data to build models as they are suitable for data and these algorithms work well with the type of data that we have. The results obtained from our machine learning algorithms on our data is given below in table 5.6.

ML Algorithm	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**ROC Area**
Naive Bayes	94.11	0.953	0.941	0.994
Logistic Regression	89.8	0.902	0.899	0.963
Random Forest	96.7	0.968	0.967	0.997
Support Vector Machine	97.0	0.970	0.971	0.976
K Nearest Neighbors (n=3)	93.7	0.940	0.938	0.980

Table 5.6: Supervised Learning techniques for ASD and VCFS comorbidity

By comparing the different supervised learning techniques both SVM and Random Forest algorithms have similar model performance. Also, both these models are better than other machine learning techniques that have been used. When the models are compared based on each of the class labels, then SVMs are better than Random Forest. Using the entire features

```
J48 pruned tree
------------------

Functional Communication <= 49: ASD (157.0)
Functional Communication > 49
|   Performance IQ <= 100
|   |   Hand and Finger mannerisms = 0
|   |   |   Repetitive use of objects or interest in parts of objects = 0: VCFS (90.0/1.0)
|   |   |   Repetitive use of objects or interest in parts of objects = 1
|   |   |   |   Spontaneous imitation of actions = 0: VCFS (7.0)
|   |   |   |   Spontaneous imitation of actions = 1: VCFS (5.0/1.0)
|   |   |   |   Spontaneous imitation of actions = 2: VCFS+ASD (3.0)
|   |   |   Repetitive use of objects or interest in parts of objects = 2: VCFS+ASD (3.0/1.0)
|   |   Hand and Finger mannerisms = 1
|   |   |   Age of first single words = 0: VCFS (2.0)
|   |   |   Age of first single words = 1: ASD (10.0/1.0)
|   |   Hand and Finger mannerisms = 2
|   |   |   Attention Problems <= 65: ASD (5.0)
|   |   |   Attention Problems > 65: VCFS+ASD (6.0/1.0)
|   Performance IQ > 100: ASD (18.0)

Number of Leaves  :     11

Size of the tree :      18
```

Figure 5.3: J48 pruned tree with ADI feature set

available for the data, the J48 algorithm summarized our model. The results obtained from our J48 algorithm are shown in figure **??** For further analysis of these comorbid disorders, SVM models are built on each of the features selected by the feature selection algorithms in chapter 3. The results obtained by these SVM models are shown in table 5.7.

Feature Selection Algorithm	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**ROC Area**
LASSO	62.09	0.389	0.621	0.498
ReliefF	78.43	0.765	0.784	0.805
RFE	90.5	0.873	0.905	0.909

Table 5.7: Feature Selection Algorithms for ASD and VCFS comorbidity

SVM model trained by RFE features is the best and the performance of the other to models is not comparable to our previous models. If this model was compared to SVM model with entire dataset, it could be seen that the performance of the model has dropped by 7%.

5.2.2 Individual Feature Sets Analysis

In this section, ASD and VCFS comorbidity is analyzed with respect to each of the individual feature sets present in our data. The SVM model which is performing well with our data is trained on each of the individual feature sets. The results of these models are shown in

table 5.8. By comparing these models, ADI parent oriented reviews model performance is better than other feature sets. However, when compared to the previous models of SVM, the performance is less, but this model could be a better generalization of our data.

Feature Set	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**ROC Area**
IQ	63.07%	0.566	0.631	0.518
ADI	93.13%	0.927	0.931	0.936
BASC	84.96%	0.841	0.850	0.874
VINE	80.3%	0.785	0.804	0.825

Table 5.8: Random Forest algorithm with different feature sets.

5.2.3 Applying Ensemble Methods

Ada Boosting and Logit Boosting techniques were applied to the ASD and VCFS data and the performance of these models has been analyzed. The accuracy of Ada Boosting technique is 92.15% and ROC Area is 0.978. The Ada model takes 5 iterations and the features selected at each iteration are given below:

1. Criteria for Repetitive behaviors and stereotyped patterns- When this feature value is equal to Yes, then child belongs to the class label ASD with 90% confidence. However, when the feature value is not Yes, then child belong to class label VCFS with 88% confidence. The weight given to this feature is 2.18.
2. Functional Communication- The subject belongs to the class ASD, when the feature value is less than or equal to 49.5 with 100% confidence. On the other hand, individuals belongs to class VCFS with less confidence of 54%, if the feature value is greater than 49.5. The weight assigned to this feature is 1.08.
3. Adaptive Daily Living- The weight for this feature is 0.67. When this feature takes a value less than of equal to 49 subjects belongs to ASD class with 100% confidence. However, when the value is greater than 49, with 50% confidence the subjects belongs

to the VCFS and ASD class.

4. Vineland Composite- The weight assigned for this feature is 0.4. The class label is VCFS with 42% confidence, if the feature value is less than 98.5. Also, the class label is ASD with 97% confidence, if feature value is greater than 98.5
5. Criteria for Communication- When this feature value is equal to Yes, then subject belongs to the class label VCFS+ASD with 50% confidence. However, when the feature value is not Yes, then child belong to class label VCFS with 64% confidence. The weight given to this feature is 0.22.

Logit Boosting has a model accuracy of 95.7% and ROC Area value of .988. It is performing better than Ada boosting. Also, for subjects that both these disorders, the model performance is better. The features selected by Login Boosting are given below:

1. Functional Communication
2. Criteria for Repetitive behaviors and stereotyped patterns
3. Vineland Daily Living
4. Criteria for Qualitative impairments in reciprocal social interaction
5. Quality of social overtures
6. Adaptive Daily Living
7. Hyperactivity
8. Performance IQ
9. Criteria for Communication
10. Spontaneous imitation of actions
11. Attention Problems
12. Verbal IQ
13. Somatization

5.2.4 Observations

As our data consists of less number of children who are diagnosed with VCFS and ASD, the models that are being trained on this data are able to precisely detect those children. However, it cannot be guaranteed that our models are able to generalize this class label.Our models are doing a good job with the individual diagnosis of ASD and VCFS especially the Support Vector Machine algorithm. The best features which help in analyzing our comorbid disorders are given below:

1. Functional Communication
2. Criteria for Communication
3. Spontaneous imitation of actions
4. Adaptive Daily Living
5. Criteria for Repetitive behaviors and stereotyped patterns
6. Repetitive use of objects or interest in parts of objects

On comparison of the feature sets, ADI parent-oriented review is better than the other two parent-oriented reviews. Also, the IQ feature set is not good at analyzing the comorbidity of ASD and VCFS. Overall, due to the difference in the class proportions, the machine learning algorithms are not able to categorize when generally. Also, RFE has selected features from ADI parent-oriented reviews and it is better than other feature selection algorithms. When comparing the boosting techniques, Logit is doing better, but it has chosen a variety of features. Therefore, for analyzing ASD and VCFS comorbidity, ADI parent-oriented review provide better generalization and so does the J48 model.

Chapter 6

Investigating Developmental Disorders Diagnosis

The target label of the data is modified to binary to examine the individual developmental disorders. After converting the label, each of the disorders are studied individually in this chapter. In the first section, ASD features and models are discovered, while in the second section, ADHD features and models are discovered. Finally, the last section VCFS features and models are built. Each of these sections help us understand these disorders and show results for children if these disorders existed exclusively in the children. This chapter discusses more specifically about their individual existence in the children. Also, impact of each of these individual parent-oriented reviews along with the IQ features on the developmental disorders will be discussed. Discussions are also done on the models that will we built for each of the developmental disorders using the features selected in chapter 3 by each of the feature selection algorithms.

ML Algorithm	Accuracy	ROC Area
Random Forest	98.64%	0.99137
Logistic Regression	98.64%	0.96875
K- Nearest Neighbors(n=3)	85.135%	0.79202
Decision Tree	97.297%	0.9375

Table 6.1: Supervised learning techniques comparison based on different metrics for ASD

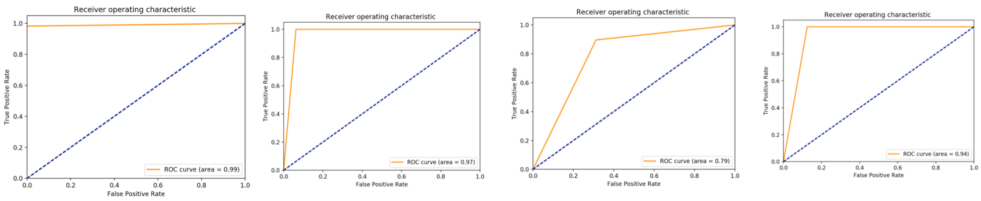

Figure 6.1: ROC curve for ASD dataset
a) Random Forest b) Logistic Regression c) KNN(n=3) d) Decision Tree

6.1 Austism Spectrum Disorder

The target label diagnosis in the subgroup data has been modified to predict if a child has ASD or not. The data is modifies to understand and build models specific to ASD. This data will analyze ASD related features and models in greater depth. On this dataset different supervised learning techniques Logistic Regression, Decision Trees, K-Nearest Neighbors and Random Forest have been applied from the sklearn package in python and the following is the ROC curves obtained in figure 6.1. The metrics for evaluation of these techniques are given in the table 6.1.

Among these supervised learning model, Random forest seems to be better fitting the ‘ASD data‘ when compared to other models. It has better Area Under Curve value than Logistic Regression model, even though the accuracy is same. By analyzing the results, it can be observed that tree based algorithms are doing a good job and the model for examining ASD can be repressed in the form of a tree. So, J48 algorithm is applied to describe the tree for our data, which can be seen in figure 6.2. The accuracy of J48 algorithm is 96% and ROC

```
J48 pruned tree
------------------

Criteria for Repetitive behaviors and stereotyped patterns = 0
|   Performance IQ <= 100
|   |   Adaptability <= 40
|   |   |   Withdrawal <= 64: 0 (12.0)
|   |   |   Withdrawal > 64
|   |   |   |   VCI <= 90: 0 (3.0/1.0)
|   |   |   |   VCI > 90: 1 (5.0)
|   |   Adaptability > 40: 0 (82.0)
|   Performance IQ > 100: 1 (6.0)
Criteria for Repetitive behaviors and stereotyped patterns = 1
|   Criteria for Qualitiative impairments in reciprocal social interaction = 0: 0 (5.0/1.0)
|   Criteria for Qualitiative impairments in reciprocal social interaction = 1
|   |   Criteria for Communication = 0
|   |   |   Offers Comfort = 0: 0 (2.0)
|   |   |   Offers Comfort = 1: 1 (15.0)
|   |   |   Offers Comfort = 2: 0 (2.0)
|   |   Criteria for Communication = 1: 1 (237.0)

Number of Leaves  :     10

Size of the tree :      18
```

Figure 6.2: J48 pruned tree for ASD

Area is 0.96. Also, for each individual group, the precision and recall are good. By combining the precision and recall, it can be seen that the F-measure for ASD subjects is 97% and the F-measure for non-ASD subjects is 93%. This shows that J48 is doing a good job at summarizing the results of the ASD data.

When training the Random Forest model, based on the features selected in chapter 3, table 6.2 shows the results of each of the feature selection algorithm. The features selected by ReliefF and RFE are performing better than the features selected by LASSO. The features selected by RFE is from the ADI parent-oriented review and hence, the ADI parent oriented review seems to be building better models when compared to the other parent oriented reviews.

Feature Selection Algorithm	**Accuracy**	**F-measure**	**ROC Area**
LASSO	82.6	0.824	0.898
ReliefF	91.3	0.915	0.967
RFE	95.3	0.954	0.959

Table 6.2: Random Forest model trained with Feature Selection Algorithms for ASD

6.1.1 Predicting Diagnosis based on IQ feature set

Based on the IQ features, different supervised learning techniques (random Forest, Naive Bayes, Logistic Regression) have been applied to predict ASD diagnosis for given dataset. These algorithms are used from WEKA by loading this data into it. The table 6.3 shows the various metrics based on which the techniques are evaluated.

ML Technique	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.878	0.878	0.878	0.930	87.80%
Naive Bayes	0.693	0.688	0.691	0.686	68.83%
Logistic Regression	0.683	0.713	0.689	0.796	71.273%

Table 6.3: Supervised Learning Techniques applied on IQ features/variables for ASD

Random Forest supervised learning technique is good at predicting ASD based on the IQ scores of the children with an accuracy of 87.8%. The accuracy for model has dropped from 96% as the features have been reduced to only IQ features. In comparison, it can be said that IQ features are not doing well compared to the entire feature set, so IQ feature set only is not a suitable method for predicting if a child has ASD or not.

6.1.2 Predicting Diagnosis based on ADI feature set

Further analysis has been done based on the ADI features, different supervised learning techniques have been applied to predict ASD diagnosis for given dataset. The table 6.4 below shows the various metrics based on which the supervised learning techniques are evaluated.

Random Forest supervised learning technique is good at predicting ASD based on the ADI feature/ variable scores of the children. Also, the ADI review is specifically designed for finding ASD behaviors in a child and hence, most models have high accuracy along with high

ML Technique	Precision	Recall	F-measure	ROC Area	Accuracy
Random Forest	0.968	0.967	0.968	0.991	96.74%
Naive Bayes	0.948	0.949	0.948	0.960	94.85%
Logistic Regression	0.918	0.916	0.917	0.954	91.59%

Table 6.4: Supervised Learning Techniques applied on ADI feature set for ASD

ROC values. The accuracy of this model is similar to that of the J48 model and hence, ADI feature set could be used individually to diagnose a child with ASD.

6.1.3 Predicting Diagnosis based on BASC feature set

Based on the BASC feature set, different supervised learning techniques have been applied to predict ASD diagnosis for given dataset. The table 6.5 shows the various metrics based on which the techniques are evaluated.

ML Technique	Precision	Recall	F-measure	ROC Area	Accuracy
Random Forest	0.920	0.908	0.910	0.965	90.78%
Naive Bayes	0.915	0.892	0.895	0.928	89.15%
Logistic Regression	0.879	0.875	0.877	0.936	87.53%

Table 6.5: Supervised Learning Techniques applied on BASC feature set for ASD

Random Forest supervised learning technique is good at predicting ASD based on the BASC feature scores of the children with an accuracy of 90.78%. Even though the model accuracy has dropped to 91% when compared to the J48 model, BASC features could still be used to predict if a child has ASD or not.

6.1.4 Predicting Diagnosis based on VINE feature set

Based on VINE feature set, different supervised learning techniques from WEKA have been applied to predict ASD diagnosis for given dataset. The table 6.6 shows the various metrics based on which the techniques are evaluated. Random Forest and Logistic Regression supervised learning techniques are good at predicting ASD based on the VINE feature scores of the children with an accuracy of 88%. However, the performance of both these models is less than 10% when compared tot he J48 model and the model with ADI features.

ML Technique	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.884	0.881	0.882	0.941	88.07%
Naive Bayes	0.792	0.786	0.788	0.869	78.59%
Logistic Regression	0.885	0.883	0.884	0.917	88.34%

Table 6.6: Supervised Learning Techniques applied on VINE feature set for ASD

On the other hand, even though the model with VINE features is not doing well, it is better than the model with IQ features. So, only the VINE parent-oriented features cannot be used to diagnose a child with ASD.

6.1.5 Observations

When diagnosing a child with ASD, it can be seen that tree-based machine learning techniques like Random Forest and J48 are doing well. Also, it possible to diagnose a child with ASD using ADI and BASC parent-oriented review, but ADI parent-oriented review seems to be performing better than both the other parent oriented reviews. This could also be seen as the RFE feature selection algorithm in chapter 3, selects features from the ADI parent-oriented review and that is the best model from our feature selection models. The important features selected by the J48 model are a combination of all three parent-oriented reviews and are

given below:

1. Offers Comfort
2. Criteria for Qualitative impairments in reciprocal social interaction
3. Criteria for Communication
4. Criteria for repetitive behaviors and stereotyped patterns
5. Adaptability
6. Withdrawal
7. Performance IQ

The features selected from the different feature selection algorithms in chapter 3 are compared to these features and 3 out of 7 of these features are same. It can be seen that J48 is converging the features selected by each of those algorithms. On an average, all the models to predict ASD have an accuracy of 90%. The model with highest accuracy is with ADI parent-oriented review features. Therefore, it is observed that rather than using all the features in the data given, only the features of ADI parent-oriented reviews are sufficient, this supports the fact that ADI parent-oriented review is used to diagnose children with ASD.

6.2 Attention Deficit/Hyperactivity Disorder

The target label diagnosis in the subgroup data has been modified to predict if a child has ADHD or not. The dataset being used for this is converted, that is the class label is modified. On this dataset different supervised learning techniques Logistic Regression, Decision Trees, Naive Bayes and Random Forest have been applied and the following is the ROC curves obtained in figure **??**. The metrics for evolution of these techniques are given in the table 6.7.

Among these supervised learning models, Random Forest seems to be better fitting the ‘ADHD dataset‘ when compared to other models. It has better Area Under Curve value

ML Algorithm	Accuracy	ROC Area
Random Forest	95.945%	0.90625
Logistic Regression	77.027%	0.604525
K- Nearest Neighbors(n=3)	79.792%	0.8028
Decision Tree	93.24%	0.8890

Table 6.7: Supervised learning techniques based on different metrics for ADHD

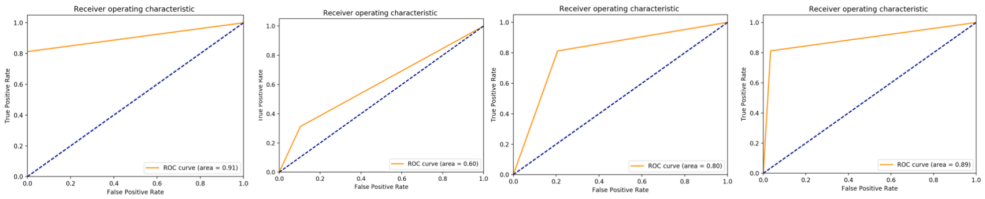

Figure 6.3: ROC curve for ADHD dataset
a) Random Forest b) Decision Tree c) Naive Bayes d) Logistic Regression

and accuracy. As the Random Forest is doing well for ADHD data, the features selected by each of the feature selection algorithm are trained using the random forest model and the performance of these three model is given in table 6.8.

Feature Selection Algorithm	Accuracy	F-measure	ROC Area
LASSO	82.9%	0.813	0.806
ReliefF	78.3%	0.764	0.722
RFE	95.9%	0.959	0.942

Table 6.8: Random Forest model trained with Feature Selection Algorithms for ADHD

When comparing the performance of models trained with the features from the three feature selection algorithms, it can be seen that RFE is doing better than the other two. Even though BASc is more commonly used to diagnose children with ADHD, this shows that ADI features can also do a good job at recognizing children with ADHD.

As the random forest is doing well, another tree-based algorithm J48 is used to summarize our model and the prunes tree obtained from J48 algorithm is given in figure6.4. The algorithm had an accuracy of 94% and it was good at diagnosing if children had ADHD or not. The

```
J48 pruned tree
------------------

Quality of social overtures = 0: 0 (84.0)
Quality of social overtures = 1: 0 (174.0/4.0)
Quality of social overtures = 2
|   Inappropriate questions or statements = 0: 0 (15.0)
|   Inappropriate questions or statements = 1: 0 (20.0/2.0)
|   Inappropriate questions or statements = 2
|   |   Offering to share = 0: 0 (2.0)
|   |   Offering to share = 1: 0 (3.0)
|   |   Offering to share = 2
|   |   |   Vineland Socialization <= 97
|   |   |   |   BPAGGR <= 68: 0 (11.0)
|   |   |   |   BPAGGR > 68: 1 (4.0)
|   |   |   Vineland Socialization > 97
|   |   |   |   Withdrawal <= 82: 1 (50.0)
|   |   |   |   Withdrawal > 82
|   |   |   |   |   Atypicality <= 67: 1 (3.0)
|   |   |   |   |   Atypicality > 67: 0 (3.0)

Number of Leaves  :     11

Size of the tree :      18
```

Figure 6.4: J48 pruned tree for ADHD

F-measure for diagnosing ADHD is 83% and without ADHD is 97%. Since our data has small percentage of children when compared to those with ADHD, it can be seen that 83% is comparable. Also, since the negative diagnosis percentage is high, it means that our model is doing well with trawl negative and can certainly till if a children doesn't have ADHD with good accuracy.

6.2.1 Predicting Diagnosis based on IQ feature set

Based on the IQ feature set, different supervised learning techniques (Random Forest, Naive Bayes, Logistic Regression) have been applied to predict ADHD diagnosis for the modifies data. The table 6.9 shows the various metrics based on which the techniques are evaluated.

Logistic Regression supervised learning technique is good at predicting ADHD based on the IQ scores of the children. When comparing the performance of our model with the J48 model, it can be seen that the accuracy has dropped by 11% and hence, only IQ features not

ML Technique	Precision	Recall	F-measure	ROC Area	Accuracy
Random Forest	0.770	0.799	0.782	0.714	79.94%
Naive Bayes	0.686	0.818	0.746	0.556	81.84%
Logistic Regression	0.792	0.832	0.772	0.687	83.19%

Table 6.9: Supervised Learning Techniques applied on IQ features/variables for ADHD

sufficient at predicting if a child has ADHD.

6.2.2 Predicting Diagnosis based on ADI feature set

Now, further analysis is done based on the ADI feature set, different supervised learning techniques have been applied to predict ADHD diagnosis for modified data. The table 6.10 shows the various metrics based on which the techniques are evaluated.

ML Technique	Precision	Recall	F-measure	ROC Area	Accuracy
Random Forest	0.981	0.981	0.981	0.996	98.10%
Naive Bayes	0.934	0.919	0.923	0.974	91.86%
Logistic Regression	0.977	0.976	0.976	0.687	97.56%

Table 6.10: Supervised Learning Techniques applied on ADI features/variables for ADHD

Random Forest supervised learning technique is good at predicting ADHD based on the ADI feature scores of the children and this also has high accuracy and ROC values for different classifier models. The Random Forest model with ADI feature set is better than the J48 model and Random Forest taking all the features into consideration. So, the J48 pruned tree using only the ADI features is given in figure 6.5. The accuracy of this model is 94.5%, which is comparable to the previous J48 model. Also, the F-measure for diagnosing children with ADHD is 84.5% which is more than our previous model diagnosis and for a children not having ADHD, the F-measure is 97%. So, the ADI feature set is doing a better job at predicting if the child has ADHD better than the entire features of the data.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

quality of social overtures = 0: 0 (84.0)
quality of social overtures = 1: 0 (174.0/4.0)
quality of social overtures = 2
|   Inappropriate questions or statements = 0: 0 (15.0)
|   Inappropriate questions or statements = 1: 0 (20.0/2.0)
|   Inappropriate questions or statements = 2
|   |   Response to approaches of other children = 0: 0 (2.0)
|   |   Response to approaches of other children = 1: 0 (11.0/2.0)
|   |   Response to approaches of other children = 2
|   |   |   Offers Comfort = 0: 1 (0.0)
|   |   |   Offers Comfort = 1
|   |   |   |   Range of Facial Expressions = 0: 0 (0.0)
|   |   |   |   Range of Facial Expressions = 1: 1 (2.0)
|   |   |   |   Range of Facial Expressions = 2: 0 (2.0)
|   |   |   |   Range of Facial Expressions = 3: 0 (0.0)
|   |   |   Offers Comfort = 2: 1 (59.0/6.0)

Number of Leaves  :     12

Size of the tree :      17
```

Figure 6.5: J48 pruned tree for ADHD with ADI feature set

6.2.3 Predicting Diagnosis based on BASC feature set

Additional analysis is done using different supervised learning techniques to predict ADHD diagnosis for our converted data based on the BASC features/variables. The table 6.11 shows the various metrics based on which the techniques are evaluated. Some of the supervised learning techniques applied from WEKA are Random Forest, Naive Bayes and Logistic Regression.

ML Technique	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.794	0.832	0.793	0.768	83.19%
Naive Bayes	0.792	0.759	0.773	0.710	75.88%
Logistic Regression	0.760	0.802	0.775	0.709	80.21%

Table 6.11: Supervised Learning Techniques applied on BASC features for ADHD

Random Forest supervised learning technique is good at predicting ADHD based on the BASC scores of the children. However, when compared to the ADI model, our Random Forest classifier trains don BASC features has 10% lower accuracy.

6.2.4 Predicting Diagnosis based on VINE feature set

Now, VINE features are applied to different supervised learning techniques to predict ADHD diagnosis for modifies data (target label modified to binary). The table 6.12 shows the various metrics based on which the techniques are evaluated.

ML Technique	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.685	0.813	0.744	0.689	81.30%
Naive Bayes	0.851	0.580	0.627	0.735	57.9946%
Logistic Regression	0.687	0.824	0.749	0.688	82.38%

Table 6.12: Supervised Learning Techniques applied on VINE features for ADHD

Logistic Regression supervised learning technique is good at predicting ADHD based on the VINE scores of the children. When the Logistic Regression model is compared to the ADI models, the performance of the model is not good, however, its performance is comparable to the model with BASC.

6.2.5 Observations

The analysis done in the above section shows that ADI parent-oriented review feature set is doing good at predicting if a child has ADHD. Moreover, our analysis shows that our models are good at eliminating true negatives that is there are better at diagnosing if the child doesn't have ADHD. The important ADI features, which can be used to diagnose a child with ADHD are given below:

1. Quality of social overtures
2. Inappropriate statements or questions
3. Group play with peers or friendships
4. Offers comfort
5. Response of approaches to other children
6. Range of facial expressions

Most of these features selected were selected by our feature selection algorithms in chapter 3. Also, most of these features are important symptoms of diagnosing a child with ADHD.

6.3 22Q Deletion Syndrome

The target label diagnosis in the subgroup data has been modified into a binary attribute to predict if a child has ADHD or not. On this dataset different supervised learning techniques like Logistic Regression, Decision Trees, K-Nearest Neighbors and Random Forest have been applied and the following is the ROC curves obtained are present in figure**??**. These techniques have been applied from the sklearn package of python. The metrics for evaluation of these techniques are given in the table 6.13.

ML Algorithm	**Accuracy**	**ROC Area**
Random Forest	98.648%	0.99107
Logistic Regression	97.297%	0.98214
K- Nearest Neighbors(n=3)	91.891%	0.88988
Decision Tree	83.7837%	0.79861

Table 6.13: Supervised learning techniques based on different metrics for VCFS

Among these supervised learning models, Random Forest seems to be better fitting the 'VCFS dataset' when compared to other models, it has better Area Under Curve value and accuracy. Even Logistic Regression algorithm is fitting the data well and has a good accuracy of 97%.

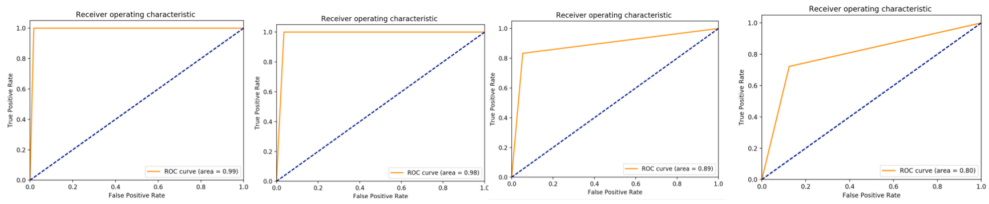

Figure 6.6: ROC curve for VCFS dataset
a) Random Forest b) Decision Tree c) Naive Bayes d) Logistic Regression

```
J48 pruned tree
------------------

Criteria for qualitative impairments in reciprocal social interaction = 0
|   Adaptability <= 35
|   |   Performance IQ <= 95: 1 (2.0)
|   |   Performance IQ > 95
|   |   |   Adaptive Skills <= 28: 0 (3.0)
|   |   |   Adaptive Skills > 28: 1 (3.0/1.0)
|   Adaptability > 35: 1 (83.0/1.0)
Criteria for qualitative impairments in reciprocal social interaction = 1
|   Vineland Composite <= 59
|   |   Neologisms/idiosyncratic language  = 0: 1 (21.0/1.0)
|   |   Neologisms/idiosyncratic language  = 1: 0 (2.0)
|   |   Neologisms/idiosyncratic language  = 2: 1 (0.0)
|   |   Neologisms/idiosyncratic language  = 3: 1 (0.0)
|   Vineland Composite > 59
|   |   Criteria for repetitive behaviors and stereotyped patterns = 0
|   |   |   Vineland Daily Living <= 82: 1 (6.0/1.0)
|   |   |   Vineland Daily Living > 82: 0 (7.0)
|   |   Criteria for repetitive behaviors and stereotyped patterns = 1: 0 (242.0/4.0)

Number of Leaves  :     11

Size of the tree :      19
```

Figure 6.7: J48 pruned tree for VCFS

Using the J48 algorithm, the pruned tree in figure 6.7 is a summarization of the model. The accuracy of this J48 is 93.4%. The F-measure of diagnosing the children with VCFS is 90% and the F-measure of not diagnosing the child with VCFS is 95.3%. The performance of this model is not has good as the Random Forest model or Logistic Regression model, but it is comparable and a good way of representing our data. Now, the features selected in chapter 3 are used to train the Random Forest model as it is performing the best with our data and the results of the models for each feature selection algorithm are given in table 6.14. Among all the feature selection algorithms, LASSO is performing the best.

Feature Selection Algorithm	Accuracy	F-measure	ROC Area
LASSO	92.4%	0.924	0.974
ReliefF	787.2%	0.873	0.922
RFE	90.2%	0.903	0.951

Table 6.14: Random Forest model trained with Feature Selection Algorithms for VCFS

6.3.1 Predicting Diagnosis based on IQ feature set

Based on the IQ feature set, different supervised learning techniques have been applied to predict VCFS diagnosis for given dataset. The table 6.15 shows the various metrics based on which the techniques are evaluated.

ML Technique	Precision	Recall	F-measure	ROC Area	Accuracy
Random Forest	0.883	0.883	0.883	0.948	88.34%
Naive Bayes	0.692	0.694	0.693	0.712	69.37%
Logistic Regression	0.747	0.759	0.746	0.825	75.88%

Table 6.15: Supervised Learning Techniques applied on IQ features/variables for VCFS

Out of all the supervised learning techniques, Random Forest is the best at predicting VCFS for children. However, compared to the previous Random Forest and J48 model, the accuracy has dropped by 7% and hence, predicting ADHD with only the IQ feature set is not a good idea.

6.3.2 Predicting Diagnosis based on ADI feature set

Based on the ADI review feature set, different supervised learning techniques have been applied to predict VCFS diagnosis for given dataset. The table 6.16 shows the various metrics based on which the techniques are evaluated.

ADI feature set is doing better than IQ feature set when diagnosing children with VCFS. The

ML Technique	Precision	Recall	F-measure	ROC Area	Accuracy
Random Forest	0.943	0.943	0.942	0.981	94.30%
Naive Bayes	0.913	0.913	0.913	0.927	91.32%
Logistic Regression	0.922	0.921	0.922	0.943	92.14%

Table 6.16: Supervised Learning Techniques applied on ADI feature set for VCFS

best learning technique is Random Forest, but the performance of this model is low when compared to the Random Forest model with the entire dataset. However, ADI parent-oriented review features could be used to predict if a child has VCFS.

6.3.3 Predicting Diagnosis based on BASC feature set

Based on the BASC review feature set, different supervised learning techniques(Random Forest, Logistic Regression, Naive Bayes) have been applied to predict VCFS diagnosis for given dataset. The table 6.17 shows the various metrics based on which the techniques are evaluated.

ML Technique	Precision	Recall	F-measure	ROC Area	Accuracy
Random Forest	0.956	0.951	0.952	0.983	95.12%
Naive Bayes	0.927	0.916	0.918	0.933	91.59%
Logistic Regression	0.896	0.894	0.895	0.954	89.43%

Table 6.17: Supervised Learning Techniques applied on BASC feature set for VCFS

Among the different supervised learning techniques for predicting VCFS, Random Forest supervised learning technique is good to predict based on the BASC feature/variable scores of the children with accuracy of 95%. So, the BASC parent oriented review is better than the ADI parent oriented review for predicting VCFS.

6.3.4 Predicting Diagnosis based on VINE feature set

Based on the VINE review feature set, different supervised learning techniquesrandom Forest, Naive Bayes, Logistic Regression have been applied to predict VCFS diagnosis for given dataset. These supervised learning techniques are applied from WEKA. The table 6.18 shows the various metrics based on which the techniques are evaluated.

ML Technique	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Accuracy**
Random Forest	0.887	0.886	0.887	0.961	88.16%
Naive Bayes	0.812	0.808	0.809	0.897	80.75%
Logistic Regression	0.891	0.892	0.891	0.928	89.15%

Table 6.18: Supervised Learning Techniques applied on VINE feature set for VCFS

Even in this case for predicting VCFS, Random Forest supervised learning technique is good to predict based on the VINE feature scores of the children with accuracy of 88% as it has a better ROC value when compared with Logistic Regression which has an accuracy of 89%. The performance of this model is comparable to the performance of the IQ feature set models, but its performance is lower than all other models. As the children who have taken VINE is only 47%, the model performance is good in comparison and it could be more generalized than other models.

6.3.5 Observations

Tree-based machine learning algorithms particularly Random Forest algorithm is doing best for diagnosing children with VCFS. Among the four different feature sets, BASC parent oriented reviews are doing the best. Even the though the performance of BASC parent-oriented review is good and comparable, the model with the entire features is out performing all the other models. The best features in the pruned tree are a combination of all the four

feature sets present in our data. Also, the best features selected by the J48 algorithm are as follows:

1. Adaptability
2. Criteria for Qualitative impairments in reciprocal social interaction
3. Performance IQ
4. Adaptive skills
5. Vineland Composite
6. Neologisms/idiosyncratic language
7. Criteria for Repetitive behaviors and stereotyped patterns
8. Vineland Daily Living

Out of the 8 features selected by J48 algorithm, 6 of them were selected by our feature selection algorithms. Hence, these features are important for diagnosing a child with VCFS and no individual feature set out of the four feature sets could be used to diagnose children with VCFS. However, these features have more importance over other features in the given data and a model trained with these features performs well for predicting VCFS in children.

Chapter 7

Conclusion

Your Conclusions here. During our research, the main problem that was trying to be solved was early intervention of developmental disorders. Researchers in the past have shown that machine learning is useful to diagnose children with various disorders. So, by applying different machine learning techniques, different models were built to diagnose different developmental disorders. Also, models were built to understand the co-occurrences of these disorders. Apart for this our research also focused on analyzing the importance of each reviews to the diagnose and more specifically features which are an indicators of these developmental disorders.

Among the various supervised learning techniques, most of the times Random Forest models performed exceptionally well with our data. On the other hand, for feature selection, RFE was able to select the important features from our feature set. The important findings from our analysis are as follows:

- Most of our models predict the diagnosis labels in the subgroup data for male children with a better accuracy of 7% when compared to female children.
- IQ features predict subgroup data diagnosis with 66.32%, diagnose ASD with 87.8%, diagnose ADHD with 83.19% and diagnose VCFS with 88% accuracy. Overall, IQ cannot be used to diagnose the subgroup data, but it could help with diagnosing ASD,

ADHD and VCFS separately.

- BASC features predict subgroup data with 75%, diagnose ASD with 90%, diagnose ADHD with 83% and diagnose VCFS with 95% accuracy. These tests have a low prediction rate when compared to other tests for predicting the diagnosis of the subgroup data.

- VINE features predict subgroup data with 69%, diagnose ASD with 88%, diagnose ADHD with 82% and diagnose VCFS with 89% accuracy. This has the least accuracy values when compared to other tests, the main reason behind this could be that the number of children who have taken VINE test is less when compared to other two tests and the features of this VINE test is less when compared to the rest two tests.

- ADI review features predict subgroup data with 96.47%, diagnose ASD with 96%, diagnose ADHD with 98% and diagnose VCFS with 94% accuracy. ADI review test is better in predicting the diagnosis labels of subgroup data when compared to both the other tests.

- The comorbid disorders ASD and ADHD could be identified with ADI parent oriented reviews and there exist some important features on which the models achieved an average accuracy of 94%.

- Models could identify ASD and VCFS individually, but identifying their co-occurrence was more complex. The models built for ASD and VCFS comorbidity had an average accuracy of 90%.

- When comparing the individual diagnosis of children, it could be seen that predicting VCFS (98%) among children with given features is better when compared to ASD and ADHD. Also, when clustering the children into different groups, the children diagnosed with VCFS were clustered appropriately (100%) when compared to the ASD cluster.

Our analysis shows machine learning is good at identifying these developmental disorders and they can help clinicians in diagnosing children with these orders. The models that have been found can also bee used to better emphasis on features more closely related to this developmental disorders. As our models identify comorbidity as well, these models would better assist clinicians when diagnosing children with multiple disorders.

The results and observations made in this research are a step towards using machine learning models to diagnose developmental disorders. Further analysis in this field will help us avoid confusions between different parent-oriented reviews and help us in justifying the importance of certain features over others during diagnosis. In the future, more studies could work on developing diagnostic specific models that will assess the disorder in children and their co-occurrences as well in an efficient and swift manner.

References

[1] *AutismSpeaks*. 2018. URL: https://www.autismspeaks.org/what-autism/symptoms (cit. on p. 3).

[2] Deborah L Christensen, Deborah A Bilder, Walter Zahorodny, Sydney Pettygrove, Maureen S Durkin, Robert T Fitzgerald, Catherine Rice, Margaret Kurzius-Spencer, Jon Baio, and Marshalyn Yeargin-Allsopp. "Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network". In: *Journal of Developmental & Behavioral Pediatrics* 37.1 (2016), pp. 1–8 (cit. on p. 3).

[3] Anne S Bassett and Eva WC Chow. "22q11 deletion syndrome: a genetic subtype of schizophrenia". In: *Biological psychiatry* 46.7 (1999), pp. 882–891 (cit. on p. 4).

[4] Anne S Bassett, Eva WC Chow, Janice Husted, Rosanna Weksberg, Oana Caluseriu, Gary D Webb, and Michael A Gatzoulis. "Clinical features of 78 adults with 22q11 deletion syndrome". In: *American Journal of Medical Genetics Part A* 138.4 (2005), pp. 307–313 (cit. on p. 4).

[5] DP Wall, J Kosmicki, TF Deluca, E Harstad, and VA Fusaro. "Use of machine learning to shorten observation-based screening and diagnosis of autism". In: *Translational*

psychiatry 2.4 (2012), e100 (cit. on pp. 4, 18).

[6] Fourth Edition. *Diagnostic and statistical manual of mental disorders*. Am Psychiatric Assoc, 2013 (cit. on p. 5).

[7] Susanna N Visser, Melissa L Danielson, Rebecca H Bitsko, Joseph R Holbrook, Michael D Kogan, Reem M Ghandour, Ruth Perou, and Stephen J Blumberg. "Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003–2011". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 53.1 (2014), pp. 34–46 (cit. on p. 5).

[8] William J Barbaresi, Robert C Colligan, Amy L Weaver, Robert G Voigt, Jill M Killian, and Slavica K Katusic. "Mortality, ADHD, and psychosocial adversity in adults with childhood ADHD: a prospective study". In: *Pediatrics* 131.4 (2013), pp. 637–644 (cit. on p. 5).

[9] Angelo M DiGeorge. "Congenital absence of the thymus and its immunologic consequences: concurrence with congenital hypoparathyroidism". In: *Birth defects original article series* 4 (1968), p. 116 (cit. on p. 5).

[10] Angelo Restivo, Anna Sarkozy, Maria Cristina Digilio, Bruno Dallapiccola, and Bruno Marino. "22q11 deletion syndrome: a review of some developmental biology aspects of the cardiovascular system". In: *Journal of Cardiovascular Medicine* 7.2 (2006), pp. 77–85 (cit. on p. 5).

[11] Donna M McDonald-McGinn and Kathleen E Sullivan. "Chromosome 22q11. 2 deletion syndrome (DiGeorge syndrome/velocardiofacial syndrome)". In: *Medicine* 90.1 (2011), pp. 1–18 (cit. on p. 6).

[12] Milton Cross. "External links". In: *Focus On: 100 Most Popular RCA Records Artists*

(2017) (cit. on p. 6).

[13] NIMH. *National Institute of Mental Health*. 2018. URL: https://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml (cit. on p. 6).

[14] Anne S Bassett and Eva WC Chow. "Schizophrenia and 22q11. 2 deletion syndrome". In: *Current psychiatry reports* 10.2 (2008), p. 148 (cit. on p. 6).

[15] Anne S Bassett, Eva WC Chow, and Rosanna Weksberg. "Chromosomal abnormalities and schizophrenia". In: *American Journal of Medical Genetics Part A* 97.1 (2000), pp. 45–51 (cit. on p. 6).

[16] Kathleen A Hodgkinson, Jillian Murphy, Sheri O'Neill, Linda Brzustowicz, and Anne S Bassett. "Genetic counselling for schizophrenia in the era of molecular genetics". In: *The Canadian Journal of Psychiatry* 46.2 (2001), pp. 123–130 (cit. on p. 6).

[17] Eva WC Chow, Mark Watson, Donald A Young, and Anne S Bassett. "Neurocognitive profile in 22q11 deletion syndrome and schizophrenia". In: *Schizophrenia research* 87.1 (2006), pp. 270–278 (cit. on p. 6).

[18] Gregory Raux, Emilie Bumsel, Bernadette Hecketsweiler, Therese van Amelsvoort, Janneke Zinkstok, Sylvie Manouvrier-Hanu, Carole Fantini, Georges-Marie M Breviere, Gabriella Di Rosa, Giuseppina Pustorino, et al. "Involvement of hyperprolinemia in cognitive and psychiatric features of the 22q11 deletion syndrome". In: *Human molecular genetics* 16.1 (2006), pp. 83–91 (cit. on p. 6).

[19] Therese van Amelsvoort, Eileen Daly, Jayne Henry, Dene Robertson, Virginia Ng, Michael Owen, Kieran C Murphy, and Declan GM Murphy. "Brain Anatomy in Adults With Velocardiofacial Syndrome With and WithoutSchizophrenia: Preliminary Results of a Structural Magnetic Resonance Imaging Study". In: *Archives of general psychiatry*

61.11 (2004), pp. 1085–1096 (cit. on p. 6).

[20] Eva WC Chow, David J Mikulis, Robert B Zipursky, Laura E Scutt, Rosanna Weksberg, and Anne S Bassett. "Qualitative MRI findings in adults with 22q11 deletion syndrome and schizophrenia". In: *Biological Psychiatry* 46.10 (1999), pp. 1436–1442 (cit. on p. 6).

[21] Opal Ousley, A Nichole Evans, Samuel Fernandez-Carriba, Erica L Smearman, Kimberly Rockers, Michael J Morrier, David W Evans, Karlene Coleman, and Joseph Cubells. "Examining the overlap between autism spectrum disorder and 22q11. 2 deletion syndrome". In: *International journal of molecular sciences* 18.5 (2017), p. 1071 (cit. on p. 7).

[22] Kevin M Antshel, Alka Aneja, Leslie Strunge, Jena Peebles, Wanda P Fremont, Kimberly Stallone, Nuria AbdulSabur, Anne Marie Higgins, Robert J Shprintzen, and Wendy R Kates. "Autistic spectrum disorders in velo-cardio facial syndrome (22q11. 2 deletion)". In: *Journal of autism and developmental disorders* 37.9 (2007), pp. 1776–1786 (cit. on p. 7).

[23] Amelia Kotte, Gagan Joshi, Ronna Fried, Mai Uchida, Andrea Spencer, K Yvonne Woodworth, Tara Kenworthy, Stephen V Faraone, and Joseph Biederman. "Autistic traits in children with and without ADHD". In: *Pediatrics* 132.3 (2013), e612–e622 (cit. on p. 8).

[24] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5)*. American Psychiatric Pub, 2013 (cit. on pp. 8, 11, 12).

[25] Nanda NJ Rommelse, Barbara Franke, Hilde M Geurts, Catharina A Hartman, and Jan K Buitelaar. "Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder". In: *European child & adolescent psychiatry* 19.3 (2010),

pp. 281–295 (cit. on p. 8).

[26] Yael Leitner. "The co-occurrence of autism and attention deficit hyperactivity disorder in children–what do we know?" In: *Frontiers in human neuroscience* 8 (2014), p. 268 (cit. on p. 9).

[27] S Hong Lee, Stephan Ripke, Benjamin M Neale, Stephen V Faraone, Shaun M Purcell, Roy H Perlis, Bryan J Mowry, Anita Thapar, Michael E Goddard, John S Witte, et al. "Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs". In: *Nature genetics* 45.9 (2013), p. 984 (cit. on p. 9).

[28] Jolanda MJ van der Meer, Anoek M Oerlemans, Daphne J van Steijn, Martijn GA Lappenschaar, Leo MJ de Sonneville, Jan K Buitelaar, and Nanda NJ Rommelse. "Are autism spectrum disorder and attention-deficit/hyperactivity disorder different manifestations of one overarching disorder? Cognitive and symptom evidence from a clinical and population-based sample". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 51.11 (2012), pp. 1160–1172 (cit. on p. 9).

[29] Chase C Dougherty, David W Evans, Scott M Myers, Gregory J Moore, and Andrew M Michael. "A comparison of structural brain imaging findings in autism spectrum disorder and attention-deficit hyperactivity disorder". In: *Neuropsychology review* 26.1 (2016), pp. 25–43 (cit. on p. 9).

[30] Nanda Rommelse, Jan K Buitelaar, and Catharina A Hartman. "Structural brain imaging correlates of ASD and ADHD across the lifespan: a hypothesis-generating review on developmental ASD–ADHD subtypes". In: *Journal of Neural Transmission* 124.2 (2017), pp. 259–271 (cit. on p. 9).

[31] Amelia Kotte, Gagan Joshi, Ronna Fried, Mai Uchida, Andrea Spencer, K. Yvonne

Woodworth, Tara Kenworthy, Stephen V. Faraone, and Joseph Biederman. "Autistic Traits in Children With and Without ADHD". In: *Pediatrics* (2013) (cit. on p. 9).

[32] Daniel Bone, Matthew S Goodwin, Matthew P Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. "Applying machine learning to facilitate autism diagnostics: pitfalls and promises". In: *Journal of autism and developmental disorders* 45.5 (2015), pp. 1121–1136 (cit. on p. 10).

[33] Catherine Lord, Michael Rutter, and Ann Le Couteur. "Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders". In: *Journal of autism and developmental disorders* 24.5 (1994), pp. 659–685 (cit. on p. 10).

[34] Michael Rutter, A Le Couteur, C Lord, et al. "Autism diagnostic interview-revised". In: *Los Angeles, CA: Western Psychological Services* 29 (2003), p. 30 (cit. on p. 11).

[35] Judith A Reaven, Susan L Hepburn, and Randal G Ross. "Use of the ADOS and ADI-R in children with psychosis: Importance of clinical judgment". In: *Clinical child psychology and psychiatry* 13.1 (2008), pp. 81–94 (cit. on p. 11).

[36] WR Kates, N Russo, WM Wood, KM Antshel, SV Faraone, and WP Fremont. "Neurocognitive and familial moderators of psychiatric risk in velocardiofacial (22q11. 2 deletion) syndrome: a longitudinal study". In: *Psychological medicine* 45.8 (2015), pp. 1629–1639 (cit. on p. 12).

[37] Rick Ostrander, Kevin P Weinfurt, Paul R Yarnold, and Gerald J August. "Diagnosing attention deficit disorders with the Behavioral Assessment System for Children and the Child Behavior Checklist: Test and construct validity analyses using optimal discriminant classification trees." In: *Journal of Consulting and Clinical Psychology*

66.4 (1998), p. 660 (cit. on p. 12).

[38] Kelly Pizzitola Jarratt, Cynthia A Riccio, and Becky M Siekierski. "Assessment of attention deficit hyperactivity disorder (ADHD) using the BASC and BRIEF". In: *Applied Neuropsychology* 12.2 (2005), pp. 83–93 (cit. on p. 12).

[39] Lauren M Gardner, Jonathan M Campbell, Andrew J Bush, and Laura Murphy. "Comparing Behavioral Profiles for Autism Spectrum Disorders and Intellectual Disabilities Using the BASC-2 Parent Rating Scales–Preschool Form". In: *Journal of Psychoeducational Assessment* (2017), p. 0734282916689438 (cit. on p. 12).

[40] Sara S Sparrow, David A Balla, Domenic V Cicchetti, Patti L Harrison, and Edgar A Doll. "Vineland adaptive behavior scales". In: (1984) (cit. on p. 12).

[41] Nancy J Roizen, Thomas A Blondis, Mark Irwin, and Mark Stein. "Adaptive functioning in children with attention-deficit hyperactivity disorder". In: *Archives of pediatrics & adolescent medicine* 148.11 (1994), pp. 1137–1142 (cit. on p. 13).

[42] Sabrina Yang, Jessica M Paynter, and Linda Gilmore. "Vineland adaptive behavior scales: II profile of young children with autism spectrum disorder". In: *Journal of autism and developmental disorders* 46.1 (2016), pp. 64–73 (cit. on p. 13).

[43] Darryn M Sikora, Parul Vora, Daniel L Coury, and Daniel Rosenberg. "Attention-deficit/hyperactivity disorder symptoms, adaptive functioning, and quality of life in children with autism spectrum disorder". In: *Pediatrics* 130.Supplement 2 (2012), S91–S97 (cit. on p. 13).

[44] Kevin M Antshel, Wendy R Kates, Nancy Roizen, Wanda Fremont, and Robert J Shprintzen. "22q11. 2 deletion syndrome: genetics, neuroanatomy and cognitive/behavioral features keywords". In: *Child Neuropsychology* 11.1 (2005), pp. 5–19 (cit. on p. 13).

[45] Arthur L Samuel. "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229 (cit. on p. 14).

[46] Ron Kohavi. "Glossary of terms". In: *Machine Learning* 30 (1998), pp. 271–274 (cit. on p. 14).

[47] Gary W Lewandowski Jr and David B Strohmetz. "Actions can speak as loud as words: Measuring behavior in psychological science". In: *Social and Personality Psychology Compass* 3.6 (2009), pp. 992–1002 (cit. on p. 14).

[48] Ian J Deary, Geoff Der, and Graeme Ford. "Reaction times and intelligence differences: A population-based cohort study". In: *Intelligence* 29.5 (2001), pp. 389–399 (cit. on p. 14).

[49] Mahiye Uluyagmur-Ozturk, Ayse Rodopman Arman, Seval Sultan Yilmaz, Onur Tugce Poyraz Findik, Herdem Aslan Genc, Gresa Carkaxhiu-Bulut, M Yanki Yazgan, Umut Teker, and Zehra Cataltepe. "ADHD and ASD Classification Based on Emotion Recognition Data". In: *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE. 2016, pp. 810–813 (cit. on p. 15).

[50] Brett Baisch, S Sean Cai, Zongming Li, and Victor Pinheiro. "Reaction time of children with and without autistic spectrum disorders". In: *Open J Med Psychol* 6 (2017), pp. 166–178 (cit. on p. 15).

[51] P Rinck. "Magnetic resonance: a critical peer-reviewed introduction". In: *Magnetic resonance in medicine. The basic textbook of the European magnetic resonance forum,* 2014, pp. 21–01 (cit. on p. 15).

[52] Sina Ghiassian, Russell Greiner, Ping Jin, and Matthew RG Brown. "Using functional or structural magnetic resonance images and personal characteristic data to identify

ADHD and autism". In: *PloS one* 11.12 (2016), e0166934 (cit. on p. 16).

[53] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. "Identification of autism spectrum disorder using deep learning and the ABIDE dataset". In: *NeuroImage: Clinical* 17 (2018), pp. 16–23 (cit. on p. 16).

[54] Daniel S Tylee, Zora Kikinis, Thomas P Quinn, Kevin M Antshel, Wanda Fremont, Muhammad A Tahir, Anni Zhu, Xue Gong, Stephen J Glatt, Ioana L Coman, et al. "Machine-learning classification of 22q11. 2 deletion syndrome: A diffusion tensor imaging study". In: *NeuroImage: Clinical* 15 (2017), pp. 832–842 (cit. on p. 16).

[55] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. Springer, 2013 (cit. on p. 17).

[56] M Duda, R Ma, N Haber, and DP Wall. "Use of machine learning for behavioral distinction of autism and ADHD". In: *Translational psychiatry* 6.2 (2017), e732 (cit. on p. 17).

Siri Chandana Sambatur

Curriculum Vitae

PERSONAL DETAILS

Birth	August 10, 1994
Address	444 Westcott Street, New York
Phone	(315) 416-2969
Mail	sambatur.siri@yahoo.com

EDUCATION

MSc. Computer Science 2016-2018

Syracuse University

GPA- 4.67

Relevant Coursework-Analytical Data Mining, Social Media Mining, Object Oriented Design, Design and Analysis of Algorithms, Natural Language Processing

BTech Computer Science and Engineering 2012-2016

VNR Vignana Jyothi Institute of Engineering and Technology

Graduated top 5%

MASTER THESIS

Computational Analysis of Developmental Disorders in Children 2018

Supervisor- Dr. Reza Zafarani

Analysis of various Developmental disorders and their comorbidity using tree-based learning techniques and feature selection on parent-oriented reviews

WORK EXPERIENCE

Teaching Assistant 2018- present

Syracuse University, Part-time

- Tutor and assist students to master their python skills through assignments
- Conducting lab sessions every week for students to learn python practically
- Evaluated student performance, provide feedback and assign grades for assignments

Web Analyst 2017- present

Institute of Veterans and Military Families, Part-time

- Deploying customized websites with WordPress using 17 different themes and templates
- Analyzing 90% of the traffic in the websites based on clicks and navigations using event tracking
- Examining the statistics of websites such as page views and pdf downloads using Google Analytics

Analytical Intern 2017

Ernst & Young LLP, Full-time

- Redesigned dashboards with KPIs for business organizations to understand their objectives and targets
- Diagnosed the customer transactions database to generate automated daily reports with 100% accuracy
- Modeled structured data to set goals for 2017 fiscal year by analyzing the 2016 fiscal year results

PROJECTS

Training a Smart cab to drive 2017

github. com/SirichandanaSambatur/SmartCab

- Engineered a smart cab using Q- Learning techniques to follow traffic rules and reach destination
- Measured Safety Rating (A+) and Reliability Rating (A+) after the learning process

Emotional Influence in Social Networks 2017

github. com/SirichandanaSambatur/EISN

- Measured sentiments in Quora for the topic- "Kashmir Conflict" for 2616 questions and 10512 answers
- Performed k-means clustering on 11200 users and 6/10 most viewed writers had negative emotions

Remote Code Publisher 2017

github. com/SirichandanaSambatur/RCP

- Implemented a NoSQL database and persisted the contents of the database with XML
- Published webpages from multiple clients interacting with GUI developed using Windows Presentation Foundation January - May 2017 January - March

SKILLS

Programming Languages	C, C++, Java, Python, R, MATLAB
Web programming	HTML5, CSS3, Angular JS, JQuery, Ajax
Operating Systems	Linux, Ubuntu, Windows, iOS, Unix
Databases	Oracle, Apache Tomcat, MySQL
Other	UML, Data Structures, Tensor Flow, Scikit-learn, NLTK, Git

LEADERSHIP EXPERIENCE

General Secretary, Computer Society Of India 2014-2016

Student Branch of VNRVJIET

- Guided a group of 35 - 40 students to organize 15 technical events and career workshops
- Designed 3 editions of the annual technical magazine CZINE launched every year