Using Machine Learning Optimization to Predict Autism in Toddlers

Arjun Singh, Zoya Farooqui, Branden Sattler, Unyime Usua, Michael Helde
ARQuest Student Science and Engineering Network
Irvine, CA
arjsingh2004@gmail.com, zoya.farooqui@gmail.com, basattler2368@gmail.com, unyimecusua@gmail.com, michaelhelde11@gmail.com

Abstract

Autism is a developmental disorder that is often identified by complications with social interaction and communication. Even though knowledge about it has increased and instruments such as the ADOS and ADI are available to assess the disorder, it is usually distinguished from observable symptoms, which makes it especially difficult to diagnose. A variety of concerns arise because of this issue, one of them being long-term symptoms, which can be mitigated with early screening and consequently early treatment. This research focuses on improving the diagnosis pipeline of autism by training and testing several state-of-the-art machine learning models with an Autism Spectrum Disorder dataset from the University of California, Irvine, and using machine learning methods to quantitatively identify the most significant indicators of autism in toddlers. We first designed a neural network classification model and Random Forests classification model, and then we trained/tested them to identify the presence of autism in toddlers. We also used feature selection through LightGBM parameter optimization to identify which physical characteristics are most significant in giving rise to autism. The outcome of this study was a highly accurate classifier for predicting the presence of autism and significant information on the importance of several physical characteristics in indicating autism. This enhanced diagnosis is critical as it leads to a more personalized and early-stage treatment, which can alleviate the effects of autism thereafter.

Keywords
Machine Learning, Autism Spectrum Disorder, Neural Networks, Random Forests, Feature Selection

1. Introduction

Autism Spectrum Disorder (ASD) (commonly known as “autism”) is a psychiatric disorder that restricts linguistics, cognitive, and social abilities of a person (Thabtah 2017). Nearly 1 in every 54 births is diagnosed with ASD, and about 1% of the entire world population suffers from it (The Autism Society, 2015). However, there is no known cause or cure for autism. As a result, parents of children with ASD often face a challenge of knowing whether their child has ASD, which makes diagnosing an autistic child early extremely difficult. Since symptoms of autism develop as a child grows, diagnostic tests performed on children ages 2-3 are less reliable than diagnostic tests on children performed when they are 4-5 years old (Moore and Goodson 2003). The worrisome part is that early diagnosis is key for an autistic person to achieve their milestones in life (Skibitsky 2019). Consequently, this paper uses machine learning techniques to unveil the characteristics in infants that correlate to ASD diagnoses to make early diagnoses more predictable.

This research project expands on previous research by Fadi A. Thabtah that aimed to improve ASD screening using a machine learning approach (Thabtah 2017). To build upon the results, we utilized different state-of-the-art machine learning models and optimization techniques, leading to an improvement in the performance of the classifiers. Utilizing a dataset from the University of California, Irvine, we trained machine learning models to predict the presence of ASD by creating neural networks and random forests, while also optimizing their hyperparameters to quantitatively identify the optimal combination of hyperparameters for each model. With this approach, we can achieve an improved performance of our machine learning models. In addition, we used feature selection through LightGBM parameter optimization to calculate importance scores for each predictor variable (also

referred to as feature), therefore revealing which physical characteristics are the most significant indicators of ASD in toddlers.

This paper will be organized into the following sections: Section II offers a detailed description in the dataset used in this research. Next, section III outlines the methodology and experimentation performed. Lastly, sections IV and V will discuss the results/conclusions of our work and potential directions for future work.

1.1 Objectives

The first research objective was to develop a highly accurate machine learning model that uses physical characteristics of toddlers to predict the presence of ASD. The second research objective was to quantitatively identify which physical characteristics are the most significant indicators of ASD in toddlers.

2. Literature Review

The detection of ASD using machine learning has been a prevalent interest in the field of ASD research. In this problem, the high-level objective is to use computational techniques to identify patterns in a dataset that can aid in developing an equation that takes a certain set of inputs to produce an appropriate output. This requires two main materials: 1) a dataset of characteristics/classifications of patients with and without ASD, and 2) a computational algorithm that manipulates itself based on its error in classifying an output based on a set of inputs, thereby "fitting" the data and eventually learning to classify an output with a certain degree of accuracy. The exact degree of accuracy depends on several factors, including the amount of training data, the integrity of the statistical relationship between the inputs and outputs, the learning mechanism of a certain machine learning algorithm, and more.

One approach to the detection of autism through machine learning has been to use Convolutional Neural Networks (CNNs), which are complex multi-stage algorithms that are designed to classify images through 2D numerical matrix representations. In this context, they have been used to interpret brain CT scans to classify whether it is highly characteristics of autism with a >90% accuracy, and in a more unorthodox sense, they have been modified to interpret tabular data and classify autistic individuals with a >95% accuracy (Lingyu et al. 2019; Raj and Masood 2020). Researchers have also attempted to create CNN models that are more efficient by employing less trainable parameters to reduce computational expense, however this has generally resulted in a sharp decrease of accuracy to ~70% (Sherkatghanad et al. 2020; Li et al. 2018). This presents an inherent issue with using CNN architectures that can be improved upon.

Another approach to the detection of autism using machine learning is using tabular data and standard machine learning models that interpret 1D characteristics data in order to predict autism. For instance, data from patients' responses to the Autism Quotient 10 test for adults have been used in coordination with standard state-of-the-art machine learning models such as Regression Trees, Support Vector Machines, and K-Nearest Neighbor Classifiers to classify adults/adolescents with autism with a ~90% accuracy (Alteneiji et al. 2020; Omar et al. 2019; Chowdhury and Iraj 2020; Varshini and Chinnaiyan 2020). The Autism Quotient 10 test is the research standard for tabular data collection, as they are questions specifically designed to identify traits that may or may not be associated with autism. However, the Autism Quotient 10 test itself has found to be inaccurate and unreliable when interpreted manually by physicians (Taylor et al. 2020; Jia et al. 2019). More specifically, although the Autism Quotient 10 test questions are still strong ASD-indication questions, their manual interpretation by human physicians may lead to an inaccurate diagnosis. This suggests that machine learning is an especially important tool in detecting ASD using such behavioral characteristics, and it increases the urgency for research that increases the accuracy of machine learning models beyond the ~90% in current research. Note that current research focuses on a wide range of ages, from toddlers to adults. For this study, we decided to focus on toddlers because early therapeutic intervention at that young of an age will lead to improved long-term outcomes compared to an adult detected to have ASD.

In terms of feature selection, much less work has been performed regarding which physical or behavioral characteristics are most significant in giving rise to autism. According to our research, basic standard feature selection methods such as Information Gain and Chi Square Testing and basic statistical analysis such as heat map analysis have been used to identify which physical/behavioral features are most significant and which are negligible in detecting ASD (Thabtah et al. 2019; Hossain and Kabir 2019; Varshini and Chinnaiyan 2020). Therefore, there is an urgent need for further work to be done regarding feature selection to more reliably rank features associated with ASD.

In summary, significant research has been performed regarding the use of machine learning techniques in predicting ASD, however much less has been performed regarding feature selection. For the former, several methods such as CNNs, Support Vector Machines, and more have been experimented with to identify which method

is best for detecting ASD in several age groups. However, the efficiency and accuracy of these machine learning models have a large room for improvement. For the latter, only basic statistical analysis techniques have been used to gain insight into the significance of several features, leaving lots of room more advanced techniques to more accurately rank features. Therefore, in this study we focus on using state-of-the-art machine learning techniques and optimization techniques to maximize classification accuracy, and on using these techniques to rank features according to significance in predicting ASD.

3. Methods

3.1 Dataset

To train machine learning models to predict the presence of ASD in toddlers, we first needed a dataset that contained records of physical attributes of toddlers and whether they had ASD.

The dataset used in this research, which was obtained from Kaggle.com, an open-source repository for machine learning datasets, is an Autism Spectrum Disorder dataset from the University of California, Irvine (Thabtah 2018). We verified with the dataset license that we were permitted to use the dataset in our research context. It contains significant features tested during the autism screening of toddlers. With 1054 records, the dataset includes 18 variables pointing to different attributes, 10 of these being questions to determine if the toddler has ASD, which are represented by items A1 through A10 in the table below. These 10 questions are from the clinical-standard Autism Quotient 10 test for toddlers (Booth et al. 2013). For items A1 through A9, if the response was "Sometimes", "Rarely", or "Never", a value of 1 was assigned, and a value of 0 was assigned for the opposite. However, for question A10, an answer of "Always", "Usually", or "Sometimes" was assigned a value of 1. These question values are added up and the score is represented by the Qchat 10 Score. A score of over 3 points would mean that the toddler has a high potential for showing ASD traits. A score of 3 or less points to no observable ASD traits. The remaining items represent specific traits of each toddler that are helpful in determining which variables affect the presence of ASD. These 10 questions and all the other values, except the Qchat 10 score, represent the predictor variables to help determine if a toddler exhibits ASD traits. The Qchat 10 score is the response variable as it is calculated using the other variables. The description of each predictor variable and the response variable is presented in Table 1.

Table 1. Description of Predictor/Response Variables of ASD Dataset

Variable Code	Variable Type	Description	Values
Case_No.	N/A	The number of the toddler patient in the dataset	1-1054
A1	Predictor	Does your child look at you when you call his/her name?	0, 1
A2	Predictor	How easy is it for you to get eye contact with your child?	0, 1
A3	Predictor	Does your child point to indicate that s/he wants something? (e.g., a toy that is out of reach)	0, 1
A4	Predictor	Does your child point to share interest with you? (e.g., pointing at an interesting sight)	0, 1
A5	Predictor	Does your child pretend? (e.g., care for dolls, talk on a toy phone)	0, 1
A6	Predictor	Does your child follow where you're looking?	0, 1

A7	Predictor	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g., stroking hair, hugging them)	0, 1
A8	Predictor	Would you describe your child's first words as:	0, 1
A9	Predictor	Does your child use simple gestures? (e.g., wave goodbye)	0, 1
A10	Predictor	Does your child stare at nothing with no apparent purpose?	0, 1
Age_mons	Predictor	Age (months)	Any value between 12-36
Qchat-10-Score	Predictor	Qchat 10 Score	Any value between 1-10
Sex	Predictor	Sex	Male or Female
Ethnicity	Predictor	Ethnicity	11 unique ethnicities
Jaundice	Predictor	Jaundice	True or False
Family_mem_with_ASD	Predictor	Family member with ASD	True or False
Who completed the test	Predictor	Who completed the test?	Family member or healthcare professional
Class/ASD Traits	Response	Class/ASD Traits	True or False

Note that for each question, there were 5 possible answers, with each answer representing a varying degree of agreement with the question's assertion. For example, if the question is "Does your child look at you when you call his/her name?", the answer choices are "Always", "Usually", "Sometimes", "Rarely", or "Never". Each question has a specific set of 5 answers that appropriately fits the question. For questions A1-A9, if the response was one of the three answers representing the three lowest degrees of agreement, the response was mapped to a 1. However, if the response was one of the two answers representing the two highest degrees of agreement, the response was mapped to a 0. For Question A10, if the response was one of the three answers representing the highest degree of agreement, the response was mapped to a 1. However, if the response was one of the two answers representing the two lowest degrees of agreement, the response was mapped to a 0. Note that this data mapping was performed by the authors of the dataset, not the authors of this paper (Thabtah 2018).

3.2 Machine Learning Model Implementation

All procedures done in this study were completed using the Python programming language, the Google Colaboratory development environment, and the sklearn machine learning library. Before any of the following procedures were performed, a seed of 7 was set in our programming environment to ensure reproducibility.

To train machine learning models, a training set and testing set had to be developed (Shah 2017). Thus, we programmatically split the dataset such that a random 20% was designated for testing, and the other 80% was designated for training. Whether a patient was put into the testing or training set was randomized.

Before developing any machine learning models, a baseline had to be set to evaluate the ability of our model to fit the data (Li et al. 2020). To set this baseline, an XGBoost classifier was designed, and trained using the training data. The XGBoost classifier achieved an accuracy of 97.04%, which was our baseline used to evaluate the neural network and random forests algorithm.

An XGBoost Classifier works by implementing a simple Decision Tree algorithm and applying extreme gradient boosting to increase the accuracy of the model. On a high-level, a Decision Tree works by evaluating each feature in a dataset using a loss function and choosing the best feature (the feature that minimizes this loss) to split the observations in the training process into different "branches", where new nodes (or "predictors") are created at the ends of these branches for more features to be evaluated. Depending on how a feature is evaluated, a predictor can either split into another predictor or make or decide on a classification for the set of inputs. In XGBoost, each

new predictor in the decision tree is fit on the pseudo-residuals of the previous predictor, but the unique aspect of XGBoost is that it uses second-order gradients and advanced regularization within its boosting framework to achieve greater stability and more accurate classifications (Chen and Guestrin 2016). Please refer to (Chen and Guestrin 2016) for a more detailed explanation on how XGBoost works and how Decision Trees work. The reason we chose XGBoost is because since its proposal, it has become the research standard for baseline model development. We used the default hyperparameters for an XGBoost model provided by the sklearn library.

Once our baseline was acquired, a neural network classifier and random forests classifier was designed. A four-layer perceptron model was first designed, a diagram of which is shown in Figure 1.

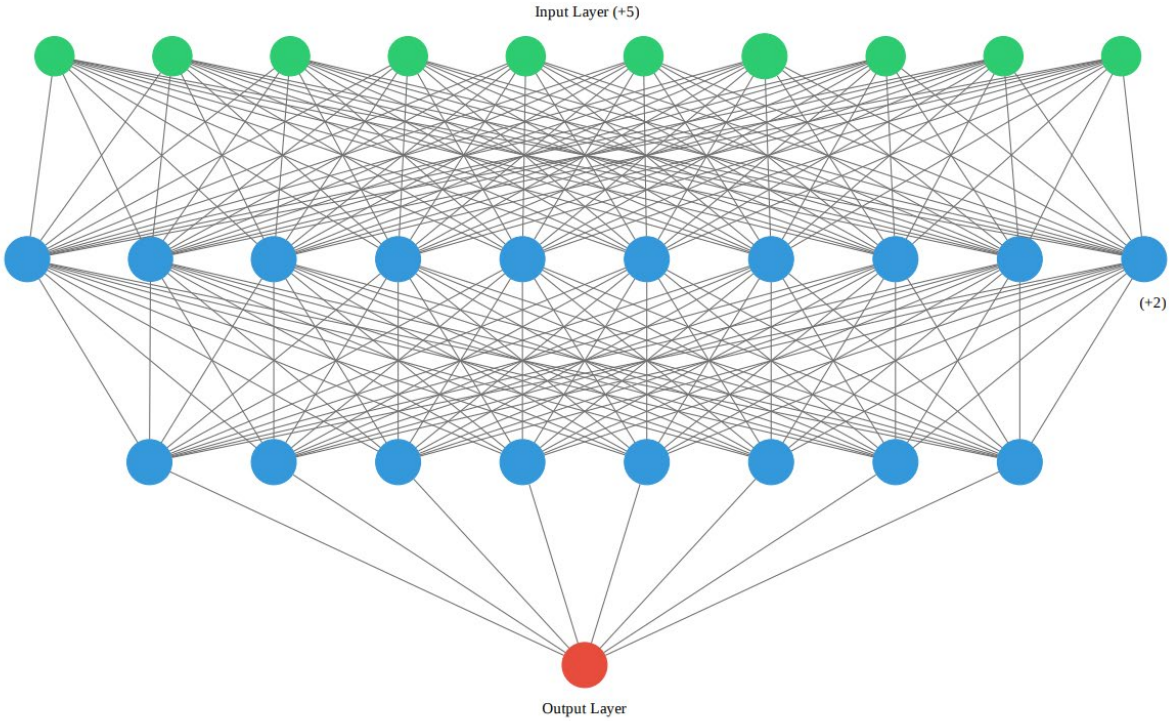

Figure 1. Visual representation of the multilayer perceptron designed in this study

How a multilayer perceptron works is through taking in a set of inputs and multiplying each input by a set certain initially randomized weights (represented by the lines connecting a node from one layer to a node from the next layer). These products are inputs into the next layer of nodes (one product to one node), which is called a hidden layer. Hidden layer nodes transform weighted sum of the products (a.k.a weighed inputs) through a non-linear activation function, such as ReLU or SoftMax. These activation functions decide the threshold at which a node is "activated" (in other words, in a non-linear context activation functions decide how significant an output will be). Such transformations are then multiplied again by the weights connecting the hidden layer to the next layer, and the process continues in the next layer. Finally, each value is inputted into the output layer, in which the output of the node is the output of the model itself (in this context, 0 or 1 depending on whether the model believes a patient has ASD). The output layer almost always uses a step-based activation function to product a value of 0 or 1, which is appropriate for binary classification tasks. In the learning process, the model's output is compared to the ground truth output, and an error is calculated. For binary classification tasks, the error is 0 or 1. This error is then used as an input into a "backpropagation" function that changes the weights of the network architecture to minimize the error. For this type of perceptron, Gradient Descent is almost universally used as a backpropagation function, and thus it was used in our model as well (Hastie et al. 2008).

The first layer was set to take 10 inputs (due to there being 10 used predictor variable values per patient), and the activation function was Rectified Linear Unit (ReLU). The second and third layer were also set to have ReLU activation, but the fourth layer was set with a Sigmoid activation function to produce a classification of 0 or 1 for a patient regarding their clinical autism status. The neural network was compiled using the binary cross entropy loss function (Gradient Descent uses the loss function to operate on the error and calculate how to change weights)

and the Adam optimizer (optimizer functions are used to change the weights). The network was also trained using a batch size of 10 (the number of samples to classify before updating weights) and 150 epoch iterations (an epoch is a full training iteration through a dataset by the neural network). Once the model was trained, it was evaluated using four evaluation metrics discussed in Section 3.1. The results of training and testing this model are discussed in the Results section, and based off these results, it was concluded that no optimization of the neural network was necessary because the network not only outperformed the baseline, but perfectly classified all patients in the test set. We anticipated on using a Grid Search optimization technique to find which combination of hyperparameters was best suited for this classification task, but due to the performance of the initial neural network it was decided that the Grid Search step was unnecessary. The hyperparameters for the model (batch size, no. of epochs, no. of hidden layers, etc.) were chosen through simply estimating what would be appropriate, as they are all relatively standard values for neural networks of this scope, and we anticipated that they would likely change through Grid Search optimization anyways. The reason we chose to use a neural network in the first place is because they are uniquely strong at tolerating noisy and meaningless data, which was beneficial for this task because of the inherent variability and inevitable mistakes in how a certain individual responds to questions on the Autism Quotient 10 test compared to other individuals.

How the Random Forest algorithm works is through designing several decision trees and having the output class be the mode of the classes of the decision trees. This helps prevent the overfitting that is common when decision trees are met with noisy data. Random Forest models use a method called “bagging” where random samples of the training set are used to fit individual trees. Random Forest models also employ a modified learning mechanism where a random subset of features is selected to be considered at each split in the decision tree process to decorrelate the decision trees. By doing this, the phenomenon of the decision tree hierarchy only considering a few highly indicative features is prevented. In addition, we employ a specific type of Random Forest algorithm that uses Gradient Boosting, where each tree is built one at a time instead of each tree being built independently, which allows for weak learners at the top of the decision tree hierarchy to be improved by learners further along the hierarchy. Further, a Gradient Boosted Random Forest model does not find its output by calculating the mode of the outputs of the individual decision trees after they are all built, but it decides the collective output of decision trees along the way, as more trees are being built (Hastie et al. 2008). We decided to use a Gradient Boosted Random Forests model because they are uniquely strong at overcoming imbalanced data and can strongly reduce overfitting compared to other models. In this sense, a Gradient Boosted Random Forests model is very similar to XGBoost, however the difference is that Random Forests builds each tree independently while XGBoost trains one tree at a time (Hastie et al. 2008).

Our Gradient Boosted Random Forest model was designed to use a “deviance” loss function, a learning rate of 0.01, and an estimator count of 1000. These were chosen since they are the default values set by the sklearn library. Then, the model was evaluated again using the four evaluation metrics discussed in Section 3.1. Based on the slight inaccuracy of the model (See Results section), it was decided that optimization via Grid Search would be valuable to perform. Grid Search works by fitting a model for every possible combination of hyperparameters specified and returning the model that has the highest average accuracy when trained/tested using a 5-fold cross validation method as the “optimized classifier”. A diagram of our Grid Search workflow is shown in Figure 2.

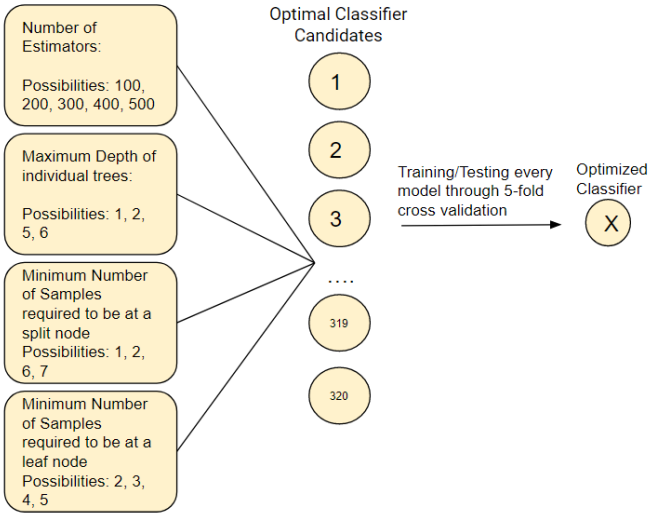

Figure 2. Visual diagram of Grid Search workflow

We decided to optimize four model hyperparameters, as shown in the diagram. Our specified possibilities for each hyperparameter were chosen arbitrarily, which resulted in 320 candidates, each one having a unique combination of the specific hyperparameters. Each one was then trained/tested using a 5-fold cross validation method, and the optimal classifier was returned. The Grid Search was performed using the standard GridSearchCV function in sklearn. Once the Grid Search was complete, the optimal classifier was returned and we re-evaluated it using the metrics described in Section 3.1 to compare it against the non-optimized model. The results of training and evaluating this optimal classifier are discussed in the Results section.

3.3 Feature Selection

This section will describe the methodology used to identify which features were most significant in giving rise to autism. On a high level, the process consisted of developing a hyperparameter optimization function, developing a shuffled K-Fold function, using these functions to develop and optimize a LightGBM Regressor model, and extracting the feature importance scores calculated by the model function.

First, a hyperparameter optimization function was developed that used LightGBM parameter optimization to calculate the optimal values for all hyperparameters for a LightGBM Regressor model. Then, a Shuffled K-Fold Cross Validation function was developed that took the calculated hyperparameters and trained/tested two LightGBM Regressors over 5 folds (one regressor to track importance split scores of each feature, the other to track importance gain scores of each feature) to calculate Out-Of Fold Mean Absolute Error scores, importance split scores (the importance of a split in the Decision Tree hierarchy), and importance gain scores (the importance of the gain made at a split) for each feature. Note that for each fold in the Shuffled K-Fold function, the importance split/gain scores calculated at each fold were aggregated to achieve a total importance split/gain score. We did this to reduce the statistical bias associated with only using one fold (in other word, only one combination of training/testing data). Also note that each LightGBM Regressor model was initialized using the calculated hyperparameters. The specific way in which these functions were used to acquire the quantitative significance of each predictor variable is discussed in Section 4.2.

Note that a LightGBM Regressor is simply a Gradient Boosted Decision Tree, where Decision Trees are produced and fit to correct the errors made by a previous Decision Tree, thereby creating an ensemble of Decision Trees. Each Decision Tree is fit using a gradient descent loss optimization algorithm. A LightGBM Regressor is very similar to XGBoost, however the key difference is that the LightGBM Regressor performs regression instead of classification and that the LightGBM Regressor employs leaf-wise tree growth whereas XGBoost employs level-wise tree growth. Thus, XGBoost would finish expanding each node in a certain part of the Decision Tree hierarchy, whereas LightGBM would not wait for each node in that part to expand before expanding nodes in a subsequent level of the hierarchy, thereby reaching maximum depth (maximum number of levels allowed in the Decision Tree hierarchy) in a different manner (Ke et al. 2017). Please refer to (Ke et al. 2017) for a more in-depth and technical explanation. The reason we used a LightGBM Regressor in this case was so that we could employ the LightGBM parameter optimization algorithm.

4. Data Collection

This section will describe how the performance of each model was calculated, and how the results data were collected using the developed feature selection methods.

4.1 Machine Learning Model Evaluation

First, each trained model was used to predict the classes of all patients in the testing set. Then, these predictions were compared to the ground truth classes of all patients in the testing set using the following four metrics.

The first metric is a Recall Score. A Recall Score is the ratio of the number of true positives to the sum of the true positives and false negatives. Therefore, a Recall Score represents the ability of the model to identify positive samples as positive (“positive” referring to patients who have autism, “negative” referring to patients who do not). The next metric is a Precision Score. A Precision Score is the ratio of the number of the true positives to the sum of the true positives and false positives. Therefore, a Precision Score is slightly different from a Recall Score, as the Precision Score more closely represents the ability of the model to not classify a negative patient as positive. The third metric is an F-1 Score, which is the ratio of the product of the precision and recall to the sum of the precision and recall, and this ratio is multiplied by two. Therefore, an F-1 Score represents the weighted average of the precision and recall. Lastly, an Accuracy Score was used, which is simply the ratio of the number of correct

classifications to the number of total classifications performed (Exsilio Solutions 2016). The reason that all four metrics were used is because they collectively give a comprehensive look at the model's performance - for all four metrics, the most ideal value is 1, and the least ideal is 0.

4.2 Feature Selection Results Calculations

To calculate the significance of each predictor variable, the first function described in Section 3.3 was first used to run 150 rounds of LightGBM parameter optimization, which returned the hyperparameters identified as creating the optimal LightGBM Regressor for this problem. Then, the second function described was used with these calculated hyperparameters as an argument to calculate the three scores mention in Section 3.3. These scores were then used to calculate an overall importance score for each feature - the importance score was equated to the natural logarithm of the product of the importance split and importance gain scores. Thus, for each feature, an importance score was calculated, with a higher value corresponding to a more significant feature. The features were then ranked from highest significance to lowest, thereby returning a list of the most important to least significant feature and their respective relative importance score.

5. Results and Discussion

5.1 Numerical Results

This section will present the numerical results gathered throughout the course of this study and the appropriate inferences we drew. Firstly, Table 2 presents the performance of the baseline XGBoost model, relative to which the neural network and random forest models were evaluated.

Table 2. Performance of Baseline XGBoost Model

XGBoost performance characteristics	Value
Mean accuracy	97.04%
Standard Deviation	1.78%

It was inferred from Table 2 that the quantitative pattern between the predictor variables and whether a subject toddler had ASD was very clear, since the baseline XGBoost model performed very well. It was also inferred from this data that the neural network and random forests model would also perform very well, since the high performance of the baseline model suggests that state-of-the-art algorithms such as neural networks and random forests should also perform well.

Table 3 presents the performance of the neural network and pre-optimized/optimized random forests model. Note that the neural network model was not optimized because, as seen in Table 3, it achieved a perfect validation fit without optimization.

Table 3. Evaluation Metric Performances of Each Model on Testing Set Patients

	Recall Score	Precision Score	F1 Score	Accuracy
Neural Network	100%	100%	100%	100%
Random Forest Classifier (Pre-Optimization)	98.10%	98.15%	98.09%	98.10%
Random Forest Classifier (Post-Optimization)	100%	100%	100%	100%

It is evident from Table 3 that the neural network model performed slightly better than the random forest classifier pre-optimization, although both performed very well. It is also evident that the neural network model performed perfectly on the testing set, which is not surprising considering the very high accuracy of the baseline XGBoost model. The optimization of the random forest model using Grid Search increased accuracy marginally, although there was not much increase that could be done due to the already very high accuracy of the random forest classifier pre-optimization. After optimization, the random forest model achieved a perfect performance on the testing set.

From this part of our research, we fulfilled the first research objective by developing two highly accurate machine learning models to predict the presence of ASD in toddlers using only physical characteristics. We experimented with two types of models and a hyperparameter optimization technique in order to increase the likelihood that a highly accurate classifier was achieved, but it is evident from the results that this was not entirely necessary because of the already high performance of both models.

Table 4 presents the most significant findings in this study, which are the most significant features in giving rise to autism. Table 4 presents this information as a table of the calculated importance scores of the 10 most important features.

Table 4. Calculated Importance Scores of 10 Most Important Features

Feature	Importance Score
A4	7.828181
A6	7.821073
A9	7.818744
A1	7.679076
A7	7.266456
A5	7.108156
A2	6.830066
A8	6.414015
A3	5.133795
A10	4.908389

It is evident from Table 4 that out of the most Qchat 10 test that the most ASD-indicative questions are: “Does your child point to share interest with you? (e.g., pointing at an interesting sight)”, “Does your child follow where you’re looking?”, and “Does your child use simple gestures? (e.g., wave goodbye)”. This inference was reached because these three questions were calculated to have the highest importance scores, all averaging ~7.82. It is evident that the least ASD-indicative questions in the Qchat 10 test was “Does your child point to indicate that s/he wants something? (e.g., a toy that is out of reach)” and “Does your child stare at nothing with no apparent purpose?” as they had the lowest importance scores out of the 10 questions, with scores of 5.133795 and 4.908389, respectively. It is also noted from the table that the 10 most importance predictor variables are also the 10 questions in the Qchat 10 test – not one of the variables that was not one of the Qchat 10 questions was also one of the 10 most significant predictor variables. This suggests that only behavioral characteristics and not inherent physical characteristics are the strongest indicators of ASD. This phenomenon is discussed further in Section 5.2.

5.2 Graphical Results

Figure 3 presents the importance scores of each predictor variable in a comparative bar graph format. The y-axis is the predictor variable’s designated code, and the x-axis is the importance score of that predictor variable.

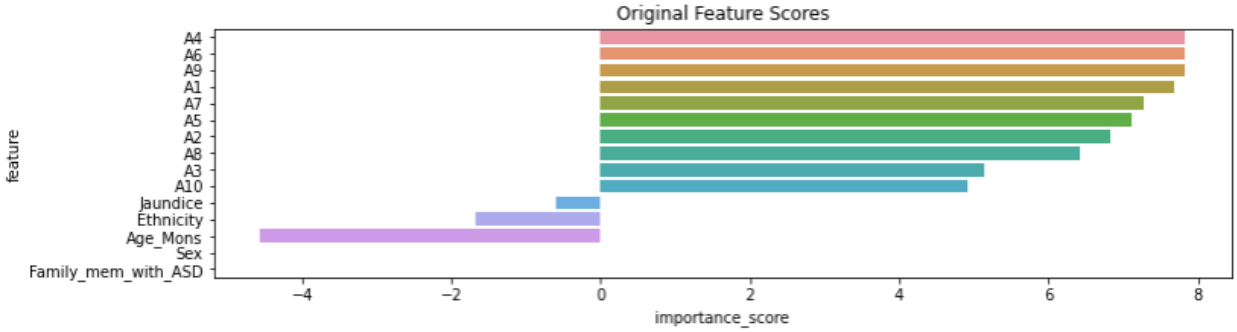

Figure 3. Bar graph representation of the calculated importance scores of each predictor variable (feature) included in training/testing models

Figure 3 not only reiterates the inferences drawn from Table 4 in Section 5.1, but it leads to new inferences regarding the importance of features that were not Qchat 10 questions. It is evident from Figure 3 that all the

features that were not Qchat 10 questions had a negative importance score, which means that the product of the importance split and importance gain scores must have been less than 1, because only in that condition would the natural log of this product (a.k.a. the overall importance score) be negative. This suggests that those features (the presence of Jaundice, ethnicity, and age) have a statistically insignificant role in indicating ASD. It is also evident from Figure 3 that the importance scores of the Sex and Family_mem_with_ASD features had an importance score of 0, which means that the product of the importance split and importance gain scores must have been 1, because only in that condition would the natural logarithm of this product be 0. This suggests that the Sex and Family_mem_with_ASD features are slightly more significant than the Jaundice, Ethnicity, and Age_mons features, but are still much less significant than any of the Qchat 10 questions. Therefore, the statistical significance of the Sex and Family_mem_with_ASD features may not be negligible but is still very small.

From these results it can be inferred that in order to predict the presence of ASD in toddlers, the Qchat 10 questions are the most indicative in the order described in Table 4, the Sex and Family_mem_with_ASD features are much less indicative but may not be negligible, and the Jaundice, Ethnicity, and Age_mons features are very likely statistically insignificant. From the work leading to Table 4 and Figure 3, we fulfilled our second research objective by quantitatively identifying which physical characteristics are the most significant indicators of ASD in toddlers.

5.3 Proposed Improvements

There are several directions for improvement that can be taken in future work. Firstly, clinical testing can be performed using machine learning models to further validate the real-world efficacy of machine learning in detecting autism in toddlers. Further, other machine learning models could be experimented with by only using the most important features, thereby elucidating whether accurate classifiers can be achieved with less data. In addition, other methods of feature selection can be used such as the Fisher's score or Pearson's Correlation Coefficient in order to validate our results. In addition, further research into exactly why certain questions is more indicative of autism than others can be done, to reveal potentially unknown characteristics of autism in general. Thus, this work is significant in facilitating further research and improvement in the field of autism study.

6. Conclusion

In this study, novel research is performed regarding the detection of autism in toddlers. First, machine learning models were designed and optimized to achieve high performance in being able to detect autism in toddlers using physical, non-invasive data. Further, computational methods are used to quantitatively determine the importance of each physical feature.

There are several notable takeaways from this work. First, it was observed that the standard neural network and optimized random forests model both achieved perfect fits on the testing set. This led us to conclude that standard machine learning models are sufficient to act as a reliable clinical decision justification system in the diagnosis of autism in toddlers. The most significant aspect of our work and the aspect that is a unique research contribution is the quantitative evaluation of the importance of the physical features used to detect autism. As discussed in Sections 5.1 and 5.2, we quantitatively identified the most to least significant features and drew inferences regarding the relation between physical characteristics and the presence of ASD.

This information is significant because it is useful for clinicians to properly analyze results in order to determine whether a child has autism. Knowing, for example, that question A4 is significantly more indicative of ASD than question A10 or that the presence of Jaundice in a child has almost no statistical indication of ASD will be useful for helping a clinician make a more accurate and informed diagnosis for a child. Therefore, highly useful information for pediatricians that make autism diagnoses is uncovered in this study.

Through or methodology, data collection, and analysis, we fully achieved both of our research objectives. We developed two highly accurate machine learning models to detect ASD in toddlers using only physical characteristics, and we quantitatively identified the physical characteristics that are the most significant indicators of ASD in toddlers.

References

Alteneiji, M., Alqaydi, L., and Tariq, M., Autism Spectrum Disorder Diagnosis using Optimal Machine Learning Methods, *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, 2020.

Booth, T., Murray, A., McKenzie, K., Kuenssberg, R., O'Donnell, M., and Burnett, H., Brief Report: An Evaluation of the AQ-10 as a Brief Screening Instrument for ASD in Adults, *Journal of Autism and Developmental Disorders*, vol. 43, pp. 2997-3000, 2013.

Chen, T., and Guestrin, C., XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
Chowdhury, K., and Iraj, M., Predicting Autism Spectrum Disorder Using Machine Learning Classifiers, *Proceedings of the 2020 International Conference on Recent Trends in Electronics, Information, Communication & Technology*, pp. 324-327, 2020.
Exsilio Solutions, Accuracy, Precision, Recall, & F1 Score: Interpretation of Performance Measures, Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/, September 9, 2016.
Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, 2nd Edition, Springer, 2008.
Hossain, D., and Kabir, M., Detecting Child Autism Using Classification Techniques, *Studies in Health Technology and Informatics*, vol. 264, pp. 1447-1448, 2019.
Jia, R., Steelman, Z., and Jia, H., Psychometric Assessments of Three Self-Report Autism Scales (AQ, RBQ-2A, and SQ) for General Adult Populations, *Journal of Autism and Development Disorders*, vol. 49, pp. 1949-1965, 2019.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017.
Li, D., Hasanaj, E., and Li, S., 3 - Baselines, Available: blog.ml.cmu.edu/2020/08/31/3-baselines/, August 31, 2020.
Li, G., Liu, M., Sun, Q., Shen, D., and Wang, L., Early Diagnosis of Autism Disease by Multi-channel CNNs, *Machine Learning in Medical Imaging*, vol. 11046, pp. 303-309, 2018.
Lingyu, X., Geng, X., He, X., Jun, L., and Jie, Y., Prediction in Autism by Deep Learning Short-Time Spontaneous Hemodynamic Fluctuations, *Frontiers in Neuroscience*, vol. 13, pp. 1120, 2019.
Moore, V., and Goodson, S., How well does early diagnosis of autism stand the test of time? Follow-up study of children assessed for autism at age 2 and development of an early diagnostic service, *Autism: The International Journal of Research and Practice*, vol. 7, no. 1, pp. 47-63, 2003.
Omar, K., Mondal, P., Khan, N., Rizvi, R., and Islam, N., A Machine Learning Approach to Predict Autism Spectrum Disorder, *Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering*, pp. 1-6, 2019.
Raj, S., and Masood, S., Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques, *Procedia Computer Science*, vol. 167, pp. 994-1004, 2020.
Shah, T., About Train, Validation, and Test Sets in Machine Learning, Available: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7, December 6, 2017.
Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U., Khosrowabadi, R., and Salari, V., Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network, *Frontiers in Neuroscience*, vol. 13, pp. 1325, 2020.
Skibitsky, J., The Importance of Early Diagnosis and Intensive Therapy for Children with Autism, Available: https://alternativebehaviorstrategies.com/the-importance-of-early-diagnosis-and-intensive-therapy-for-children-with-autism/, January 10, 2019.
Taylor, E., Livingston, L., Clutterbuck, R., and Shah, P., Psychometric concerns with the 10-item Autism-Spectrum Quotient (AQ10) as a measure of trait autism in the general population, *Experimental Results*, vol. 1, 2020.
Thabtah, F., Abdelhamid, N., and Peebles, D., A Machine Learning Autism Classification Based on Logistic Regression Analysis, *Health Information Science and Systems*, vol. 7, no. 1, pp. 12, 2019.
Thabtah, F., Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment, *Proceedings of the 1st International Conference on Medical and Health Informatics*, New York, USA, May, 2017.
Thabtah, F., Autism Screening Data for Toddlers, Available: https://www.kaggle.com/fabdelja/autism-screening-for-toddlers/version/1, July, 2018.
The Autism Society, Facts and Statistics, Available: https://www.autism-society.org/what-is/facts-and-statistics/, August 26, 2015.
Varshini, D., and Chinnaiyan, R., Optimized Machine Learning Classification Approaches for Prediction of Autism Spectrum Disorder, *Annals of Autism and Developmental Disorders*, vol. 1, no. 1, 2020.

Biographies

Arjun Singh is a junior in high school from Warren, New Jersey, who is interested in combining the fields of computer science and biology to solve challenging problems in the current healthcare industry. Arjun has leadership positions at multiple clubs at his school, including Computer Science Club and the Science Bowl club. Arjun also takes part

heavily in entrepreneurship, serving on the Board of Directors at two student-started nonprofit companies and starting his own international online tutoring firm in March 2020. Arjun also has lots of experience in scientific research, pursuing professional research projects focused on machine learning with professors from Kean University, a PhD student at Cornell University, and more. Arjun has also taken initiative of his scientific research and has completed several research projects independently, winning him several awards at regional and national science fairs as well as a student membership to the American Association of Cancer Research. At ARQuest SSERN, Arjun serves as the Head of Research, leading two research teams on machine learning projects involving problems in healthcare and setting up the research internship process that connects like-minded students from around the world. Lastly, Arjun has recently found an interest in the use of machine learning for the environment, creating two apps in his sophomore and junior year of high school that won him recognition from the Spellman High Voltage Electronics Corporation and the United States Congress. In his free time, Arjun likes to play tennis, watch movies, and drive with friends.

Branden Sattler is a junior in high school from Long Island, New York, who has a passion for drawing, playing tennis, exploring mental health topics, and helping others. To dive into his interest in art, Branden has self-taught himself how to draw and even started his own art gallery on Instagram with a following of over 900 people. His interest in art has also resulted in him becoming an artist of a non-profit that spreads awareness and talk politics. Branden's love of tennis gave rise to him becoming a varsity tennis player as a freshman at East Meadow High School. In addition, Branden's love for helping others motivated him to volunteer at a local nursing home. While at the local nursing home, he witnessed people with AD and dementia. As a result of this experience, Branden developed an academic interest in psychology and mental health. This new academic interest has given rise to his role as the Head of the Blog Department at ARQuest Student Science and Research Network. Besides overseeing the blog department, Branden has written about mental and psychology in these blogs, which get posted on our official website. Additionally, Branden completed a research internship in which he wrote about predicting Autism cases in toddlers. Branden plans to continue to explore and research mental health and neurobiological related topics. Other interests that Branden has are his love for spending time with his family, listening to music, and exploring his very mixed cultural roots.

Michael Helde is a high school junior in Washington, USA. He has a rigorous course load consisting of only college-level courses which have piqued his interest in the following subjects: mathematics, biology, computer science, and physics. To further pursue his interests, Michael has taken maxed out the math courses he can take at his local community college. Also, he has taken part in research with the assistance of a WSU Professor on computational neuroscience which is a multidisciplinary research topic that has allowed him to pursue all his interests, as it is inclusive of all the topics. In addition, at the ARQuest Student Science and Research Network, Michael worked in tangent with other motivated students to form a research group and writes monthly blogs for the ARQuest SSERN website as another way to pursue his interests. Outside of his STEM interests, however, Michael has a leadership position as the President of his high school's Key Club where he gives back to his community through volunteer service. With the advent of COVID-19, Michael has created an online tutoring service to help students struggling to keep up with online learning. In his free time, Michael enjoys going on hikes in the Pacific Northwest, investing, and spending time with his family.

Zoya Farooqui is a junior in high school from Texas, who always finds ways to take initiative and help others. She loves programming and robotics, spending time each week competing with her robotics team and teaching young kids how to code. Some of her other hobbies are art and graphic design, which recently got her into book illustrations. Zoya has illustrated the cover for a book published on Amazon and is currently working on creating the graphics for a children's storybook. Moreover, she is the co-host of her own podcast, which has provided her with the opportunity to empower young minds with leadership skills through her own experiences. As the Head of the Technology Department at ARQuest Student Science and Research Network, she became involved in research by completing an internship under them, discovering that her research interests include computer science, technology, and physics, and she is looking forward to diving further into these passions in the future. Other than these interests, Zoya also enjoys spending time with her family and friends, playing video games, and travelling.

Unyimeabasi Usua is a sophomore in high school from Houston, Texas. She is engaged in her school's Science Olympiad team, Computer Science Club and UIL Science team. Currently, she works as a student researcher at the Green Bank Laboratory where she is analyzing the data and writing a paper on magnetars. Unyimeabasi is passionate about the fields of science and technology. She takes advanced science and computer courses at her school. In her free time, she learns python and works on coding projects with her friends.