Diagnosing Autism Spectrum Disorder with Machine Learning

By

DEVASHISH KASHIKAR

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Chen-Nee Chuah, Chair

Sen-ching Samson Cheung

Lifeng Lai

Committee in Charge

2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Diagnosing Autism Spectrum Disorder with Machine Learning

Autism Spectrum Disorder (ASD) affects 1 in 59 children in America and the current methods of diagnosis are cumbersome, time intensive, and also require substantial resources which create a delay in achieving a diagnosis. The lost time is crucial for behavioral interventions and treatments that have proven to reduce the core symptoms of ASD. Techniques to effectively and promptly assess risk of ASD and other developmental disorders are highly necessary to reduce the delay. Using forward feature selection, oversampling, undersampling, and 5-fold cross-validation, we trained and tested five machine learning models on twelve features extracted from videos recorded at the MIND Institute. We found that out of the twelve features, only seven were necessary to achieve a diagnosis and the top performing models had sensitivity ranging from 81% to 88%. Our ability to achieve higher performance was limited by the small amount of ASD samples in the dataset, but the results show that machine learning algorithms have the ability to capture relevant information about the condition. Machine learning approaches can assist in streamlining the clinical diagnostic process of ASD as well as reach a larger community that may not have access to required resources.

ACKNOWLEDGMENTS

I'm extremely grateful to my parents for their guidance and motivation. I would not have made it all the way through without them by my side.

I would also like to thank the professors on the committee for their mentorship and support on this thesis.

Chapter 1

Introduction

ASD is a neurodevelopmental disorder which causes difficulty in maintaining social interaction and communication as well as displaying repetitive patterns. In 2017, CDC recorded 1 in 68 children with ASD and in 2018, 1 in 59 children were diagnosed with ASD [1]. Factors such as availability of trained clinicians and diagnostic tools are causing a rise in the delay of receiving a diagnosis and contributing to the growing prevalence. The average age of diagnosis in the United States is after 4 years old, but there are studies showing that ASD can be diagnosed at 2 years of age [2]. Even though the 2-year delay may seem trivial, this time period is extremely critical for children because studies show that this might be a 'pre-symptomatic' period where interventions can have a more powerful effect on gene expression and symptom emergence [3].

Due to the complexity of the disorder and the overlap of symptoms from other developmental disorders, there still has not been an agreement as to how a cure should be defined for ASD [4]. Therefore, the current option that remains is for behavioral therapies to take place after diagnosis to reduce the effects of the core symptoms of ASD [5]. Diagnosing ASD itself remains an intricate process generally requiring a trained clinician to examine the behaviors of the child. This process can be expensive and time consuming, making it inaccessible for people in rural areas. [6].

As an alternative to the current diagnostic methods, approaches in machine learning and computer vision are being explored so that an early diagnosis can be achieved with less cost and increased availability to people in rural and semi urban areas [7], [8]. This thesis illus-

trates the machine learning approaches taken towards diagnosing ASD in 18 to 36-month-old children as well as the effect that age has on the ability to achieve a diagnosis. The diagnosis of ASD is treated as a binary classification task and various machine learning models are trained with features extracted from video data provided by the MIND Institute. Some of the techniques explored differ from current machine learning approaches because the features used for model training are extracted from videos and they also explore the effects of achieving a diagnosis based on the age of the patients. These effects are analyzed in two different sets of experiments. The first set separates the video data based on the age the video is recorded and the second set separates the data by the age the patient is diagnosed. The results, on average, have shown that older children exhibit more recognizable ASD characteristics than younger children and when Random Forest models are training with age-based data and tested on data from 18-month-olds, they are able to classify ASD samples with an average AUROC of 66%.

1.1 Motivation for Early Intervention

Children on the spectrum usually start to exhibit reduced gestures and initiations in social communicative environments by the time they are two years old [3]. It's crucial to identify these symptoms as they develop so that they can be appointed to interventions that can reduce the growing developments of core ASD symptoms and steer development towards a more typically developing pathway [3]. An example of such interventions is the Early Start Denver Model (ESDM). Its behavioral therapy that creates a positive environment between children, parents and therapists. Through the joint activities that take place, children are encouraged to enhance cognitive, social and language skills [1]. Children who received ESDM for two years showed a significant increase in IQ, social, language and adaptive behavior [9]. The brain is at a malleable phase during early development of ASD and interventions can alter the course of this development. If intervention is initiated before the onset of core ASD symptoms, it's highly possible to prevent these effects from developing further and affecting other functions [1]. In order for children to start intervention, they have to be diagnosed as early as possible, around 18 or 24 months. Due to the extensive resources required for

screening methods, its beneficial to use parts of diagnostic modules to train machine learning models [10].

1.2 State-of-the-Art Diagnostic Screening Methods

Even though early interventions have shown to improve social-brain circuitry, achieving early diagnosis can still be a challenge. Isolating ASD symptoms and recognizing them as ASD is difficult because there is a high degree of overlap between symptoms of other developmental disorders. ASD has approximately a 30% overlap with attention-deficit-hyperactivity disorder (ADHD), 20% overlap with epilepsy and 10% overlap with an identifiable single gene disorder. This overlap creates an even bigger challenge to diagnose ASD and there has been no evidence showing effective medicine treatment options available to cure the symptoms of ASD. Due to the genetic complexity of ASD combined with its similarities among other developmental disorders, diagnostic screening methods have been proven to be the most effective in distinguishing ASD patients from non-ASD patients. [4]

Three of the most common screening methods used are the Autism Diagnostic Observation Schedule (ADOS), Autism Diagnostic Interview – Revised (ADI-R), and Childhood Autism Rating Scale (CARS). ADOS is a screening method that consists of four modules. One module is chosen to evaluate the behaviors of the child depending on his/her age and verbal skills. Each module consists of structured and unstructured activities and interactions which provide standardized contexts to communicative, social, and other behaviors relevant to development disorders. The observation period lasts for about 30 to 45 minutes and in the end, each categorized observation is combined to produce a quantitative score. Using this score as a method of assessment, ADOS is able to extract subtle behaviors indicating ASD behavior and understand the social nuances of the people being evaluated. [11]

Another common screening method is the ADI-R which is a meeting between the parents of the patient and an examiner that extracts historical and current information about the behavior of the patient. This test covers 93 items that focus on three primary areas: communication and language (e.g. stereotyped utterances, usage of language in a social setting), repetitive, restricted and behaviors (e.g. unusual sensory interests and hand and finger man-

nerisms), and the quality of social interactions (e.g. social smiling and emotional sharing). A score, on a 4-point scale (0 – 3), is given for each behavior with the lowest score representing typical developing behavior and the highest score representing ASD behavior. Creating categories for behaviors allows ADI-R to break down ASD behavior into levels so it's easier for the examiner to group behaviors in certain groups and analyze scores for each behavior separately. [12]

CARS is another method, similar to ADOS and ADI-R, that is utilized because it measures a large range of behaviors that indicate autistic behavior. It assesses 14 major behaviors and each behavior is given a score between 1 – 4. Similar to ADI-R, the lowest score represents typical developing behavior and the highest score illustrates ASD behavior that may affect functioning of other area. Each individual score is summed to form a total score that is used to convey whether the behaviors of the patient indicate ASD or not. A total score below 30 indicates the patient is in the non-autistic range and scores from 37 to 60 indicate sever autism. [13].

1.2.1 Drawbacks

The screening methods described in the previous section are able to distinguish an ASD patient from a non-ASD patient, but these techniques have limitations which prevent them from being used globally. The biggest factor contributing to the delay in ASD diagnosis is the need for professional clinicians to conduct these tests [14]. Clinicians seem to be available in developed countries, but in places like Nepal, there is a dearth of resources which causes the average ASD diagnosis to occur around 58 months [15]. The late diagnosis creates a higher risk of no recovery for the patient and an imbalance in the coverage of the population around the world [10, 14].

Another contributing factor to the delay in diagnosis is the time that it takes to perform these screening tests [15]. The ADI-R method discussed above consists of 93 items and takes up to 2.5 hours long. Research shows that only 7 of the 93 questions are required to train machine learning models to achieve nearly 100% statistical accuracy [14]. Classification algorithms used in these studies have been able to eliminate the time and cost that current diagnostic methods require. Similar to the features used from the ADI-R method, [10] used

8 of the 29 items in the Module 1 of the ADOS. Their findings are impressive with models achieving 100% sensitivity and 100% accuracy. These studies show that by creating features from commonly used diagnostic tests, machine learning models can be implemented at any clinic diagnosing ASD.

1.3 Our Contributions

We developed supervised machine learning models to diagnose children with ASD. Achieving an early diagnosis is extremely important for the child to start intervention. Currently, the screening techniques being used for diagnosis involve clinicians, but we explored approaches that eliminated the need for professionals to be involved during the diagnostic process. Many of the projects discussed in Chapter two require training models on data from screening methods such as ADOS and ADI-R. Even though these projects have explored the advantages of machine learning algorithms, the features used as input still required clinician support. The experiments discussed in this thesis involve using extracted features from videos as input to machine learning algorithms. Our methods have shown that it is possible to achieve an ASD diagnosis by automating the diagnostic screening process.

We first investigated the performance of five different classification models, discussed in Chapter 3.3, with relevant features extracted from videos. After conducting a variety of different experiments to determine which model maximizes output, we study the performance of random forest when its trained with data separated by age. Along with using features extracted from videos, our approach differs from current work because we also explore the effects that age has on achieving a diagnosis. Chapter 4.2.3 discusses experiments that involve data from different ages to be used for model training. From some of these experiments, we were able to show that since older children have developed symptoms, the machine learning models were able to learn information from them much more effectively.

Chapter 2

Related Work

Since obtaining an early diagnosis is essential for receiving early intervention, there is a pressing need for diagnostic tools that are open-source and do not require professional training [8], [7]. Machine learning algorithms meet the criteria and have the potential to reach semi-urban and rural population that lack resources for diagnosing ASD. Classification models have used observation-based features like eye contact, imaginative play, and reciprocal communication to detect ASD [16]. Some of the models have achieved an accuracy of 97% and have cut down the time it takes to receive a diagnosis by 72% [16]. This highlights the capability and the power that machine learning algorithms can have on diagnosing ASD. It is important to note that the input features for these models are sought out carefully with expert guidance. It is highly important for the features to represent the characteristics of ASD otherwise the model will not learn useful information.

Machine learning algorithms help alleviate difficulties that arise in formulating an ASD diagnosis, but they can also be used to differentiate between developmental disorders such as ASD and ADHD [17]. Duda et al. uses the data of 2925 subjects from the Social Responsiveness Scale (SRS) to train six machine learning models to distinguish between ASD and ADHD [17]. The SRS is a parent-completed screening questionnaire that is used to assess children suspected to have ASD or other mental disorders [18]. The study adopted 65 features from the SRS and implemented forward feature selection to reduce the dimensionality of the data to less than ten features. During training and testing classification models, the data maintained a ratio of 1.5:1 (ASD to ADHD). Support Vector Machine and logistic regres-

sion achieved an AUROC of 0.96 while decision trees and random forest could only reach an AUROC of 0.93 and 0.95. Even though the area is almost close to 100% for these models, these models still may not be very accurate and precise. The models were evaluated on the same data as they were trained on, so there is not much evidence on how well these models will be able to generalize to different datasets.

In an eye gaze study, Madipakkam, et al. proves the reported lack of eye contact in ASD patients by showing the TD group had an unconscious bias towards faces with direct gaze and the ASD group focused towards faces with averted gaze [19]. Clear differences in eye contact between TD and ASD has sparked interest in effectively measuring and distinguishing gaze into two distinct groups. Chong et al. presents a novel approach to eye contact detection during adult-child interaction by proposing the adult wearing a POV camera to capture an egocentric view of the child's behavior [20]. Head pose is a key factor in interpreting gaze direction based on analysis of the eye regions and this study focuses on estimating head pose by introducing Pose-Dependent Egocentric Eye Contact (PEEC) and Pose-implicit Convolutional Neural Networks (PiCNN) Detections. The PEEC uses an egocentric video as input and predicts the direction of eye contact. This method relies on head pose estimation and facial landmark detection to determine the direction of eye contact. Dependence in detecting other events creates a bottleneck, making the entire PEEC pipeline unreliable. PiCNN is also proposed in this paper and it takes a rectangle of pixels corresponding to a face bounding box as the input and outputs a prediction of three axes of head rotation (yaw, pitch, roll) as well as a binary classification of eye contact. This method achieved precision of 0.75 and recall of 0.78, higher than AlexNet, in classifying whether or not the ASD group maintained eye contact.

Along with a lack in the ability to maintain eye contact, several studies have shown that children with ASD are perceived as less engaging by observers. Guha et al. focuses on quantifying details of facial expressions of children with high functioning autism (HFA). This study was able to show that there is reduced complexity in the facial behavior of the HFA group arising from the eye region. Complexity in facial dynamics is defined as the rate at which new information is produced and it is investigated with the multiple scale entropy method.

Facial behavior was studied by placing 32 markers on the face of the person being evaluated and the Facial Action Coding System (FACS) is implemented to study local facial movements. The study shows that HFA group produces the highest complexity in the cheek region and the TD group in the eye region. This study believes that since children with ASD avoid looking at the eye region of the face, they are able to produce fewer intricate movements in the eye region because they lack experience in perceiving and processing this dynamic information. There is evident difference in the facial complexion of TD and ASD children. Using this as a feature in machine learning will be able to further capture the characteristics of ASD and help the performance of the models. [21]

Pratap et al. compares performances of various machine learning models, including Naïve Bayes, K Means clustering, and Artificial Neural Network (ANN), on data from the CARS diagnostic tool. The goal of this study was to use a 16-feature dataset consisting of 50 cases of children with ASD and 50 cases of TD children. There were four target classes that the model tried to predict: Normal, Mild-Moderate, Moderate-Severe, and Severe. The results of the experiments indicated that when probabilistic models or ANN models were integrated with unsupervised learning methods such as K-Means clustering, the results of detecting the presence of ASD improved. The results of this study only report the accuracy of the models which creates doubt about the true performance of the algorithms discussed in these models. The dataset used in the experiment represents a really small sample space for autism and may not have been able to capture the true characteristics of the group. [22]

To demonstrate the effectiveness of machine learning models in diagnosing ASD, Tariq et al. uses eight models, including Support Vector Machine, Logistic Regression, and Decision Trees, to identify ASD by utilizing 30 behavioral features from ADOS and ADI-R. The data analyzed includes 116 home videos of children with ASD and 46 videos of TD children. Additional 66 videos, 33 ASD and 33 TD, were used to evaluate the performance of models on data never seen before. Out of all the models implemented, Logistic Regression performed the highest with eight features and achieved 0.93 AUROC. This study concludes that feature tagging home videos for machine learning classification of ASD yields optimistic results. It's difficult to be confident with model performances because these models were tested on sim-

ilar data as they were trained on. [7]

In our machine learning approach, we use annotated video data from interactive sessions conducted at the UC Davis MIND Institute. This Institute is a research center dedicated to understanding, preventing, and treating challenges associated with neurodevelopmental disabilities. It conducts research on many disorders including the Autism Spectrum Disorder (ASD). Parents bring children to be evaluated by interactive sessions if they find unpredictable behavior in their children [17]. The dataset consists of video data from 25 ASD and 154 TD children. Majority of children have videos of sessions from when they were 6 to 36 months old, but only data from 18 to 36 months was used because earlier ages do not exhibit clear properties of ASD.

Chapter 3

Methodology

3.1 Data

The dataset used in this project was collected from children between the ages of 6 months to 36 months that participated in a sibling study. The children that are a part of this study are siblings of children who have a formal diagnosis of ASD. The results from the study show that there is an increased risk of developing ASD if their older sibling has already been diagnosed with ASD. It is an interactive study that consists of an examiner observing the behavior of children in a one-on-one session. It takes place in a room with a privacy glass window and has a team on the opposite side of the window that records the session. During the session, an examiner, with a box of toys next to him/her, sits in front of a child. If the children are too young, they are accompanied by their parents. The examiner grabs the attention of the child with a toy and assesses the behavioral response exhibited towards him/her and the toy. In a similar manner, all children participating in the study have a video recorded of them while interacting with an examiner. Once all videos have been recorded, a team of coders mark the start and end times of the events listed in Table 3.1. These events are the main components of the study because they are deemed important by researchers in behavior recognition for ASD. Examples of these events are provided in Table 3.2.

Almost all the children in the dataset visited the MIND Institute when they were 18, 24, and 36 months old. Tables 3.3 and 3.4 show the breakdown of the dataset highlighting the imbalance of data between the two classes. Data from 25 unique ASD and 142 unique Non-

ASD patients was analyzed to create a feature set for classification models. In each video, the mean duration and the frequency of each event, as described in Table 3.1, are used as the input to five machine learning models. The frequency counts the occurrences of the event and the mean duration is the ratio of the total duration of the event in the video to the frequency of the event. For all the experiments conducted on this dataset, each sample used for training and testing is event data extracted from a single video. Since distinguishing between ASD behavior or a general delay in development can be a challenge until children are at least 2 years old [3], only data from children 24 and 36 months old is used for model training.

Event Name	**Description**
Look Face	Child maintains eye contact with the examiner
Look Object	Child maintains eye contact with a toy
Vocalization	Child is vocally active
Smile	Child is smiling
Shared Smile	Child is smiling as he or she looks at the examiner
Shared Vocalization	Child is vocally active as he/she looks at the examiner

Table 3.1. Description of Events Extracted From Videos

3.2 Distinctions Between ASD and Non-ASD Feature Distributions

Figures 3.1 and 3.2 display the distributions of the features. The y-axis shows the range of values for the features and '0' on the x-axis represents the distribution for the Non-ASD class and '1' represents the ASD class. To quantify the differences in mean and variance between events of both classes, an analysis of variance test (ANOVA) is conducted [23]. This test computes a f-value that measures the ratio of variance between the two groups to the variance within the groups and a p-value that determines whether or not the null hypothesis will be

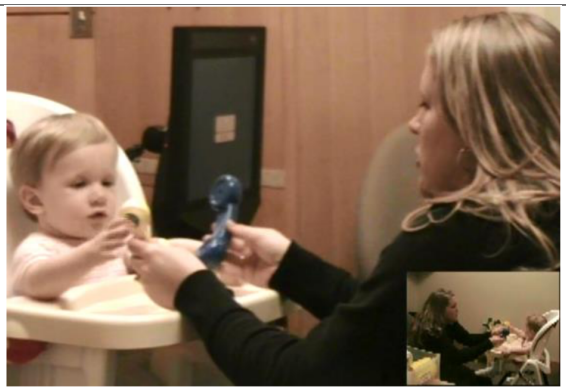
Event	Example Image
Look Face	
Look Object	
Smile	

Table 3.2. Examples of Look Face, Look Object, and Smile Events

rejected or accepted. The null hypothesis states that the two independent samples used as input have identical expected values and the populations have identical variances. If the p-value is less than a certain threshold, typically 0.05, ANOVA provides strong evidence towards rejecting the null hypothesis and if the p-value is greater than this threshold, then the test fails to provide evidence against the null hypothesis and it is accepted. Table 3.5 shows the f and p-values of features extracted from 24 and 36-month-old data. Except for the vo-

Class	Age (in Months)	Number of Patients	Number of Videos
ASD	**18**	21	57
	24	19	52
	36	22	61
Non-ASD (CV set)	**18**	105	285
	24	94	259
	36	89	245
Non-ASD (Test set)	**18**	35	90
	24	28	77
	36	30	77

Table 3.3. Breakdown of the Dataset Based on Availability of Video Files

Class	Age (in Months)	Number of Patients	Number of Videos
ASD	**18**	11	82
	24	7	25
	36	7	14
Non-ASD (CV set)	**18**	105	285
	24	94	259
	36	89	245
Non-ASD (Test set)	**18**	35	90
	24	28	77
	36	30	77

Table 3.4. Breakdown of the Dataset Based on the Age of Diagnosis of Patients

cal duration, all other features have p-values that are significantly lower than the threshold of 0.05 and relatively high f-values. The low p-values illustrate that these particular features are statistically significant in the ASD and the Non-ASD classes and the high f-values imply that the means of the groups greatly differ from each other compared to the variation of the individual observations in each group [24]. Similarly, Table 3.6 shows the f and p-values of features used for children that are less than 24 months old. The p-values in Table 3.5 are

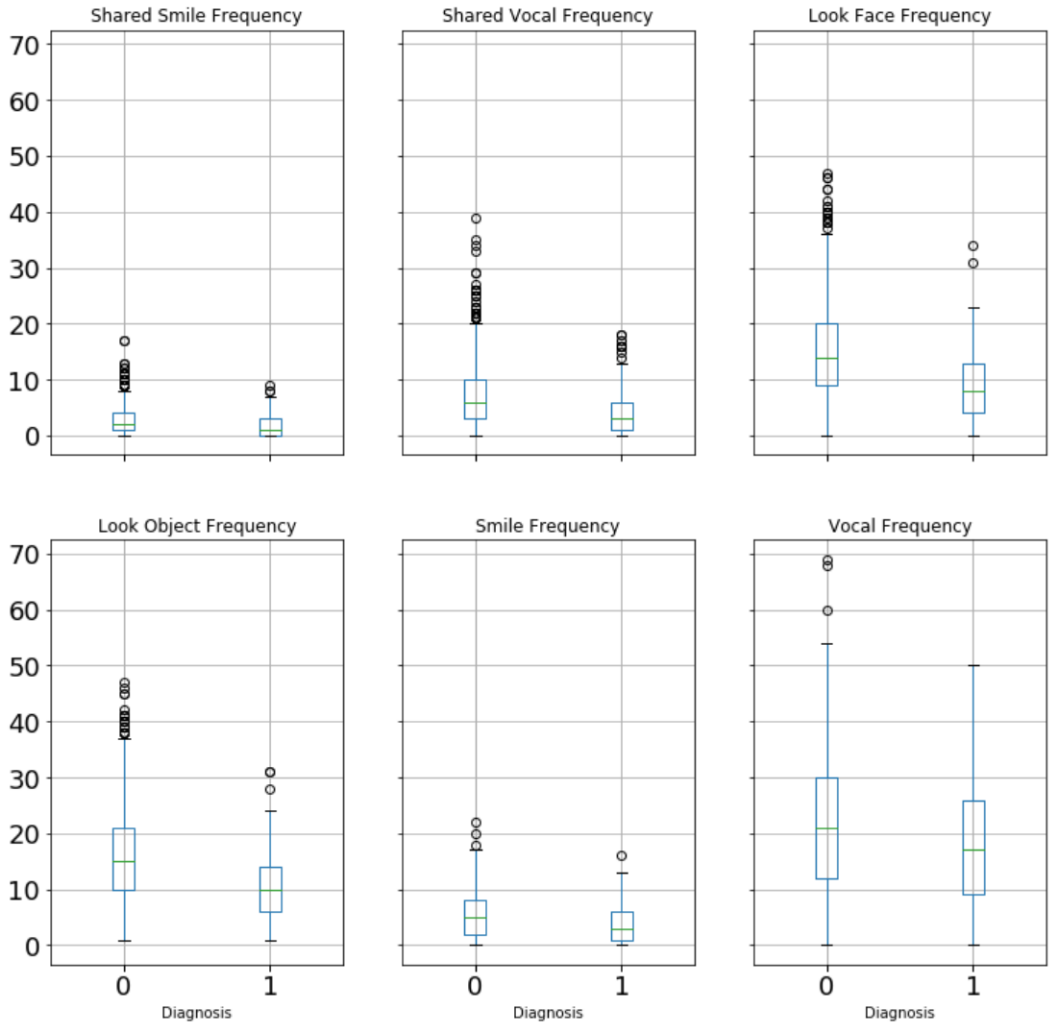


Figure 3.1. Distribution of Frequency Events

much higher than the p-values in Table 3.6 because typically, children with ASD start developing symptoms around 24 months [3] and the high p-values show that before turning 24 months old, ASD and Non-ASD children interact in a similar manner. The f-values are lower than the ones from Table 5.5 illustrating that the variance of events for ASD and Non-ASD children is similar for earlier ages.

Even though the features used to represent the classes show significant difference in ASD and Non-ASD children, the small sample size of ASD is unable to capture the true characteristics of the general ASD population due to the law of large numbers being true for small numbers [25]. There is an inherent bias present in this dataset, especially for the ASD class,

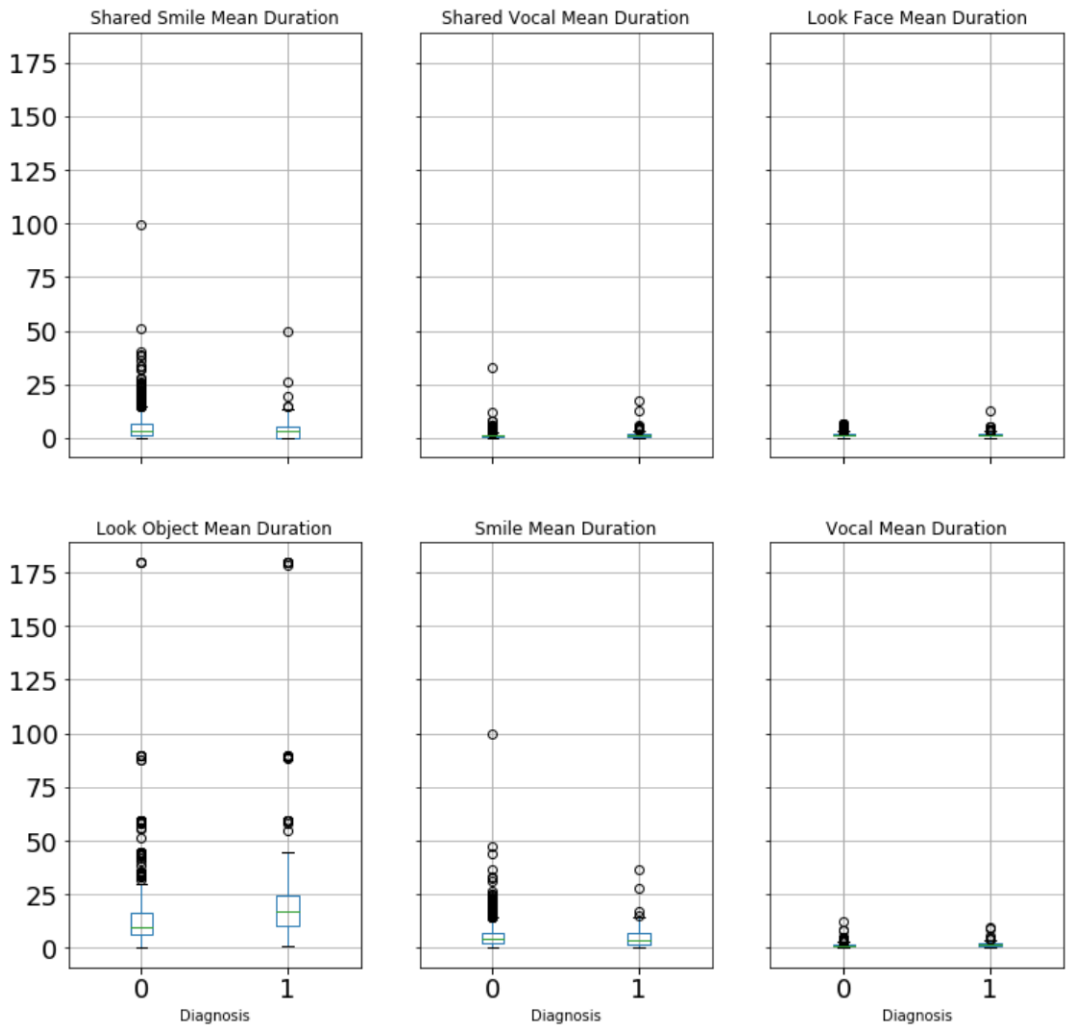


Figure 3.2. Distribution of Duration Events

because ASD features are extracted from only 25 subjects which is an extremely small sample space to represent Autism. The annotation techniques for the videos created by the MIND Institute adds a potential observer bias because the events in the videos are hand-labeled. In order to mitigate the effects of high entropy in the performance of models, various approaches are experimented with. These experiments include training and evaluating Random Forest, Balanced Random Forest, Weighted Random Forest, Extreme Gradient Boosting, and Adaptive Boosting, selecting a subset of optimal features, removing the imbalance by keeping the sample sizes of both classes equal and finally exploring the effects of age on achieving a diagnosis.

Feature	p-value	f-value
Look Face Duration	2.44×10^{-7}	27.13
Look Object Duration	5.90×10^{-4}	11.92
Smile Duration	2.40×10^{-4}	13.60
Vocal Duration	0.97	0.0015
Shared Smile Duration	5.20×10^{-5}	16.55
Shared Vocal Duration	0.0015	10.20
Look Face Frequency	7.09×10^{-13}	53.32
Look Object Frequency	5.95×10^{-11}	44.08
Smile Frequency	2.95×10^{-5}	17.66
Vocal Frequency	3.70×10^{-4}	12.81
Shared Smile Frequency	4.0×10^{-6}	21.56
Shared Vocal Frequency	2.0×10^{-6}	22.89

Table 3.5. p-value and f-value for Children 24 and 36 Months Old

3.3 Machine Learning Models

3.3.1 Random Forest (RF)

In an attempt to eradicate the impact of overfitting due to the class imbalance, models that could subside the effects of high variance were needed and decision tree-based models such as random forest was a fitting choice. Random forest is an ensemble method that constructs a myriad of decision trees during training and averages the prediction of each tree for its final prediction. Each decision tree votes for a class and these votes are averaged to get one final vote in the end. Before training the model, the number of decision trees that will be created needs to be defined. All decision trees start with a single node and keep splitting until the maximum depth of each tree is reached. Each node in random forest is split by the value

Feature	p-value	f-value
Look Face Duration	0.12	2.39
Look Object Duration	0.12	2.42
Smile Duration	0.30	1.06
Vocal Duration	0.34	0.93
Shared Smile Duration	0.40	0.71
Shared Vocal Duration	0.16	1.93
Look Face Frequency	1.34×10^{-6}	23.55
Look Object Frequency	9.13×10^{-7}	24.29
Smile Frequency	0.021	5.32
Vocal Frequency	0.012	6.39
Shared Smile Frequency	0.011	6.47
Shared Vocal Frequency	0.0042	8.21

Table 3.6. p-value and f-value for Children Less Than 24 Months Old

of the Gini impurity, which is also a modifiable parameter in the model. At each node, the impurity in the dataset is calculated by estimating the probability of incorrectly classifying a randomly chosen sample as if it were to be labeled according to the class distribution in the dataset. Due to the random sampling of the dataset and the creation of several decision trees, the random forest model performs relatively well on the given dataset.

3.3.2 Balanced Random Forest (BRF)

Training random forest on a highly imbalanced dataset can lead to decision trees having few or none of the minority class samples resulting in poor classification performance. A naïve way of fixing this problem is to control the number of samples from the majority class in each decision tree. For each iteration in random forest, first a bootstrap sample from the

minority class is selected and then the same number of samples from the majority class are randomly drawn. Adding this constraint to the splitting process transforms a random forest into balanced Random Forest. [26]

3.3.3 Weighted Random Forest (WRF)

Another approach to make random forest well suited for imbalanced data follows the idea of cost sensitive learning. RF is generally biased towards the majority class so it's important to place a heavier penalty on misclassifying the minority class. Weights are assigned to each class, with the higher weight placed on the minority class, and are incorporated in two places, terminal nodes of each tree and during tree induction. As each sub-decision tree is created, the class weights are taken into account to weight the Gini impurity for finding the best splits. The class predictions for each tree are also weight by taking the product of the weight for the class and the number of cases for that class at the terminal node. The final class prediction is formulated by aggregating the weighted vote from each tree.[26]

3.3.4 Extreme Gradient Boosting (XGB)

Similar to random forest, XGB is also an ensemble method that aggregates predictions from weak classifiers to create a strong classifier. By combining predictions from independent models, the errors of the previous model are corrected by the next model. Instead of assigning different weights to classifiers after every iteration, XGB fits the new model to residuals from the previous prediction and minimizes the loss when as it adds another prediction. The model updates itself using gradient descent which helps alleviate the problem of overfitting by penalizing or even removing nodes that do not contribute in the classification task. [27]

3.3.5 Adaptive Boosting (AdaBoost)

AdaBoost is another boosting algorithm created to reduce error rate by forcing the algorithm to focus on labels from the minority class. This boosting algorithm is similar to XGBoost because it also creates multiple weak classifiers and averages their results to produce the final prediction. However, the method used for updating weights is slightly different. Initially, all samples are assigned to equal weights of 1 divided by the total number of training samples. During training, each weak classifier starts to classify samples as positive or negative. Once

all learners have made a prediction, the total error is calculated by adding all the weights of samples that have been misclassified divided by the weight of all the samples. In the next iteration, this error is used to update the weights of each weak classifier by assigning higher weights to classifiers that misclassified the samples and lower weights to classifiers that predicted the class correctly. This method ensures that incorrect predictions are given more importance because they can lower the overall performance of the model. [28]

3.4 Class Imbalance

3.4.1 Oversampling and Undersampling

There is a significant imbalance between the ASD and Non-ASD samples in the dataset. For 24 and 36-month-old data, there are 24 unique ASD subjects that have a total of 113 videos and there are 142 unique Non-ASD subjects with a total of 658 videos. A detailed breakdown of the data is shown in Tables 3.3 and 3.4. Almost all subjects in the dataset have videos taken of them when they are 18, 24, and 36 months old. Initially, Synthetic Minority Oversampling Technique (SMOTE) and Tomek links are used to alleviate the problem of overfitting, but due to high bias in the ASD data, these methods were ineffective. To resolve this imbalance, the Non-ASD sample size was downsampled to match the ASD sample size which is discussed in the equal sampling section of results.

SMOTE is an oversampling technique that generates more data points for the minority class [29]. This technique first takes a point from the data and examines its k nearest neighbors. A vector between the data point and one of its k neighbors is drawn and the distance between these two points is multiplied by a random number between 0 and 1. This number is then added to the current data point, creating a new synthetic sample in the dataset. Tomek links is an under-sampling method used to remove data points from the majority class. A Tomek link is formed when two samples from different classes are nearest neighbors of each other [30]. Once these links are established, the majority class sample is removed, leaving only the minority class sample. Due to the non-linearity of the ASD class and the similarity of distribution with the Non-ASD class, these methods were not as effective in improving model performance.

Chapter 4

Results & Discussions

The results are divided into two sections, training the models with an imbalanced dataset and a balanced dataset. Accuracy, precision, sensitivity, specificity, and area under receiver operating curve (AUROC) metrics are used to evaluate the performance of the models. Accuracy is the ratio of correct predictions to the total predictions made and precision is the ratio of true positives to the total number of predicted positives, assessing the quality of the positive predictions made. Sensitivity, also known as recall, is the ratio of true positives to true positives plus false negatives and it analyzes the ability of the model to identify all the positive samples. Specificity, on the other hand, is the true negative rate of the model that determines the ability of the model to identify all the negative samples and the AUROC calculates the product of sensitivity and 1-specificty.

4.1 Unequal Class Sampling

4.1.1 Comparison Between Five Models

The two main goals of this experiment were to find the best fitting model and analyze the effects that undersampling and oversampling have on class imbalance. On average, undersampling the dataset with Tomek links resulted in better performance than oversampling because SMOTE was only able to create synthetic point by linear interpolation. Removing Non-ASD samples that had overlapping values with ASD samples provided models with clear boundaries distinguishing both classes.

Performances of the five machine learning models are shown in Table 4.1. Before train-

Classifier	Accuracy	Precision	Sensitivity	Specificity
Random Forest	0.73	0.48	0.86	0.37
Random Forest w/ SMOTE	0.69	0.42	0.77	0.46
Random Forest w/ Tomek links	0.74	0.54	0.86	0.45
Balanced Random Forest	0.65	0.38	0.69	0.56
Balanced Random Forest w/ SMOTE	0.52	0.33	0.47	0.68
Balanced Random Forest w/ Tomek links	0.66	0.47	0.63	0.74
Weighted Random Forest	0.70	0.52	0.85	0.49
Weighted Random Forest w/ SMOTE	0.68	0.45	0.78	0.44
Weighted Random Forest w/ Tomek links	0.73	0.51	0.88	0.36
XGBoost	0.68	0.43	0.70	0.51
XGBoost w/ SMOTE	0.60	0.34	0.64	0.48
XGBoost w/ Tomek links	0.71	0.48	0.81	0.46
AdaBoost	0.79	0.48	0.78	0.75
AdaBoost w/ SMOTE	0.61	0.33	0.64	0.49
AdaBoost w/ Tomek links	0.68	0.45	0.77	0.46

Table 4.1. Comparing Model Performances

ing these models, a grid search with all parameters and a range of their values is conducted to find hyperparameters that optimize performance. When the experiment requires both

classes to have an equal sample size, a small subset of the Non-ASD class is chosen at random. During model training, the samples used are vectors of all the features representing a video of a child that is 24 or 36 months old. Before oversampling or downsampling, the models are trained with 84 samples (75% of ASD data) from each class and tested on 29 samples (25% of ASD data) from each class. Before SMOTE upsamples the ASD class, the training data consists of 493 samples (75% of Non-ASD data) and 84 samples from the ASD class. Once SMOTE resamples the data, the ASD training data increases to 493 samples. The test data for both classes remains the same as before, completely unaffected by resampling from SMOTE. Similarly, when Tomek links are implemented, the test data remains the same, but the Non-ASD training size is modified to 103 samples and the ASD training size is kept at 84 samples. It's odd that after applying SMOTE, the sensitivity of all the models decreases. The additional synthetic ASD samples were expected to contribute to the sensitivity, but this decrease illustrates that the synthetic samples created by SMOTE were unable to represent the properties of the ASD class. Downsampling the Non-ASD class increased (on average) the sensitivities of the models and the Weighted Random Forest had the highest sensitivity, 88%. Removing Non-ASD samples to match the ASD sample size led to higher sensitivities because it established equal importance for both classes in the dataset.

Overall, the results from Table 4.1 show higher sensitivity than specificity in all the models. The weighted random forest had the highest sensitivity and the third highest precision out of all the models. This is expected because Tomek links downsampled the Non-ASD class and higher weights were also placed on the ASD class. Since we are solving a classification problem, it is important to distinguish model performance for each class. We particularly focused on the sensitivity and the precision to evaluate the confidence in the models' predictions for the ASD class. We believe that it is more important to identify the slightest symptoms of ASD early on than to miss them completely to ensure that children are appointed to early intervention. Even though the sensitivity is high, the precision is below 50% which indicates that the predictions made by the models is unreliable. We aim to increase the precision and sensitivity by expanding the dataset.

4.1.2 Feature Selection Part 1

After analyzing results from the previous experiment, we proceeded to conduct all future experiments on random forest because it offered the highest sensitivity indicating that it can effectively recognize ASD characteristics in high-risk infants. The goal of this experiment was to quantify the effects of class imbalance on the feature selection process as well as the overall performance of the model. There are a total of 12 features in the dataset, but to optimize performance, we first find the ideal amount of features that should be used for training. We also study how the performance of RF alters as the total number of Non-ASD samples in the dataset vary.

Even though applying SMOTE and Tomek links improved a few metrics for some of the models, in order to maximize performance even further, an optimal number of features is chosen by using the forward feature selection method. This method iterates through all the possible features one at a time and during each iteration, it fixes one feature that performed the best from the previous iteration. This technique produces a list of features that is sorted from the most contributing to the least contributing towards model performance. Similar to the previous experiment, random forest had hyperparameter values based on a grid search. The goal of this experiment was to examine the effects of the number of features and the Non-ASD sample size on the performance of Random Forest.

The ASD training and testing sets remained constant throughout the experiment. Out of the 24 ASD subjects, 18 were used for training and 6 were used for testing. The Non-ASD training and testing sample sizes varied from 18 subjects for training and 6 for testing, equal to ASD sample size, to 106 subjects for training and 36 subjects for testing, all the Non-ASD subjects in the dataset. During feature selection, different Non-ASD sample sizes were used for training and testing for each number of features. Figure 4.1 shows a graph of five evaluation metrics vs. number of features and Figure 4.2 shows a graph of the same evaluation metrics vs. the total number of Non-ASD sample size used. Both these Figures illustrate that as the number of features and the number of Non-ASD sample size increases, the performance of Random Forest tends to decline. Figure 4.2 illustrates that the best performance, on average, occurs when the Non-ASD and the ASD sample sizes are equal. Table 4.2 shows

the order of features in terms of importance as the total number of samples in the Non-ASD class increase. It's interesting that the model changes its most important features as the Non-ASD sample size changes. This shows that these subsets are not able to capture the true characteristics from the Non-ASD population. Due to the high variance in each subset, the next experiment is conducted with equal sample sizes from both classes.

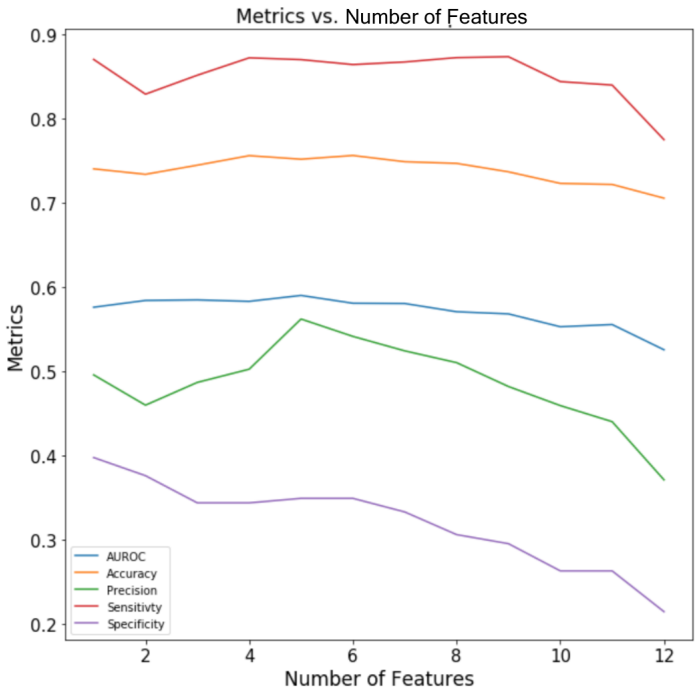


Figure 4.1. Effects of the Number of Features on RF with Unequal Non-ASD Sample Size

	119	**205**	**305**	**405**	**506**	**658**
1	Shared Vocal Frequency	Shared Vocal Frequency	Smile Mean Duration	Smile Mean Duration	Look Face Mean Duration	Look Face Mean Duration
2	Shared Smile Frequency	Look Object Frequency	Look Object Mean Duration	Shared Smile Frequency	Look Face Frequency	Look Face Frequency
3	Smile Mean Duration	Vocal Frequency	Shared Smile Mean Duration	Shared Smile Mean Duration	Shared Smile Mean Duration	Shared Vocal Frequency

4	Look Face Mean Duration	Shared Vocal Mean Duration	Vocal Frequency	Look Object Frequency	Shared Smile Frequency	Shared Smile Mean Duration
5	Look Face Frequency	Look Face Frequency	Look Face Frequency	Vocal Mean Duration	Shared Vocal Frequency	Shared Smile Frequency
6	Shared Vocal Mean Duration	Shared Smile Frequency	Shared Vocal Frequency	Look Object Mean Duration	Look Object Mean Duration	Look Object Mean Duration
7	Vocal Frequency	Look Object Mean Duration	Smile Frequency	Shared Vocal Frequency	Smile Mean Duration	Vocal Mean Duration
8	Look Object Frequency	Smile Frequency	Look Object Frequency	Shared Vocal Mean Duration	Vocal Mean Duration	Look Object Frequency
9	Shared Smile Mean Duration	Smile Mean Duration	Shared Vocal Mean Duration	Look Face Mean Duration	Look Object Frequency	Shared Vocal Mean Duration
10	Look Object Mean Duration	Shared Smile Mean Duration	Look Face Mean Duration	Smile Frequency	Smile Frequency	Smile Frequency
11	Vocal Mean Duration	Look Face Mean Duration	Shared Smile Frequency	Look Face Frequency	Vocal Frequency	Vocal Frequency
12	Smile Frequency	Vocal Mean Duration	Vocal Mean Duration	Vocal Frequency	Shared Vocal Mean Duration	Smile Mean Duration

Table 4.2: Order of Features Based on AUROC for Varying Non-ASD Sample Sizes

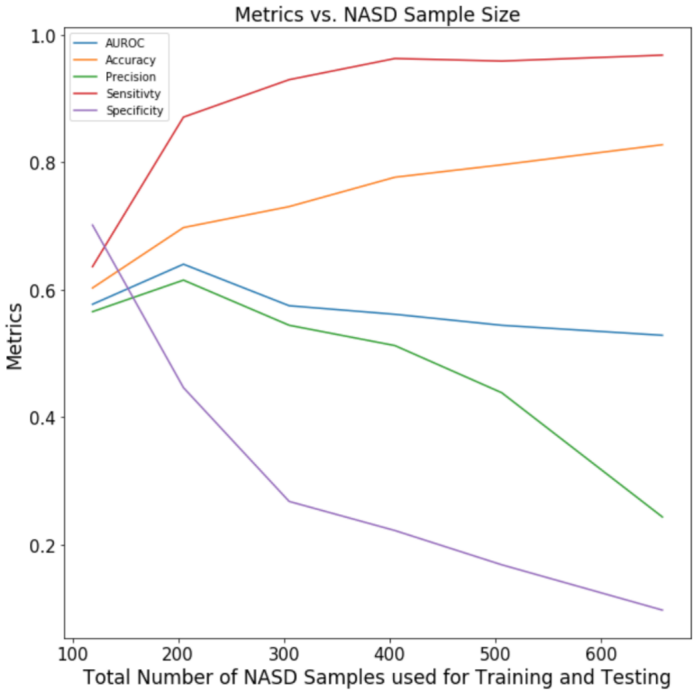


Figure 4.2. Effects of Varying Non-ASD Samples on RF

4.2 Equal Class Sampling

4.2.1 Feature Selection Part 2: Equal Sample Sizes for Both Classes

From the previous results, we concluded that the model performed relatively the best when both classes had an equal number of samples. The goal of this experiment is to find the optimal number of features that should be used when classes are balanced and also to determine which features enhance model performance the most. The data is first divided into training and testing sets based on unique subjects for both classes. 75% of the subjects are chosen for training and 25% of the subjects are chosen for testing. All videos from these subjects are aggregated and used as samples for random forest. Equal number of subjects are chosen for both classes and since the Non-ASD class has more subjects, ten ensembles are created with the same ASD data, but different Non-ASD data.

To determine the optimal number of features for equal sample sizes, the forward feature selection method was implemented once again. Since only a subset of the total Non-ASD data is used for training and testing, the feature selection method is repeated ten times and each iteration uses different samples of Non-ASD. This creates ten ensembles that are trained

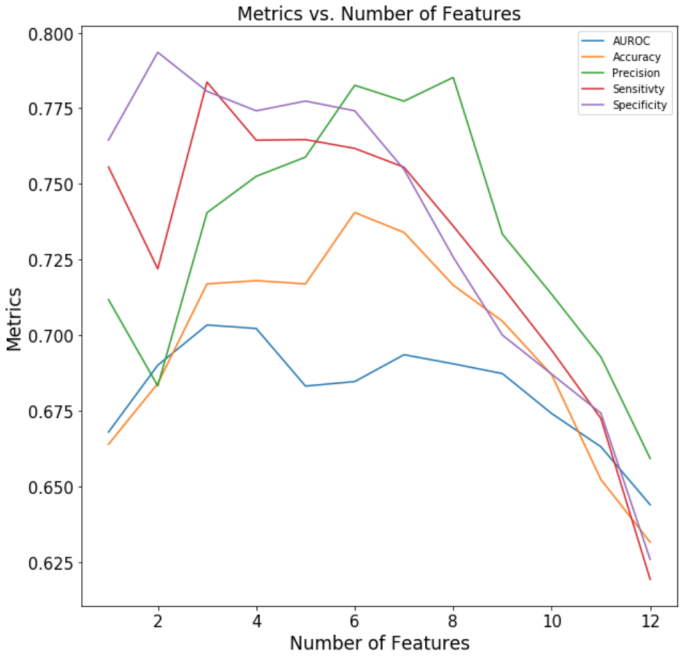


Figure 4.3. Effects of Varying Feature Size on RF

and tested on the same ASD data, but different Non-ASD data. Results from these ensembles are averaged and produce Figure 4.3. Similar to Figure 4.1, Figure 4.3 also illustrates that as the number of features increase, the model performance starts to decline. The peak performance occur when a subset of seven features are used for training.

4.2.2 Feature Selection Part 3: Cross-Validation During Each Iteration

The goal of this experiment is to prevent overfitting by first performing cross-validation to fix hyperparameters, then choosing the most advantageous features for training and then evaluating the model on the test set. During the hyperparameter tuning stage, only four expert-derived features that include look object mean duration, look face frequency, shared smile frequency, and shared vocal frequency are used to train the model. Since these features have shown the most importance in ASD behavioral research conducted at the MIND Institute, they were treated as the baseline features.

Up until now, the data was divided into training and testing sets. Due to the lack of ASD data, cross-validation was not performed which led us to doubt the true performance of the models. In this experiment, the data is again divided into two sets, training and testing, but

during hyperparameter tuning and feature selection, 5-fold cross validation is performed on the training set and repeated ten times to account for the varying Non-ASD samples. Then, the held-out test set is used to evaluate the performance. First, cross-validation is executed to find the optimal values of the hyperparameters, number of trees in the forest and depth of the trees. Then, these hyperparameter values are fixed in random forest and cross-validation is carried out to find the optimal feature set.

Figure 4.4 shows the cross-validation results when finding the optimal number of trees and Figure 4.5 shows the cross-validation results when finding the maximum depth of the trees. There is high variance in the AUROC for both Figures because of the different Non-ASD samples used during each run of cross-validation. Based on these Figures, the number of trees is set to 300 and the max depth is set to 13. With fixed hyperparameter values, the forward feature selection method is performed once again to find the optimal set of features. Figure 4.6 shows the cross-validation results when finding the optimal amount of features. Tables 4.3 and 4.4 list the order of features for each of the ten independent runs. Similar to the order from Table 4.2, the feature set highly varies from run to run. This could possibly be due to either the lack of information present in the ASD or the Non-ASD data which effects the features that the model considers important. Since most metrics reach their peak values when the number of features is seven, the top seven features rated by AUROC are used when evaluating the model on the test set in Figure 4.7. Model evaluation is repeated ten times, each time a different subset of Non-ASD samples is used from the test set. Figure 4.7 shows similar variation in performance as Figure 4.2. The variance of the AUROC is approximately 0.15 which represents that the model is unsure about the predictions made on never before seen data.

After conducting the previous feature selection experiments, the top seven optimal features are:

1. Shared Smile Frequency
2. Shared Vocal Frequency
3. Shared Vocal Mean Duration
4. Look Face Frequency

5. Look Object Frequency

6. Vocal Frequency

7. Look Object Mean Duration

These top features were common among all three of the experiments. The performance of random forest with these features is shown in Table 4.3. The sensitivity is 61% and the precision is 58% indicating that the predictions made on the ASD samples are correct most of the time. The specificity is 68% indicating that the model is able to identify the Non-ASD samples with certain confidence. It is expected for the specificity to be higher than the sensitivity due to the larger Non-ASD sample size in the dataset.

Accuracy	**Precision**	**Specificity**	**Sensitivity**	**AUROC**
0.60	0.58	0.68	0.61	0.62

Table 4.3. Performance of RF with the Top Seven Features

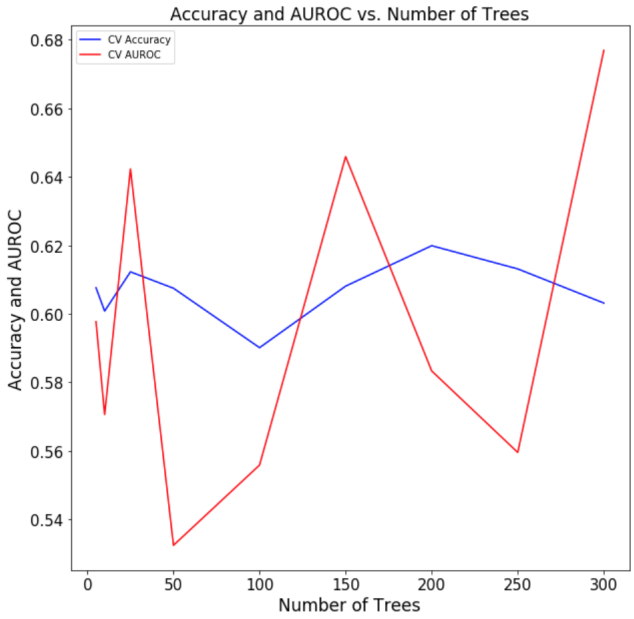


Figure 4.4. Cross-Validation to Find Optimal Number of Trees

	1	**2**	**3**	**4**	**5**
1	Shared Smile Frequency	Shared Vocal Frequency	Look Face Frequency	Look Object Frequency	Look Face Frequency

2	Look Object Frequency	Look Face Mean Duration	Look Object Mean Duration	Shared Vocal Mean Duration	Shared Vocal Mean Duration
3	Vocal Mean Duration	Shared Smile Mean Duration	Shared Vocal Frequency	Look Object Mean Duration	Smile Frequency
4	Vocal Frequency	Look Object Frequency	Shared Smile Frequency	Look Face Mean Duration	Vocal Frequency
5	Look Face Frequency	Vocal Mean Duration	Look Object Frequency	Vocal Frequency	Shared Vocal Frequency
6	Look Face Mean Duration	Look Object Mean Duration	Look Face Mean Duration	Shared Smile Frequency	Look Object Mean Duration
7	Smile Mean Duration	Smile Mean Duration	Vocal Frequency	Vocal Mean Duration	Smile Mean Duration
8	Smile Frequency	Look Face Frequency	Shared Vocal Mean Duration	Smile Frequency	Look Object Frequency
9	Shared Vocal Frequency	Vocal Frequency	Smile Frequency	Shared Vocal Frequency	Shared Smile Frequency
10	Look Object Mean Duration	Smile Frequency	Smile Mean Duration	Look Face Frequency	Look Face Mean Duration
11	Shared Vocal Mean Duration	Shared Smile Frequency	Shared Smile Mean Duration	Shared Smile Mean Duration	Vocal Mean Duration
12	Shared Smile Mean Duration	Shared Vocal Mean Duration	Vocal Mean Duration	Smile Mean Duration	Shared Smile Mean Duration

Table 4.4: Order of Features Based on AUROC for First Five Iterations

	6	**7**	**8**	**9**	**10**
1	Look Face Frequency	Smile Mean Duration	Look Face Frequency	Look Object Frequency	Smile Frequency
2	Shared Vocal Mean Duration	Shared Vocal Mean Duration	Shared Vocal Mean Duration	Vocal Mean Duration	Vocal Mean Duration
3	Smile Frequency	Look Face Frequency	Shared Vocal Frequency	Smile Frequency	Look Object Mean Duration
4	Shared Smile Mean Duration	Vocal Mean Duration	Look Object Mean Duration	Shared Vocal Frequency	Vocal Frequency
5	Shared Smile Frequency	Look Object Frequency	Vocal Frequency	Vocal Frequency	Look Face Frequency
6	Vocal Frequency	Shared Smile Frequency	Smile Mean Duration	Shared Smile Mean Duration	Shared Smile Frequency
7	Look Face Mean Duration	Vocal Frequency	Shared Smile Frequency	Shared Vocal Mean Duration	Shared Vocal Frequency
8	Vocal Mean Duration	Shared Smile Mean Duration	Look Face Mean Duration	Shared Smile Frequency	Shared Vocal Mean Duration
9	Look Object Frequency	Shared Vocal Frequency	Shared Smile Mean Duration	Look Face Mean Duration	Look Object Frequency

10	Look Object Mean Duration	Look Object Mean Duration	Vocal Mean Duration	Look Face Frequency	Shared Smile Mean Duration
11	Shared Vocal Frequency	Smile Frequency	Smile Frequency	Look Object Mean Duration	Smile Mean Duration
12	Smile Mean Duration	Look Face Mean Duration	Look Object Frequency	Smile Mean Duration	Look Face Mean Duration

Table 4.5: Order of Features Based on AUROC for Second Five Iterations

4.2.3 Using Age-Based Data for RF Training

The next few experiments take the age of the patients into account during the model training process. Much of the previous work pertaining behavioral analysis of ASD shows that patients tend to have a developed set of symptoms by the age of 24 months old [3]. Since our dataset consists of children ranging from 18 to 36 months old, we wanted to experiment with the age and examine how it affects model performance. The first set of experiments only takes into account if video data was available for a patient at a certain age and the second set focuses on the age that the patient was diagnosed at. These experiments are conducted to stress-test the machine learning models to identify early onset of ASD symptoms in infants. A variety of combinations with different age is experimented with and the results are displayed in Tables 4.6 and 4.7. The results from these tables show that it is possible to utilize trained machine learning models with data from older children to recognize ASD symptoms in children that are 18 months old.

Due to the high variance in previous results, approaches based on the age of subjects are considered in the next couple experiments. It has been shown in [2], [3] that achieving a diagnosis for ASD can occur at 24 months and p-value Tables 3.5 and 3.6 also show the clarity of ASD symptoms once children are 24 months old. The performance of RF is compared

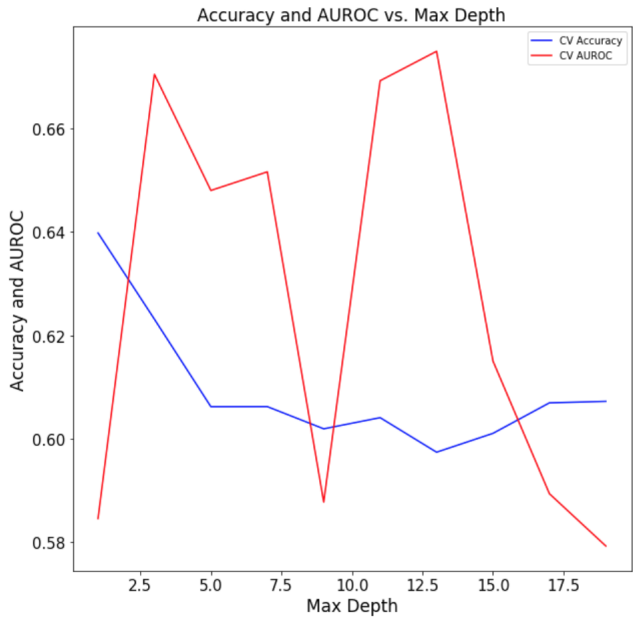


Figure 4.5. Cross-Validation to Find Optimal Depth of Trees

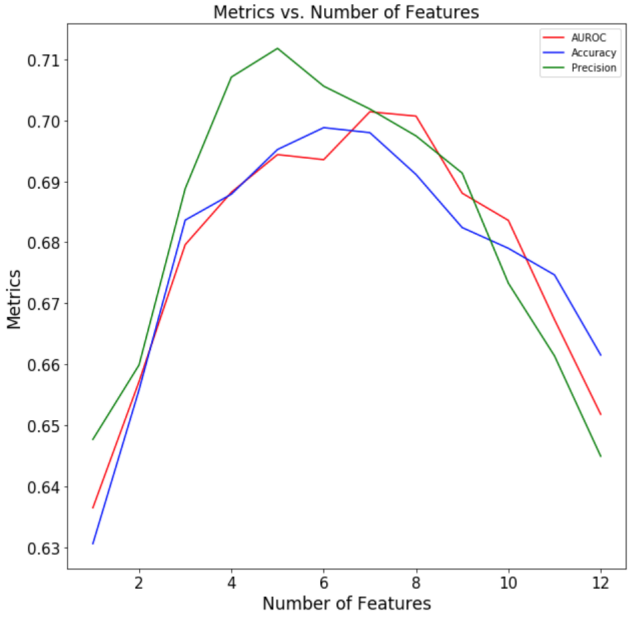


Figure 4.6. Cross-Validation to Find Optimal Number of Features

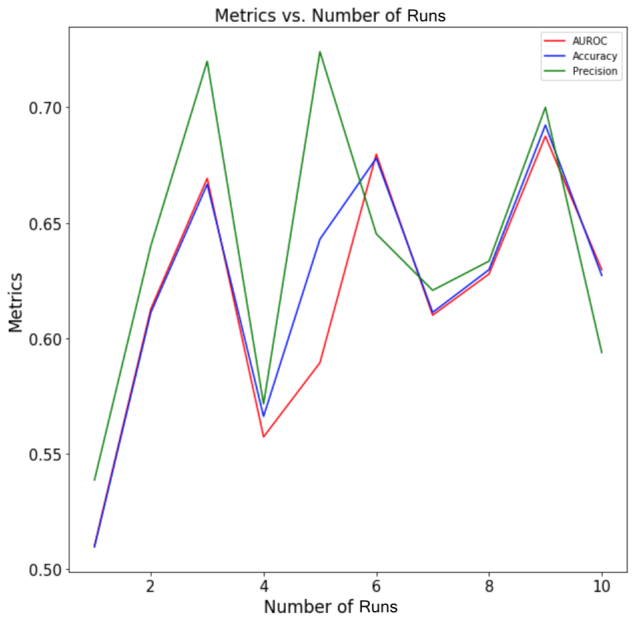


Figure 4.7. Validating RF on Test Set with Varying Non-ASD Samples

when trained on data from 18, 24, and 36 months independently to show the effect of age on achieving a diagnosis. Only the top seven features, as discussed in the previous sections, are used since Figure 4.6 shows the performance decreasing after more than seven features are added.

4.2.3.1 Training Data Based on Availability of Videos

The first set of experiments was conducted based on using videos taken at different ages. Random forest is first trained independently on data from children aged 36, 24, and 18 months and it is tested on a variety of different sets as shown in Table 4.6. Hyperparameter tuning for number of trees and the maximum depth is conducted before training and testing. Theoretically, when the model is trained on data from children aged 36 months, the model should perform the highest because the older children exhibit the core symptoms of ASD, but in Table 4.6, the results do not completely agree. As seen in Table 4.6, AUROC is particularly higher when random forest is trained on 18 months and tested on 18 months and also when it is trained on 24 months and tested on 18 months. These results show that our hypothesis about the data is wrong because we expected the performance of random forest to be the highest when trained on data from 36 months. The ROC curves generated by training

random forest, independently, on data from 36 months, 24 months, and 18 months and testing these models on data from 18 months is shown in Figure 4.8. Even though data from 36 months did not have the best performance, it is important to note that the model was able to classify samples from 18 months even when it was trained on 24 months and 18 months. This shows that with the current feature set, random forest is able to learn information about ASD from older children and recognize ASD behaviors in infants. This demonstrates the capability of machine learning algorithms to obtain early ASD diagnosis.

Even though the average AUROC of all the experiment is approximately 65%, the average precision is 26% which indicates that the predictions made by the model are unreliable. It is also unusual that there is not much difference in performance when RF is trained on data from videos taken of children aged 36 months and 18 months. There could be many reasons as to why this occurred including that the features used to acquire the characteristics of ASD may not be exhibiting the ASD behaviors as expected or it may be due to the limited sample space representing the ASD class.

Figures 4.9 and 4.10 illustrate the effects that the number of trees and the maximum depth, respectively, have on AUROC and precision. The pink shade in the background of the graph represents the variance of the precision and the green shade represents the variance of the AUROC. The brown represents the intersection of the two variances. Same as the previous experiment, different Non-ASD samples are used per iteration for cross-validation. The variance, illustrated in Figures 4.9 and 4.10, displays that it could be possible that the subsets of the Non-ASD samples may be coming from different distributions.

4.2.3.2 Training Data Based on the Age of Diagnosis

The second set of experiments focuses on training RF based on the age that the children were diagnosed. Breakdown of the data based on the age of diagnosis is shown in Table 3.4. When a child is diagnosed at 18 months, videos for this child from 18, 24, and 36 months are aggregated for the training data. Similarly, when a child is diagnosed at 24 months, videos from 24 and 36 months are aggregated, but if the child is diagnosed at 36 months, videos from the past are not used. Same as the previous experiment, hyperparameter tuning for the number of trees and the maximum depth is conducted before training and testing. Table 4.7

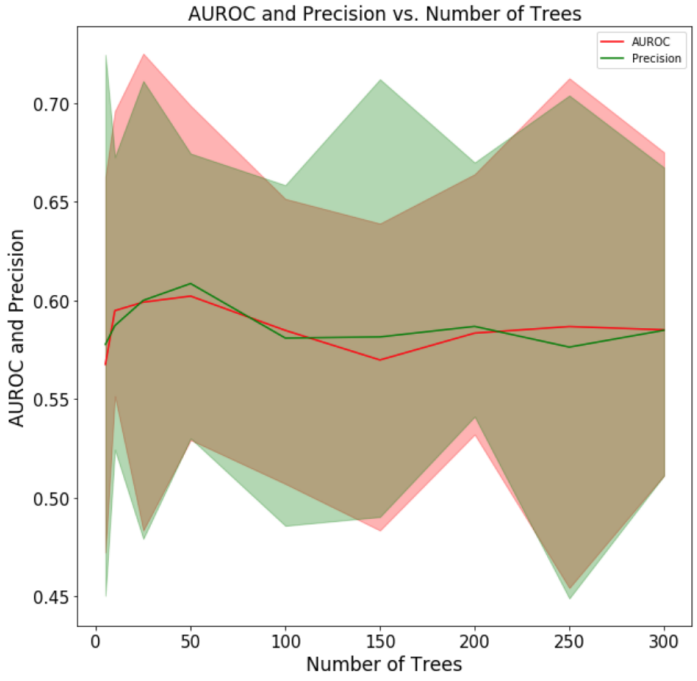


Figure 4.8. Cross-Validation Performance with Varying Tree Sizes

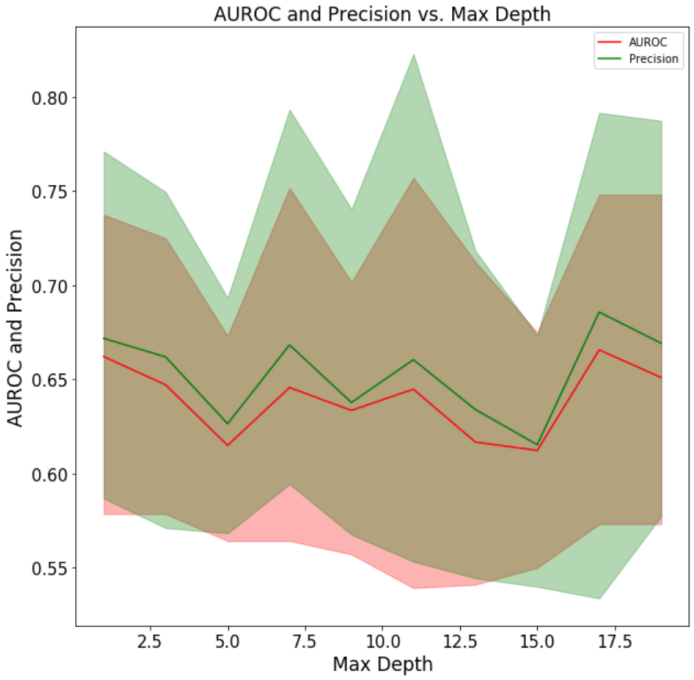


Figure 4.9. Cross-Validation Performance with Varying Depth Sizes

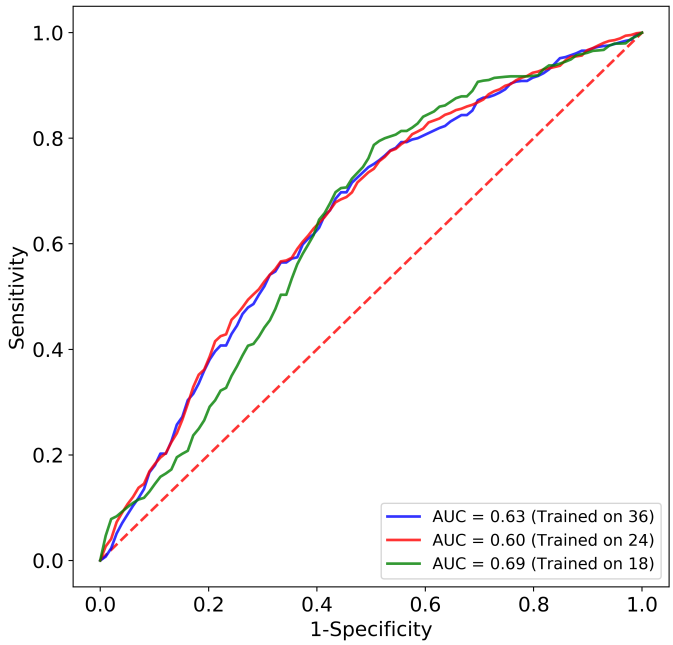


Figure 4.10. ROC curve for RF trained on three distinct ages based on availability of videos

Training Data (In Months)	**Testing Data (In Months)**	**Accuracy**	**Precision**	**Specificity**	**Sensitivity**	**AUROC**
36	36	0.54	0.25	0.73	0.49	0.61
	24	0.42	0.21	0.96	0.31	0.63
	18	0.52	0.23	0.88	0.45	0.66
	18,24,36	0.52	0.24	0.81	0.46	0.64
24	36	0.69	0.34	0.61	0.71	0.66
	24	0.60	0.25	0.69	0.59	0.64
	18	0.65	0.28	0.74	0.64	0.69
	18,24,36	0.65	0.29	0.73	0.63	0.68
18	36	0.59	0.28	0.74	0.56	0.65
	24	0.53	0.23	0.79	0.48	0.64
	18	0.66	0.28	0.75	0.64	0.69
	18,24,36	0.58	0.26	0.75	0.55	0.65

Table 4.6. RF Performance when Trained and Tested on Particular Ages

displays the results from this experiment and it shows that regardless of which age the model was trained on, it had the highest AUROC score when it was tested on data from 18 months. For this particular experiment, when the model is tested on the same age as it is trained on, the training data is based on the age of diagnosis, but the testing data consists of videos of subjects taken at that age due to the lack of age of diagnosis data.

When the model is trained on subjects diagnosed at 18 months and tested on data from 18 months, the AUROC is the highest at 69%. When RF is trained on subjects that have been diagnosed at 18 months, video data from 24 and 36 months is also included in training. The accumulation of data from later years can be contributing to the elevated performance. The training sets for 24 and 36 months consists of merely 25 and 14 samples. It would be unjust to compare performances across these ages directly, but from Table 4.7, we see that RF was able to learn relevant information indicating ASD behavior. Figure 4.11 shows ROC curves for RF trained on data from children diagnosed at 18, 24, and 36 months, independently, and tested on data from children diagnosed at 18 months. It is interesting that regardless of when the children are diagnosed, RF is able to classify ASD behaviors in 18-month olds.

The precision varies from 16% to 65% which displays the instability of the training data. This could be due to the lack of information missing in the ASD features or the lack in quantity of ASD training data. Figures 4.12 and 4.13 show the effects of number of trees and the maximum depth, respectively, have on AUROC and precision as the model is trained. The shaded area of the Figures represents the variance of the metrics, same as Figures 4.9 and 4.10.

Tables 4.6 and 4.7 show results when random forest is trained using data based on the availability of video data for a particular age and the age that the children were diagnosed. These results display much lower sensitivity and higher specificity for all experiments compared to the sensitivity and specificity from Table 4.1 because there are fewer ASD samples present in the training set. This discrepancy occurs due to the difference in technique used to separate the data into training and testing sets. The experiments that produced results in Table 4.1 used data that was solely based on individual video files. 75% of the total ASD videos were used for training and 25% were used for testing. The same split was used for

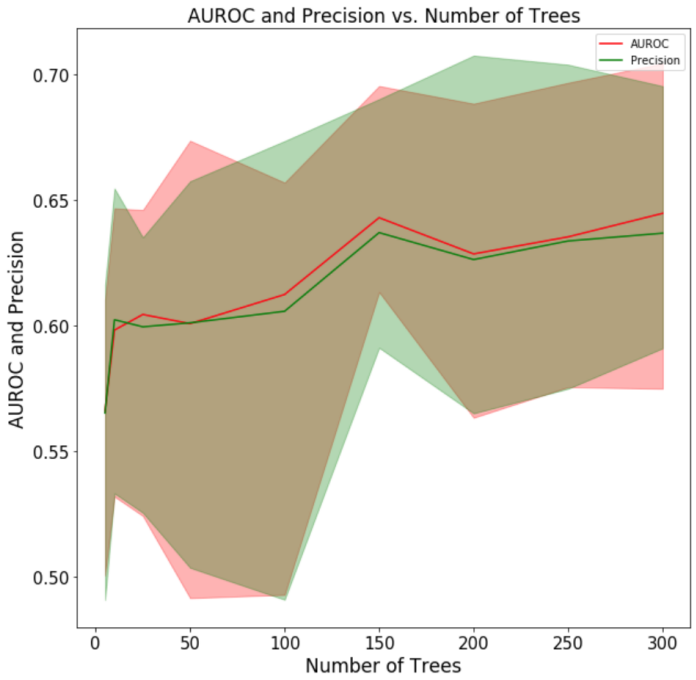


Figure 4.11. Cross-Validation Performance with Varying Tree Sizes (Age of Diagnosis)

the Non-ASD class, but the training set was capped to match the ASD sample size. The age-based experiments used data that first involved splitting the individual subjects into training and testing sets and then extracting video data of those particular subjects as training and testing samples. The previous 75-25 split is kept for training and testing for both classes in this experiment as well. By splitting data based on the subjects first, the sensitivity declines because fewer number of ASD samples are used when compared to splitting directly based on individual video files. This difference explains the increased specificity in Tables 4.6 and 4.7 because more Non-ASD samples are present in the dataset.

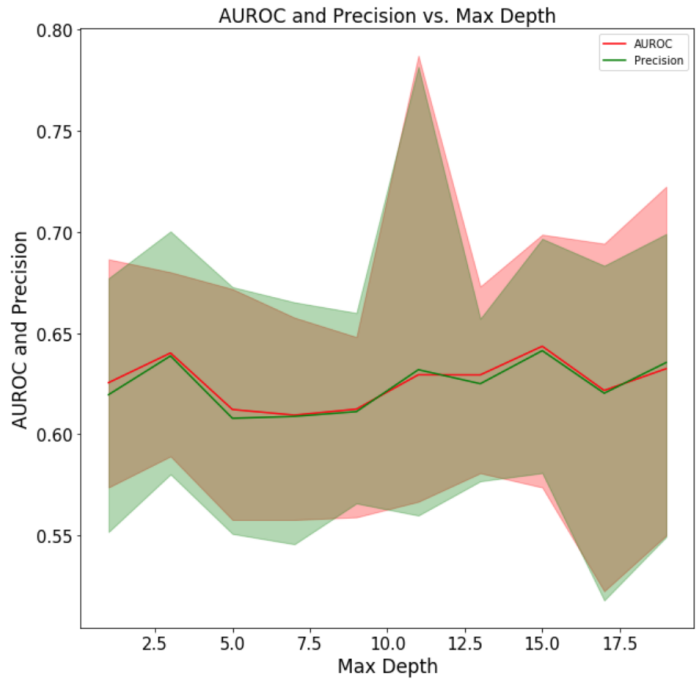


Figure 4.12. Cross-Validation Performance with Varying Tree Sizes (Age of Diagnosis)

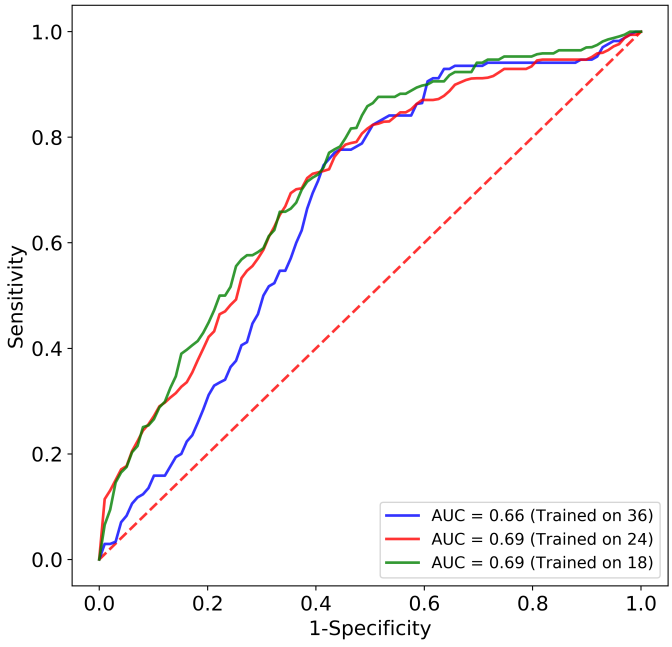


Figure 4.13. ROC curve for RF trained on three distinct ages based on age of diagnosis

Training Data (In Months)	Testing Data (In Months)	Accuracy	Precision	Specificity	Sensitivity	AUROC
36	36	0.60	0.46	0.43	0.70	0.56
	24	0.49	0.25	0.55	0.47	0.51
	18	0.62	0.60	0.67	0.58	0.63
	18,24	0.59	0.48	0.65	0.55	0.60
24	36	0.46	0.17	0.65	0.43	0.54
	24	0.60	0.48	0.54	0.63	0.59
	18	0.60	0.59	0.58	0.63	0.60
	18,36	0.58	0.44	0.62	0.56	0.59
18	36	0.47	0.19	0.75	0.42	0.59
	24	0.51	0.29	0.67	0.46	0.56
	18	0.69	0.65	0.74	0.64	0.69
	24,36	0.47	0.16	0.64	0.45	0.54

Table 4.7. RF Performance when Trained and Tested Based on Age of Diagnosis

Chapter 5

Conclusion & Future Directions

Autism Spectrum Disorder (ASD) denotes a range of conditions characterized by repetitive and restricted behavior and challenges with nonverbal communication and social skills [1]. There has been an increase in the number of children diagnosed with ASD which could be due to the lack of diagnosis at an early age. Machine learning approaches are being explored to alleviate the current issues of availability and cost to achieve an early diagnosis [10]. This thesis examines various different approaches towards diagnosing ASD and treats the ASD diagnosis as a classification task. Several models are trained with features extracted from video data provided by the MIND Institute. Even though models from previous studies have shown to achieve high accuracies and sensitivities, due to the biased and limited dataset, the experiments conducted on this dataset were unable to devise a conclusion about achieving early ASD diagnosis.

Given the model performances with the current feature set, there are several directions to explore. The primary focus should be collecting data from ASD patients or collaborating with institutions conducting research in ASD. It's important for the dataset to contain a large ASD sample size so that all possible variations of the ASD characteristics are captured. In future machine learning models, eye gaze data measured by scanpath should be added to the feature set. The current annotation technique exposes the look face and look object events to a higher possibility of error because the video is manually annotated. Leveraging scanpath analysis for eye gaze will be more efficient and effective in demonstrating ASD behavior [31]. It would also be interesting to add recorded responses from the ADOS test for each patient to

the dataset. Training models independently first on the ADOS data and then on the current feature set will allow us to directly compare the model performance of the current feature set to the state-of-the-art technique.

Moving forward, implementing machine learning algorithms to diagnose ASD has substantial scope. This scope is largely dependent on the limitations placed on the amount of data and resources available. If ample amount of ASD data is provided for supervised machine learning models, as shown in experiments discussed in the related work section, they have the capability to reach performances comparable to professional clinicians.

BIBLIOGRAPHY

[1] Autism Facts and Figures. *https://www.autismspeaks.org/autism-facts-and-figures*, 2019. [Online; accessed 1-November-2019].

[2] Children diagnosed with autism at earlier age more likely to receive evidence-based treatments. =https://www.psychiatry.org/newsroom/news-releases/children-diagnosed-with-autism-at-earlier-age-more-likely-to-receive-evidence-based-treatments. [Online; accessed 11-November-2019].

[3] Sara Jane Webb, Emily JH Jones, Jean Kelly, and Geraldine Dawson. The motivation for very early intervention for infants at high risk for autism spectrum disorders. *International journal of speech-language pathology*, 16(1):36–42, 2014.

[4] Sven Bölte. Is autism curable? *Developmental Medicine & Child Neurology*, 56(10): 927–931, 2014.

[5] Geraldine Dawson. Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. *Development and psychopathology*, 20(3):775–803, 2008.

[6] Sebastien Levy, Marlena Duda, Nick Haber, and Dennis P Wall. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. *Molecular autism*, 8(1):65, 2017.

[7] Qandeel Tariq, Scott Lanyon Fleming, Jessey Nicole Schwartz, Kaitlyn Dunlap, Conor Corbin, Peter Washington, Haik Kalantarian, Naila Z Khan, Gary L Darmstadt, and Dennis Paul Wall. Detecting developmental delay and autism through machine learning models using home videos of bangladeshi children: Development and validation study. *Journal of medical Internet research*, 21(4):e13822, 2019.

[8] Maureen S Durkin, Mayada Elsabbagh, Josephine Barbaro, Melissa Gladstone, Francesca Happe, Rosa A Hoekstra, Li-Ching Lee, Alexia Rattazzi, Jennifer Stapel-Wax, Wendy L Stone, et al. Autism screening and diagnosis in low resource settings: chal-

lenges and opportunities to enhance research and services worldwide. *Autism Research*, 8(5):473–476, 2015.

[9] Geraldine Dawson, Sally Rogers, Jeffrey Munson, Milani Smith, Jamie Winter, Jessica Greenson, Amy Donaldson, and Jennifer Varley. Randomized, controlled trial of an intervention for toddlers with autism: the early start denver model. *Pediatrics*, 125(1): e17–e23, 2010.

[10] Dennis Paul Wall, J Kosmicki, TF Deluca, E Harstad, and Vincent Alfred Fusaro. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry*, 2(4):e100, 2012.

[11] About the ADOS Exam. =https://research.agre.org/program/aboutados.cfm. [Online; accessed 11-November-2019].

[12] 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. *Journal of autism and developmental disorders*, 24(5):659–685, 1994.

[13] Colby Chlebowski, James A Green, Marianne L Barton, and Deborah Fein. Using the childhood autism rating scale to diagnose autism spectrum disorders. *Journal of autism and developmental disorders*, 40(7):787–799, 2010.

[14] Dennis P Wall, Rebecca Dally, Rhiannon Luyster, Jae-Yoon Jung, and Todd F DeLuca. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PloS one*, 7 (8):e43855, 2012.

[15] Rena Shrestha, Cheryl Dissanayake, and Josephine Barbaro. Age of diagnosis of autism spectrum disorder in nepal. *Journal of autism and developmental disorders*, 49(6):2258–2267, 2019.

[16] M Duda, JA Kosmicki, and DP Wall. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational psychiatry*, 4(8):e424, 2014.

[17] M Duda, R Ma, N Haber, and DP Wall. Use of machine learning for behavioral distinction of autism and adhd. *Translational psychiatry*, 6(2):e732, 2016.

[18] John N Constantino. *Social responsiveness scale*. Springer, 2013.

[19] Apoorva Rajiv Madipakkam, Marcus Rothkirch, Isabel Dziobek, and Philipp Sterzer. Unconscious avoidance of eye contact in autism spectrum disorder. *Scientific reports*, 7(1): 13378, 2017.

[20] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):43, 2017.

[21] Tanaya Guha, Zhaojun Yang, Ruth B Grossman, and Shrikanth S Narayanan. A computational study of expressive facial dynamics in children with autism. *IEEE transactions on affective computing*, 9(1):14–20, 2016.

[22] Anju Pratap, CS Kanimozhiselvi, R Vijayakumar, and KV Pramod. Soft computing models for the predictive grading of childhood autism-a comparative study. *Int. J. Soft Comput. Eng.(IJSCE)*, 4(3), 2014.

[23] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

[24] Hae-Young Kim. Analysis of variance (anova) comparing means of more than two groups. *Restorative dentistry & endodontics*, 39(1):74–77, 2014.

[25] Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105, 1971.

[26] Taghi M Khoshgoftaar, Moiz Golawala, and Jason Van Hulse. An empirical study of learning from imbalanced data using random forest. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 310–317. IEEE, 2007.

[27] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[28] Balázs Kégl. The return of adaboost. mh: multi-class hamming trees. *arXiv preprint arXiv:1312.6086*, 2013.

[29] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[30] Min Zeng, Beiji Zou, Faran Wei, Xiyao Liu, and Lei Wang. Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data. In *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, pages 225–228. IEEE, 2016.

[31] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. State-of-the-art of visualization for eye tracking data. In *EuroVis (STARs)*, 2014.