An Automated Machine Learning Pipeline for Predicting Autism in Adults using EEG Data

Michelle Barboure
m.c.barboure@student.vu.nl
Vrije Universiteit Amsterdam

ABSTRACT

The current diagnosis of autism spectrum disorder (ASD) relies solely on behavioural observations, leaving room for subjectivity and misdiagnosis. Studies have begun to address this issue by predicting ASD from brain imaging data using machine learning. However, few have focused their efforts in adults. This study aimed to determine whether electroencephalography (EEG) data could be used to train machine learning models that could accurately classify ASD individuals from a cohort of Dutch adults performing two visual tasks. We used a highly-automated preprocessing and feature engineering pipeline to limit biases in the methodology, then used grid search to select the highest performing models from a collection of five well-established learning algorithms. Of the two visual tasks, the boundary detection (BD) task resulted in the higher classification accuracy. By modifying the learning algorithms to include an intermediate regression step, our study was able to achieve higher mean classification performance than using traditional classification. It also provided a variable tolerance for sensitivity and specificity by tweaking the threshold parameter – making this technique relevant for clinical applications. These enhancements were possible because both ASD diagnosis and Autism Quotient-Short scores were available as target features in the dataset. Tree-based learning algorithms obtained the highest performance, with the decision tree achieving a highest mean accuracy of 83%. The most robust model was a random forest ensemble with 7 features and an area under the ROC curve of 0.887. Overall, this study provides strong evidence that ASD can be predicted with considerable accuracy in adults using EEG data. The paper also details a flexible and reproducible pipeline that can be used in future work for developing usable models.

KEYWORDS

Autism, Electroencephalography, Autism-Quotient Short score, Supervised learning

1 INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder associated with a range of phenotypes that vary in severity of social, communicative and sensorimotor deficits [1]. This variability, together with its unknown etiology [2], makes ASD notoriously difficult to diagnose. Currently, psychiatrists diagnose ASD based solely on behavioural observations, using the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) as their guide. This diagnostic process leaves room for subjectivity and ultimately carries the risk of individuals being misdiagnosed [3]. No quantitative tests currently exist to resolve this issue.

However, it is well-recognised that ASD is associated with abnormal patterns of connectivity in different regions of the brain [4, 5]. The early development of white matter tracts – the interconnected highways that link different parts of the brain – suggest that abnormal connectivity could be amongst the earliest markers of ASD, with initial signs emerging within the first year of life [6, 7]. Brain imaging techniques, particularly functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), have been able to discern these differences in functional connectivity, making them a promising tool for early ASD diagnosis.

However, the key to the progress in ASD diagnosis has not been improvements to brain imaging techniques themselves, but rather advancements in the machine learning methods that model the resulting data. The implication is that the complex time series data from brain imaging recordings contains information about the neural network structure, and that machine learning models are able to make sense of it [8, 9]. Fittingly, the use of machine learning in mental health research has grown rapidly in recent years, with progress evident in ASD, but also depression, schizophrenia, and Alzheimer's disease [10].

So far, studies using fMRI data for ASD diagnosis have been promising: A recent study was able to classify subjects with ASD from healthy controls with a maximum accuracy of 82% [11]. A hybrid deep learning approach was implemented, combining an autoencoder with a single-layer perceptron. Neural networks are becoming a very popular approach for diagnosing ASD, and many studies have used autoencoders, long short term memory, and convolutional neural networks to do so [12–16]. Others have relied on more traditional supervised approaches, using machine learning techniques such as support vector machines (SVMs) and random forest (RF) ensembles to classify ASD [17, 18]. Although all approaches vary in their level of subjectivity and computational cost, all rely on fMRI data as input – an expensive and time-consuming technique not well suited for clinical use.

By contrast, EEG is an ideal alternative for clinical settings, since it is quick, easy to use, and inexpensive [19]. Subsequently, researchers have increasingly explored EEG recordings as a viable tool for ASD diagnosis. However, no universal preprocessing standards currently exist for removing eye and muscle movement artifacts, with cleaning protocols varying from study to study [20]. Moreover, studies have engineered a variety of features from preprocessed EEG data – ranging from entropy measures, wavelet and Fourier transforms, and various statistical measures. Together, this makes cross-study comparisons difficult and has resulted in vastly different approaches across studies.

VU Masters Biomedical Sciences, 2020, AM_178

For instance, Bosl et al. [21] decomposed resting state EEG recordings from 188 infants across varying ages into frequency bands using Daubechies (DB4) wavelets. From these, nine nonlinear invariant features were computed across 19 scalp sensors – multiscale entropy analysis, detrended fluctuation analysis, and several recurrence quantitative analyses providing the bulk of the features. An SVM model achieved high specificity, sensitivity, and positive predictive values, exceeding 95% in some age groups.

A study by Haputhanthri et al. [22] used a similar approach, decomposing resting state EEG data from 15 children into the beta frequency band (16–32 Hz) using discrete wavelet transforms. The mean and standard deviation were extracted from the raw and transformed data, and further sub-selected based on a correlation-based feature selection algorithm. A primary focus of their study was to classify ASD with as few sensors as possible to simplify the downstream real-world implementation and increase its affordability. A random forest model using only five sensors yielded the best result with an accuracy of 93%.

Pham et al. [23] implemented a novel approach. They converted the resting state EEG signals from 77 children to two-dimensional images using the higher-order spectra bispectrum. Various nonlinear features were extracted, then reduced using locality sensitivity discriminant analysis, and further condensed using Student's t-tests. A probabilistic neural network classifier was used and achieved an accuracy of 98.7% with only five features.

The above is by no means an exhaustive list of studies that have used EEG data to classify ASD. However, they clearly illustrate the range of feature extraction and machine learning approaches used in this domain. They also highlight the fact that the majority of the field focuses on ASD classification in children rather than adults. Indeed, Pham et al. [23] provide a comprehensive, up-to-date list of studies that substantiate this. Focusing on children is entirely understandable, as an early diagnosis would assist early interventions, which in turn may increase the child's response to treatments. Furthermore, an emerging view is that the behavioural symptoms displayed by ASD individuals may be the end result of early brain-wide adaptations, rather than the direct consequence of ongoing neural pathology [24]. This would suggest that the initial neural abnormalities that lead to ASD may be transitory, and thus difficult to detect at later stages in life [25].

If the neural abnormalities in ASD are indeed transitory, then the question remains whether neuronal adaptations can be detected using EEG in adults. Snijder et al. [26] showed that adults with ASD showed a reduced steady-state gamma response to contextual modulation in the visual domain compared to controls. In other words, when participants were presented with a visual task, a smaller evoked response in the gamma frequency range (30–90 Hz) was detected over parieto-occipital sensors in those with ASD compared to controls. These results are in line with previous studies that suggest atypical visual perception in ASD is related to impaired lateral connectivity within primary visual areas of the brain [27, 28]. All these results indicate that abnormal connectivity patterns may be evident in the parietal-occipital region in adults with ASD – a promising insight that may be captured by EEG and applied to machine learning models for ASD diagnosis.

Therefore, the goal of this study was to determine whether visual-task-based EEG data from parieto-occipital sensors can successfully classify ASD in a cohort of adults. Our visual tasks were similar to those in Snijder et al. [26], using Gabor patches as stimuli known to drive early visual activity in a controlled fashion. We investigated two learning types, approaching the problem as a binary classification and an intermediate regression task. The intermediate regression model took on an unusual approach by using AQ-Short scores as an intermediate prediction target. AQ-Short is an abridged version of the established 50-item Autism-Spectrum Quotient (AQ), a self-report questionnaire that assesses autistic traits in individuals with normal intelligence [29]. AQ-Short is made up of only 28 questions and has been confirmed to retain its validity and factor structure with a recommended cut-off above 65 to classify ASD [30]. As Bosl et al. [21] note, "because ASD occurs along a spectrum, training a classifier to make a binary decision (ASD or not) with subjects that have essentially the full range of ASD characteristics leads to problems with the very definition of ASD." The intermediate regression step – targeting AQ-Short scores – takes this into account.

This paper is structured as follows: Section 2 describes the dataset, the visual tasks, the EEG recording procedure, the preprocessing workflow, the learning algorithms, and the experimental design. Section 3 details the results obtained on different combinations of visual task data and learning types. Section 4 discusses the implications of our findings. Finally, we outline limitations of this study and our conclusions in Sections 5 and 6 respectively.

2 METHODOLOGY

2.1 Participants

This study forms part of the larger ongoing project *BECAUSE (from Behaviour to Cell in Autism Sensory Processing)*, which recruited participants with ASD from the Netherlands Autism Registry (NAR) database (www.nederlandsautismeregister.nl). From all NAR members, only members aged between 18 and 55 years, and with a self-reported intelligence equal or higher than average, were invited. The ASD diagnosis was confirmed using the Developmental, Dimensional and Diagnostic Interview (3di), an interview assessment tool for ASD diagnosis [31]. The interview was performed with a person that knew the participant since childhood, such as a parent, a (previous) caregiver, or a sibling.

In total, 100 ASD subjects and 92 control subjects were recruited. Of these, 75 subjects were excluded (25 ASD, 50 controls) due to at least one of the following reasons: the visual task was not attempted, the visual task was not fully completed, a visual impairment (e.g. lazy eye) was reported, construction noise was present during recording, the task was recorded with a broken sensor, a sensor replacement was done prior to recording, or no AQ-Short score was available. A further 33 ASD subjects were excluded to balance the dataset. This was necessary because class imbalances can be exploited by learning algorithms, resulting in low-quality models – a problem which is exacerbated when their are only 2 classes. Table 1 presents summary statistics for the included subjects' characteristics.

The experiment was approved by the medical ethics committee of the Vrije Universiteit Amsterdam, and all subjects gave written informed consent before participation.

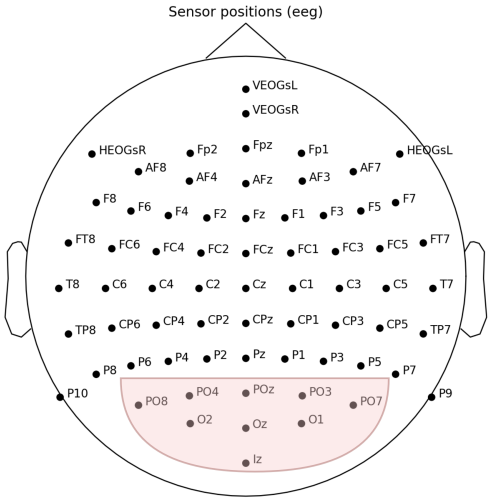

Figure 1: Position of the sensors according to the 10-20 international system. Sensors used for this study are highlighted in red.

2.2 EEG recording

Recordings were conducted at the Vrije Universiteit Amsterdam, where each session took approximately 3.5 hours to complete. After a brief hearing test, participants were seated in a dimly-lit, soundproof booth for recording. The computer screen was positioned 80 cm from their eyes, so that 1° of visual angle corresponded to 1 cm on the screen. In total, five sensory tasks and two 5-minute resting-state conditions were completed – two of which were the focus of this study. EEG activity was recorded with a *Biosemi 64-channel Active Two EEG system* (Biosemi Instrumentation BV, Amsterdam, the Netherlands), and sampled at 2048 Hz. Two sensors in the electrode cap provided an active ground. Electrooculogram (EOG) data was recorded from sensors above and below the eye, as well as at the outer canthi of the eyes. The subset of sensors chosen for our analysis were positioned in the region of the occipital lobe, namely *PO3*, *PO4*, *PO7*, *PO8*, *POz*, *O1*, *O2*, *Oz*, and *Iz* as per the standard 10-20 montage (see Figure 1).

Table 1: Subject characteristics of control and Autism Spectrum Disorder (ASD) groups used in model training and validation. Means and standard deviations are given. AQ = autism quotient; F = female; M = male

	Control group	ASD group
Sample size	42	42
Gender	25 F, 17 M	25 F, 17 M
Age	36.3 (13.1)	44.1 (8.3)
AQ-Short	52.6 (9.8)	84.0 (11.1)

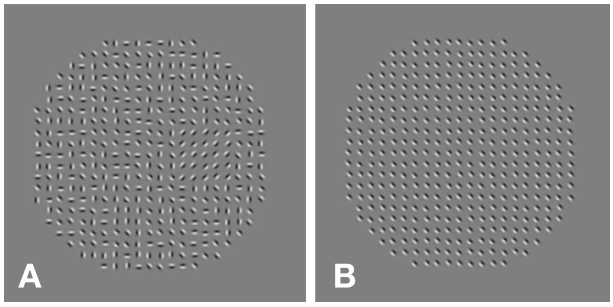

Figure 2: Example visual stimuli used in this study. (A) shows a stimulus from the boundary detection task. The target (square is positioned in the right side with 100% homogeneity, the surrounding circle has a homogeneity of 25%. (B) shows a stimulus from the orientation discrimination task. Gabor patches are homogeneously aligned (100%) at an angle of 45°.

2.3 Visual stimuli

Visual stimuli were generated using *OpenSesame*, a Python-based stimulus presentation program [32]. Small, tightly-arranged Gabor patches – sine wave gratings seen through a Gaussian window – were arranged in a circle on a gray background, and their directions were manipulated to suit each task. Participant responses were recorded with a standard keyboard and mouse.

2.3.1 Boundary detection (BD) task. Participants were asked to identify a square within a circle consisting of Gabor patches. The square – also made up of Gabor patches – was either more or less homogeneously-orientated compared to the circle surrounding it. Homogeneity of the square varied (100%, 80%, 60%, 40%, 20% or 0%), as well as that of the circle surrounding it (75% or 25%). An example is shown in Figure 2. The stimulus was presented for 500 ms, after which the participant was required to indicate (with a mouse click) whether the square was located to the right or to the left of the circle centre. Twenty-four different stimulus combinations (6 square homogeneities x 2 circle homogeneities x 2 target locations) were presented 16 times each, resulting in a total of 384 trials. Assuming the participant clicked after each stimulus presentation, that resulted in a total of 768 events per subject. The time between the participant's response and the next stimulus was varied randomly between 800 and 1200 ms.

2.3.2 Orientation discrimination (OD) task. Participants were presented with Gabor patches with 100% homogeneous orientation but with varying angles of orientation. The stimulus was presented for 1000 ms, after which the participant needed to decide whether the direction of the Gabor patches in the current stimulus presentation was turned clockwise or counterclockwise compared to the previous presentation. The first stimulus was presented at an angle of 45 degrees, followed up with rotational degrees of 2, 4, 6, 10, or 14 degrees clockwise or counterclockwise. An example is shown in Figure 2. This introduced 10 different stimulus combinations, each of which was presented 36 times, resulting in a total of 360 trials. Assuming the participant clicked after each stimulus presentation, this resulted in a total of 720 events per subject. Once again,

the time between the response and the next stimulus was varied randomly between 800 and 1200 ms.

2.4 EEG preprocessing

The data was preprocessed with *MNE-Python* version `0.19.2`, an open-source software library [33]. The preprocessing steps described below were deliberately chosen for their principled or automated characteristics so as to reduce bias and enhance the reproducibility of our cleaning protocol.

The raw data was downsampled to 256 Hz to reduce processing time. A 50 Hz notch filter was applied to suppress power-line interference, and a bandpass filter between 1 and 70 Hz was used to remove low-frequency drifts and subselect to brainwave frequencies of interest.

The filtered data was epoched into varying lengths depending on the task. BD task data was segmented into 2000 ms epochs ranging between 500 ms before and 1500 ms after stimulus onset. OD task data was segmented into 3000 ms epochs with 2500 ms after stimulus onset. Next, version `0.2.1` of the *Autoreject* Python library [34] was used to automatically reject bad trials and repair bad sensors. The epoched data was then re-referenced to the average of all EEG electrodes, since a high coverage was achieved with 64 electrodes and over 50% of the head surface was covered.

EOG artifacts were removed by first detecting and subsequently removing EOG related components. Detection was based on Pearson correlation between the filtered data and the filtered EOG sensors (*HEOGsR*, *HEOGsL*, *VEOGsR*, *VEOGsL*), where thresholding was based on adaptive z-scoring. The components were computed using the preconditioned ICA for real data (PICARD) algorithm. This independent component analysis (ICA) algorithm has shown fast convergence on real data and is known to be more robust than other algorithms in cases where the sources are not completely independent [35]. Finally, the *Autoreject* random sample consensus (RANSAC) method was applied to detect and interpolate any remaining bad epochs.

2.5 Dataset

When considering the data at an epoch level, a vast amount of training data was available. The dataset contained over 500,000 measurements per sensor for each participant, since each of the task recordings consisted of around 750 events, each with more than 500 recordings.

Initially, we hypothesised that the models would benefit from learning from these thousands of samples per individual. This would have reflected the current trends in contemporary applications of machine learning [36]. However, our preliminary testing indicated that prediction accuracies remained only marginally above chance. We concluded that too much noise was present at the level of individual epochs in EEG data, obscuring any possible signal and inhibiting the models from learning the underlying patterns. To overcome this, signal averaging was performed. Unfortunately, this dramatically reduced the number of available data points and made "deep" algorithms like multi-layer neural networks infeasible.

2.5.1 Signal averaging. Event-related potentials (ERPs) are direct brain responses triggered by sensory, cognitive, or motor stimuli. These ERPs contain much of the useful signal that we wished to model. However, these ERPs are small in relation to the brain's ongoing electrical activity, making them difficult to detect. The easiest technique currently used to detect ERPs is signal averaging, which is based on calculating the average waveform over all ERPs, with the assumption that the ERPs are time- and phase-locked [37]. Figure 3 presents this concept visually using a specific event from the epoched BD task data. The critical assumption of signal averaging is that the noise in the EEG data is randomly distributed and is therefore cancelled out when averaging, whilst the ERP signal follows an underlying pattern and is thus amplified during averaging.

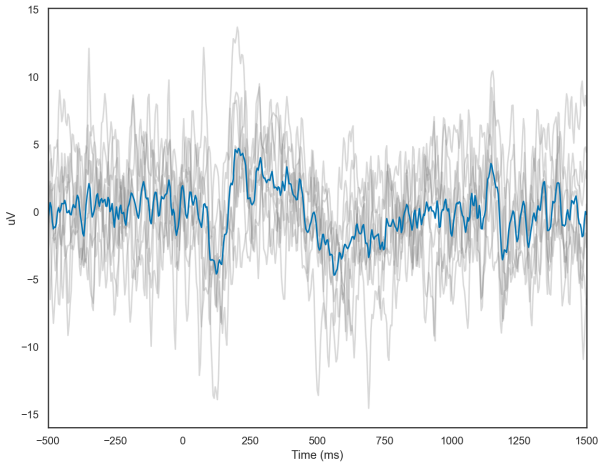

Figure 3: Example of performing signal averaging on repeated event stimuli from the BD task. Signal averaging is based on calculating the average waveform (blue) of all ERP waveforms (grey), with the assumption that the ERPs are time- and phase-locked. The goal is to boost the ERP signals, whilst cancelling out the random noise. ERP = event-related potential

2.5.2 Target feature. The goal of supervised machine learning is to approximate some unobservable function $f(x) \rightarrow y$ by learning from pairs of observations x_i, y_i in our dataset. To minimise the difference between the unobservable function $f(x)$ and our predictive model $h(x)$, we must have some method of quantifying the loss. We do this by measuring the error between the predicted value $\hat{y}$ and the true value y for each of the N data points in the test set:

$$\text{error}(f, h) = \sum_{i=1}^{N} \text{error}(y_i, \hat{y}_i) \tag{1}$$

The better we are able to measure this error, the more the model can be optimised. Researchers thus wish for the highest quality target signal available in their dataset. In our study, the primary target was the binary categorical variable indicating ASD status. Because the severity of ASD is highly varied, a binary variable of this kind was a low quality signal. For instance, a mild case of ASD was hard to differentiate from an extreme case of ASD. This low quality signal made optimising the model difficult. Therefore,

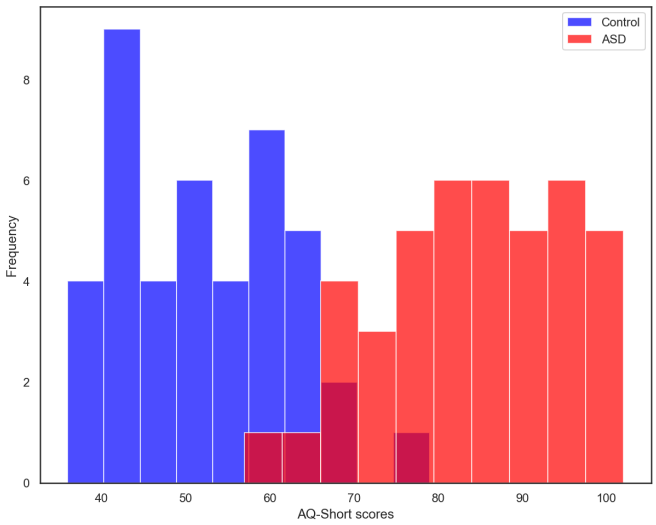

Figure 4: Autism Quotient (AQ)-Short scores successfully distinguished most ASD individuals from controls ($r = 0.83$), making it a strong predictor of ASD status.

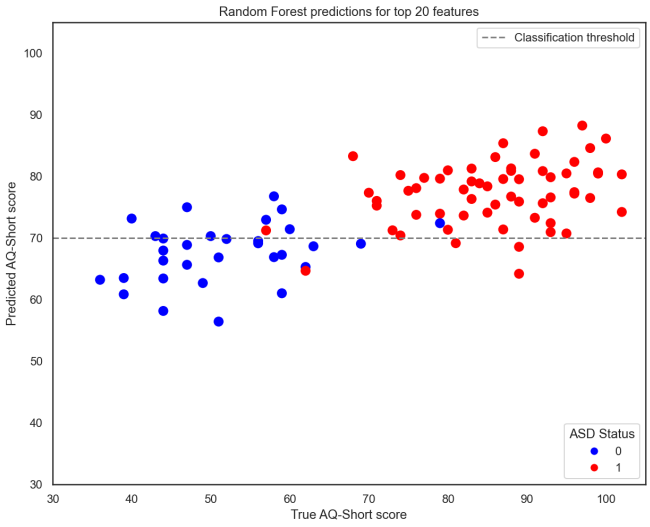

Figure 5: Scatter plot of predicted vs. true AQ-Short scores, with the adjustable classification threshold set to > 70 to define ASD. Data points represent ASD participants (red) or controls (blue) based on their diagnosed ASD status. A Random Forest regressor with 20 features is used to illustrate how an intermediate regression step using AQ-Short scores is able to separate ASD and controls.

we explored the use of a higher-quality numeric signal (the AQ-Short score) as an intermediate target feature. We expected this to produce higher precision in the calculation of model error, thus boosting performance. In our dataset, AQ-Short scores successfully distinguished most ASD individuals from controls (see Figure 4), making it a strong predictor of ASD status ($r = 0.83$). The regression task of the AQ-Short score estimation could then be converted back to a binary classification by using a threshold value ($\tau > 65$) [30] to classify subjects as ASD or control.

2.6 Feature engineering

The averaged ERP signals represented time-series data for each individual and task. These sequences of temporal data had to be collapsed into non-temporal summaries so that the learning algorithms could model them as independent observations. This process of engineering representative features is an integral part of learning from time series data. We used the open-source Python package *tsfresh* `0.16.0` to engineer both time-domain and frequency-domain features from the epoched data [38]. This exhaustive feature engineering approach was chosen to (i) eliminate bias by not preselecting particular features and (ii) enhance reproducibility by engineering features using open-source software with pre-selected parameters. All 72 available methods for calculating features were used, resulting in 763 features per sensor. This included fast Fourier transform (FFT) coefficients, wavelet coefficients, autocorrelation, entropy, and linear trends. The complete list of methods can be found in the tsfresh documentation.

2.7 Feature selection

A core goal of this study was to minimise methodological biases and enhance the reproducibility of our approach. Instead of hand-picking a few techniques and parameters to transform the ERP signals into learnable features, we used a full suite of techniques and parameter values. The total feature set computed from each EEG session consisted of 763 engineered values computed across 9 channels, totaling to 6,867 features. Since all available *tsfresh* feature calculators were used, it was likely that some features were uninformative (e.g. constant values) or highly correlated, thus requiring considerable feature selection. In order to select the optimal features (and number of features) for modelling, another automated approach was employed.

Recursive feature elimination (RFE) was used to rank the importance of all features. The method selected features by recursively training and evaluating a small model on ever-smaller feature sets [39]. A `DecisionTreeRegressor` algorithm was used as the base estimator. Unlike forward- and backward-selection, RFE exploits the learned weights of the base estimator to efficiently rank all features. RFE was performed on the entire dataset to obtain the feature ranking, using the rfe method in the *scikit-learn* framework (version `0.22.2`). Features were standardised by removing the mean and scaling to unit variance using the `StandardScaler` method in *scikit-learn*.

2.8 Learning algorithms

We implemented two types of learning approaches to classify ASD. One was a conventional binary classifier, using ASD status as the target feature. The other learning type performed classification via an intermediate regression step – using AQ-Short scores as the intermediate target feature. The predicted AQ scores were then divided by a predefined threshold ($\tau > 65$) to classify ASD and controls (See Figure 5).

Five machine learning algorithms were used: Decision tree (DT), random forest (RF), XGBoost (XG), generalised linear model (LR), and support vector machine (SVM). We chose these algorithms to obtain a range of tree-based, kernel-based, and linear approaches to compare how algorithm classes handled our problem. Each algorithm was implemented as a classifier or regressor depending on the learning type used. All algorithms, except XGBoost, were implemented with *scikit-learn* version `0.22.2`. XG was implemented with the *xgboost* library (version `1.1.1`). For fair comparison, and to prevent overfitting, all algorithms were instantiated with their default parameters. Descriptions of each algorithm and further reasons for inclusion are found below. Even though many studies have successfully used neural networks to classify ASD, we were limited by the small cohort size and variety of events. Early testing showed that even multi-layer perceptrons with a single hidden layer – the simplest version of a neural network – failed to converge with less than 100 examples to learn from. Thus, neural networks were untenable given our dataset and exhaustive feature engineering process.

2.8.1 Decision tree (DT). These are the simplest form of a tree-based algorithm. The training data is split into nodes, where a predefined metric chooses a feature at each step that best splits the set of observations [40]. DTs were included to investigate whether they would be outperformed by their ensemble counterparts, RF and XG, and because a DT was the base estimator used for recursive feature elimination – which could have conferred some advantage.

2.8.2 Random forest (RF). RF is an ensemble learning method that trains a large number of small decision trees in parallel and takes the average of all the individual decision tree estimates to make a prediction. Because RF implements bootstrapping aggregation (bagging), each tree in the ensemble is built from different re-sampled subsets of the training data and predictions are combined. This has been shown to increase generalisability and reduce the risk of overfitting [41]. RF has also performed well in previous studies of a similar use case to ours [22].

2.8.3 XGBoost (XG). Like RF, XGBoost is also an ensemble learning method based on averaging the predictions of many decision trees. However, instead of a parallel bagging approach, XGBoost use an iterative gradient boosting algorithm. This approach trains each subsequent tree in a way that emphasises the examples that the previous tree failed to predict accurately. The goal of boosting is to minimise both bias and variance, and has been show to do so effectively even with noisy data [42]. XGBoost expands on traditional boosting by implementing sophisticated penalisation of trees, Newton Boosting, proportional shrinking of leaf nodes, and extra randomisation parameters.

2.8.4 Support vector machine (SVM). SVM is a kernel-based model that aims to find the higher-dimensional hyperplane that maximises the margin between data points while minimising the generalisation error. In the case of classification, this means finding a hyperplane that best divides a dataset into two classes. In the case of regression, the hyperplane is set to predict the continuous target value as accurately as possible. A predefined maximum error margin (ϵ) restricts the space in which the hyperplane can be defined. SVM has shown to be tolerant to noise [43], and has achieved high accuracy in predicting ASD in previous studies [21].

2.8.5 Linear models (LR). In the case of regression, Lasso regression was used. Lasso (least absolute shrinkage and selection operator) is a regression method that implements regularisation to reduce variance and assist feature selection. During the model fitting process, Lasso selects a subset of the provided features by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value – effectively penalising model complexity. In the case of classification, logistic regression was used. It utilises the sigmoid function to map predicted values to probabilities. A defined probability threshold then classifies the predicted values. We set the `penalty` parameter as `l1` and the `solver` to `liblinear` to make Lasso and Logistic regression as closely comparable as possible. These models were included to assess how well generalised linear models perform in this problem domain, compared with tree-based and kernel-based algorithms.

2.9 Validation and performance measures

To estimate the general performance of the learning algorithms, resampling was needed. Thus, all performance measures were estimated through 100 rounds of Monte Carlo cross-validation (MCCV). At each iteration, the dataset was split into a train set and test set in a 70 : 30 ratio. All models were trained and tested on identical splits in each iteration so that direct comparisons could be made between them. These splits were stratified on ASD status, thus ensuring that the ratio of ASD and controls was kept equal across splits. Because the dataset was already balanced by downsampling, this ensured that each split was exactly half ASD individuals in both the train set and the test set. All 100 performance results for each model configuration were recorded. From these distributions, the mean, 5th percentile, and 95th percentile were used to establish an expected value and 95% confidence intervals (CIs). These were considered robust estimates of the general-case performance of the model configurations and allowed us to infer performance on other (similar) datasets with some degree of confidence, determined by the CIs. The percentile method was used, as it makes no assumptions about the shape of the underlying distributions. Different *Scikit-learn* metrics were used to score all model configurations to determine which achieved the best mean predictive performance. We report the mean absolute error (MAE), mean square error (MSE), classification accuracy, sensitivity, and specificity of the different models to evaluate their performance. Because we made use of both classification and regression approaches, we required both categorical (accuracy, sensitivity, specificity) and continuous (MAE, MSE) metrics. We used multiple metrics for each type to verify that top-performing models were using the underlying patterns to predict ASD and not exploiting any undetected data leaks or imbalances. These extra precautions allow us to be confident that our top results are legitimate.

Accuracy measures the ratio of correct predictions to total predictions, where the correct prediction is the sum of the true positives (TP) and true negatives (TP), and the total prediction is the sum of the correct prediction plus false positives (FP) and false negatives (FN).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

Sensitivity measures the ratio of correctly identified ASD subjects to total ASD subjects (true positive rate).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Specificity measures the ratio of correctly identified controls to total controls (true negative rate).

$$\text{Specificity} = \frac{\text{TN}}{\text{TP} + \text{FP}} \tag{4}$$

MAE measures the average absolute difference between the predicted value and the actual value of the test sample (difference between the predicted AQ-Short score and the actual AQ-Short score).

$$\text{MAE} = \frac{1}{N}\sum_{i=0}^{N}|\hat{y}_i - y_i| \tag{5}$$

MSE measures the average of the squares of the errors – that is, the average squared difference between the predicted AQ-Short score and the actual AQ-Short score.

$$\text{MSE} = \frac{1}{N}\sum_{i=0}^{N}(\hat{y}_i - y_i)^2 \tag{6}$$

where N is the number of data points. $\hat{y}_i$ is the predicted value and y_i is the actual value for observation i.

We also compared the best models using their receiver operating characteristic (ROC) curves to quantify performance under different tolerances for sensitivity and specificity. The area under the ROC curves (AUC) showed the capability of the models to distinguish between ASD and healthy subjects based on different thresholds (τ-values). Higher AUC values indicate models that are better in distinguishing between ASD and control subjects across tolerances – making them better-suited to clinical applications.

2.10 Experimental Design

The study consisted of two distinct experiments, each with its own dependent and independent variables.

2.10.1 Experiment 1. The visual tasks and learning types were compared in a multi-variable analysis to determine which configurations resulted in the best predictive models. Each of the 5 learning algorithms was iteratively trained and tested in a grid search that varied the three independent variables:

(1) The visual task: BD vs. OD
(2) The learning type: regression vs. classification
(3) The number of (ranked) features to use: 2 - 150

The primary measure of performance was mean classification accuracy (after 100 rounds of MCCV), but sensitivity and specificity were considered to ensure models had no critical biases.

2.10.2 Experiment 2. The best-performing configurations from the previous experiment were selected and compared as the AQ-Short score threshold (τ) was varied from 0 to 112. This allowed us to construct a ROC curve and determine which algorithms performed the best when the sensitivity and specificity tolerances were varied (i.e. in real-world clinical applications). The primary measure of performance was the AUC value).

2.11 Source code

The Python 3 code for the preprocessing and learning pipelines is open-sourced under a GNU GPL v3 license and is available at the project repository (github.com/michbarboure/autism-classifier). It also includes *Jupyter* notebooks showing all the statistical analyses performed. This should allow others to replicate the findings of the study on similar datasets.

3 RESULTS

3.1 Visual task comparison

We compared the mean accuracies of both tasks between different algorithms and learning types to determine whether a particular task resulted in a better predictive performance. Wilcoxon signed-rank tests were performed for each pair of algorithms. Figure 6 shows a compilation of mean accuracies across 2–150 ranked features from different visual tasks and learning types. Only the results from the top 40 features were taken into consideration when doing the comparisons, since all models performed best within this range. The BD task performed equally or significantly better across all configurations, and DT consistently achieved the highest mean accuracy. In regression, DT achieved an accuracy of 0.81 with the BD task, compared to 0.70 with the OD task ($p < 0.001$). In classification, DT achieved an accuracy of 0.79 with the BD task compared to 0.72 with the OD task ($p < 0.001$). Results for all comparisons are provided in the supplementary material. Because of its superior performance, only the BD task was used for further analysis.

3.2 Learning type comparison

Next, we compared model performance for classification and regression to determine which learning type resulted in the best overall performance. Again, only the top 40 features were taken into consideration and Wilcoxon signed-rank test were used for comparisons. Regression resulted in significantly higher accuracy for DT and RF – the learning algorithms achieving the best overall performance. DT achieved an accuracy of 0.81 with regression compared to 0.79 with classification ($p < 0.01$). RF achieved an accuracy of 0.79 with regression compared to 0.76 with classification ($p < 0.001$). Results for all comparisons are provided in the supplementary material. Because of its superior performance and other benefits, only the regression learning type was used for further analysis.

3.3 Model performance

To select the optimal number of features per model, we examined the model performance using the top ranked features and identified the point at which the MAE was at its lowest while maintaining a high accuracy score. The optimal number of features varied per model, but all performed best within the top 12 ranked features.

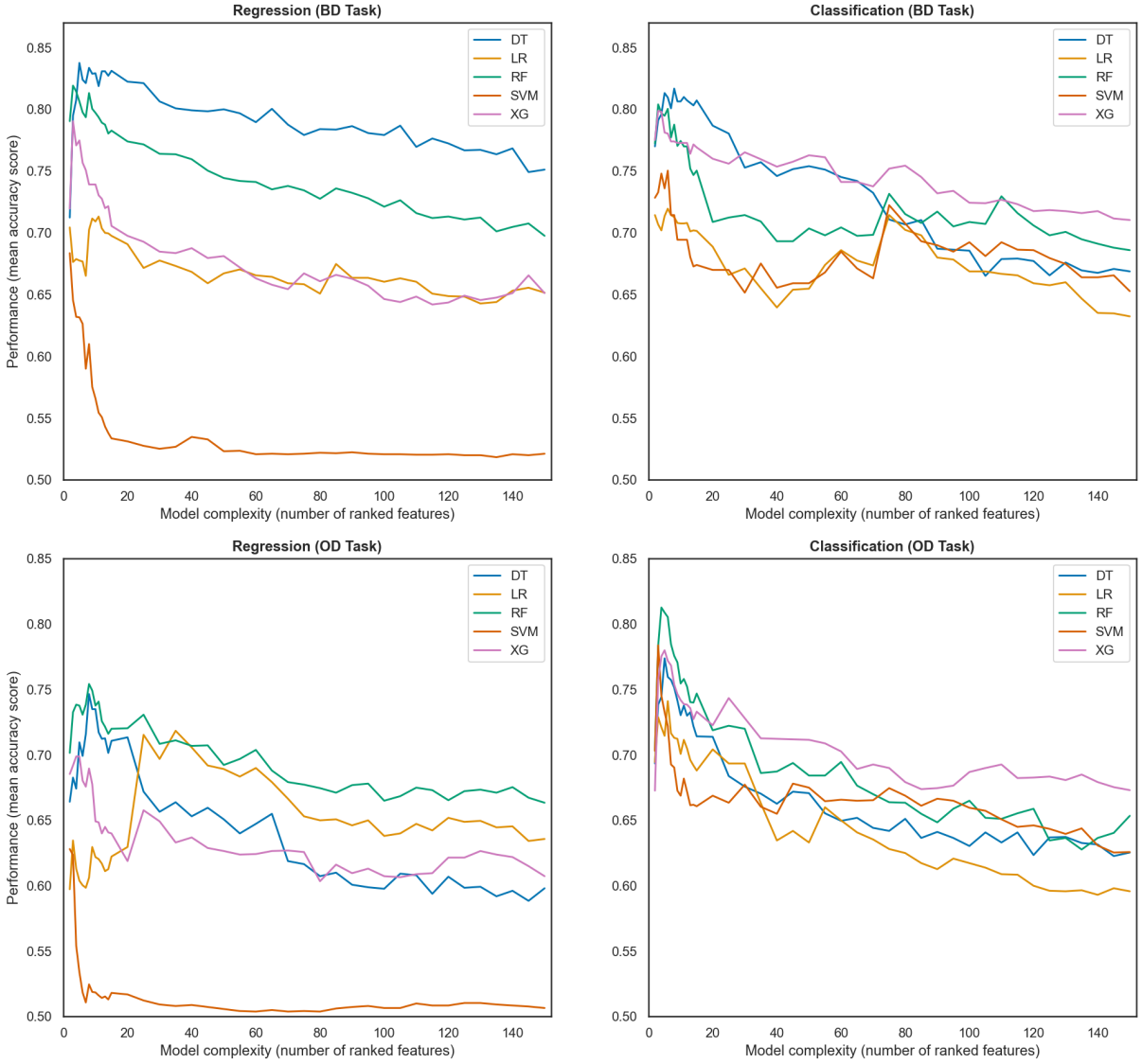

Figure 6: Prediction accuracy compared across algorithms for regression and classification approaches, across both BD and OD visual tasks, as the number of (ranked) features was varied. Results represent the means after 100 rounds of Monte Carlo cross-validation. BD = boundary detection; OD = orientation discrimination; DT = decision tree; LR = Lasso/logistic regression; RF = random forest; SVM = support vector machine; XG = XGBoost

Table 2 reports the MAE, MSE, accuracy, sensitivity, and specificity of the different machine learning algorithms, along with their optimal number of ranked features. The tree-based algorithms (RF, XG and DT) outperformed Lasso and SVM, with DT reporting the highest mean accuracy of 0.83 and lowest MAE of 10.29. However, upon further inspection, RF exhibited better performance than DT in 4 of 5 measurements (sensitivity: 0.87 vs. 0.85; specificity: 0.29 vs. 0.19; MAE: 0.880 vs. 0.781; MSE: 168.13 vs. 212.69), suggesting RF to be more robust than DT. This was further validated after plotting the ROC curve for each algorithm (Figure 7), where RF had a more stable curve and higher AUC than DT (0.887 vs. 0.875). Therefore, of the 5 algorithms analysed, the RF model seems best suited for the ASD classification problem (via intermediate regression).

3.4 Features related to ASD

Table 3 presents the top 10 ranked features used by all of the best-performing learning algorithms. FFT coefficients contribute most to the list, consisting of a mix of 'angle', 'absolute', and 'imaginary' attributes. The features originate from a variety of parietal-occipital sensors, with no single sensor dominating the selection. We compared the mean feature values between ASD and control subjects

Table 2: Mean performance of intermediate regression algorithms after 100 rounds of Monte Carlo cross-validation on BD visual task data, with 95% confidence intervals.

Algorithm	Features	MAE (95% CI)	MSE (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
DT	12	10.29 (6.00–15.02)	212.69 (69.03–422.04)	0.83 (0.64–0.96)	0.85 (0.58–1.00)	0.19 (0.00–0.42)
RF	7	10.38 (7.22–14.74)	168.13 (79.20–300.42)	0.79 (0.64–0.92)	0.87 (0.62–1.00)	0.29 (0.00–0.58)
XG	5	11.55 (8.60–15.52)	210.92 (123.67–346.64)	0.77 (0.60–0.92)	0.81 (0.54–1.00)	0.26 (0.00–0.54)
Lasso	9	13.89 (11.15–17.40)	282.95 (165.66–411.67)	0.71 (0.60–0.88)	0.83 (0.62–1.00)	0.42 (0.17–0.67)
SVM	2	15.10 (12.27–18.14)	300.25 (208.45–394.31)	0.68 (0.52–0.84)	0.84 (0.62–1.00)	0.49 (0.17–0.75)

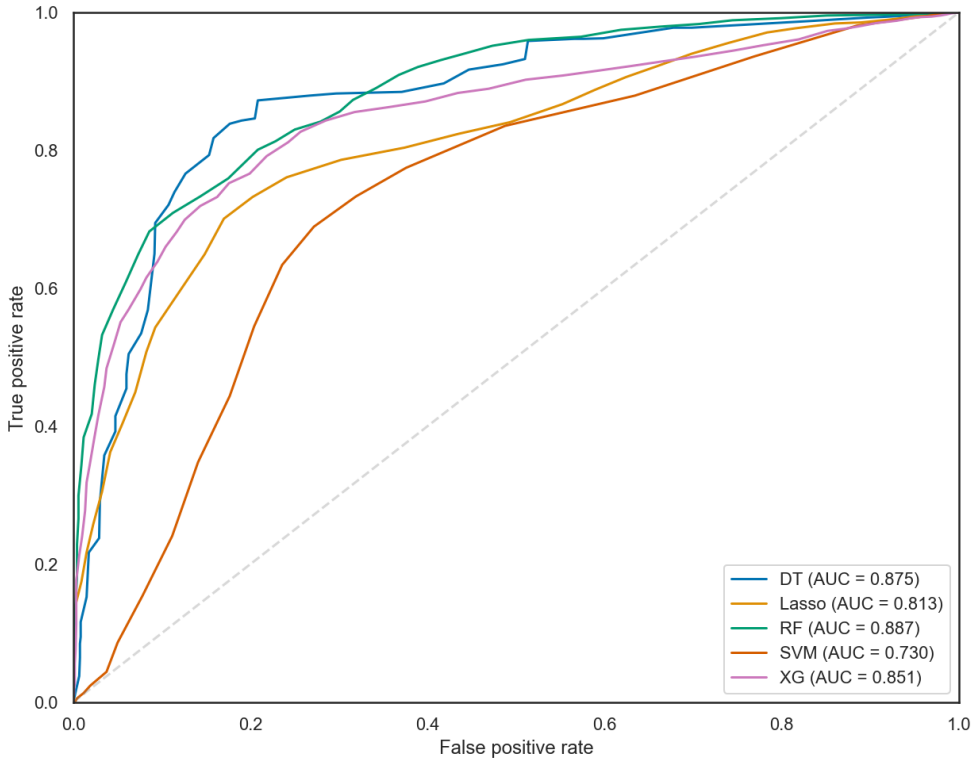

Figure 7: ROC curves of top-performing model configurations on the BD visual task after 100 rounds of Monte Carlo cross validation. AUC = Area under the curve; DT = decision tree; Lasso = Lasso regression; RF = random forest; SVM = support vector machine; XG = XGBoost

using Wilcoxon rank-sum test, of which only 4 features were significantly different at $\alpha = 0.05$ (see Table 3).

4 DISCUSSION

In this study, we targeted the problem of classifying subjects with ASD from controls. We used EEG data from two visual tasks, focusing on readings recorded from sensors located above the parieto-occipital brain region. By using an exhaustive feature engineering workflow followed by a performance-driven feature selection process, a comprehensive exploration of the problem was achieved. Two learning types – binary classification and classification via an intermediate regression step – were implemented. A collection of 5 different learning algorithms were trained on each visual task separately. The main goal of this study was to determine which combination of learning type, visual task, and learning algorithm resulted in the best overall predictive performance to classify ASD. In this section, we discuss the implications of our results. First, we present reasons why including an intermediate regression step in a classification model resulted in better performance and the benefits of this approach. Next, we discuss the advantages of our preprocessing and feature selection workflow. Thereafter, we speculate as to why the BD task resulted in better predictive models than the OD task. Finally, we present the learning algorithms that achieved the best mean performance and compare our results with those of similar studies.

Table 3: Top 10 predictive features for regression after recursive feature elimination.

Feature	Sensor	Ranking	Controls	ASD	Statistic (Z)	P-value
fft_coefficient__attr_"angle"__coeff_17	PO3	1	-80.82	-5.02	3.92	0.00009
fft_coefficient__attr_"abs"__coeff_3	PO3	2	51.80	49.98	-0.78	0.43345
autocorrelation__lag_7	Oz	3	0.59	0.66	2.26	0.02389
fft_coefficient__attr_"angle"__coeff_61	PO8	4	13.29	50.78	1.27	0.20550
fft_coefficient__attr_"abs"__coeff_30	O1	5	16.12	19.80	0.90	0.36721
fft_coefficient__attr_"abs"__coeff_40	O2	6	11.02	11.29	0.53	0.59731
fft_coefficient__attr_"angle"__coeff_55	PO8	7	7.54	-8.91	-0.89	0.37207
fft_coefficient__attr_"imag"__coeff_13	PO3	8	8.92	23.84	2.77	0.00562
fft_coefficient__attr_"imag"__coeff_92	Iz	9	1.52	-2.06	-2.40	0.01660
fft_coefficient__attr_"angle"__coeff_38	PO8	10	41.24	21.23	-1.00	0.31639

4.1 Learning types

Converting the problem of classifying ASD to an intermediate regression task resulted in the best overall models. The AQ-Short scores proved to be a good intermediate target feature, providing more resolution than a binary categorical target, while remaining strongly correlated with ASD status. We suspect that using the AQ-Short scores in an intermediate regression step increased performance by extending the range and quality of the error signal and thus allowing for superior optimisation. Future work should investigate this concept further. A secondary benefit to an intermediate regression model is that the AQ threshold value (τ) is adjustable, enabling real-world applications where the tolerances for sensitivity and specificity may vary. The model was also less sensitive to class imbalance, and was therefore less likely to overfit on unbalanced datasets than classification [44]. When using intermediate regression, predictions could also be easily combined across different sensory tasks, allowing for more robust predictions. Overall, using an intermediate regression model shows great promise to assist in classifying ASD in a clinical setting. Future work could attempt to verify these results on unseen clinical data.

4.2 Preprocessing workflow

Performing signal averaging, exhaustive feature engineering and recursive feature elimination (RFE) generated features that resulted in accurate predictive models. We aimed to keep the workflow as automated and unbiased as possible, whilst minimising computational time and retaining interpretable features. Signal averaging proved effective at reducing noise and radically minimising the computational steps required for feature engineering. Therefore, we recommend it for machine learning studies working with EEG data with repeated events. Our exhaustive feature engineering approach provided an unbiased collection of features that may not have been previously considered in this context. Of these, FFT coefficients appear almost exclusively in the top ranked features after RFE, and are well-represented across all parieto-occipital sensors. Thus, FFT proved to be a rich feature engineering method for this problem. Overall, our highly-automated workflow proved effective and could be of great value to future studies in similar domains.

4.3 Visual task performance

Of the two visual tasks, the boundary detection (BD) task led to better ASD prediction accuracy than the orientation discrimination (OD) task. This result was consistent across classification and intermediate regression models. This may be because the epochs of the OD task were longer than those of the BD task, potentially diluting the brain response signal during the feature engineering step. It is also possible that the BD task produces a more informative brain response for predicting classification, or that the task was more easily executed by the participants and therefore provided more consistent repeats for signal averaging. A previous study was able to make conclusions about horizontal brain connectivity within visual areas based on a similar object boundary detection (BD) task [28], indicating that the BD task may induce significant differences in brain activity that are detectable via EEG. Even though we are unsure why the BD task leads to better prediction accuracy than the OD task, it provides evidence that some sensory tasks may be more predictive for ASD than others.

4.4 Model performance

The tree-based learning algorithms performed the best across both visual tasks and learning types. The standard decision tree and the ensembles (RF and XG) were all based on an underlying tree structure and were all more accurate than support vector machines and linear models. Using the intermediate regression approach together with BD task data, the decision tree model achieved the highest mean accuracy of 83%, while the random forest model was the most robust across all performance parameters and had the largest area under the ROC curve of 0.887. RF is known to benefit more from the use of RFE than other learning algorithms [45], potentially explaining why it performs best. RF has also been shown to perform well in other ASD classification studies. For example, an accuracy of 93% has been achieved from resting-state EEG data in children using only five sensors [22]. We suspect that the tree-based algorithms provide a conducive architecture for this type of problem and warrant further investigation.

5 LIMITATIONS

Although our results support the conclusion that ASD can be classified based on visual tasks in adults, some limitations of this study should be considered.

Firstly, the sample size used for model development and evaluation was small. Many participants had to be excluded because of replaced sensors or missing AQ-Short scores, which can be easily rectified in further analysis. Although comprehensive resampling was performed to establish robust results, further research with more data is required to validate the generalisation capability of the top-performing models. Our cohort was also not representative of the world population in general, and thus may have lacked various forms of diversity that could make detecting ASD easier or more difficult.

Secondly, due to the limited computing hardware available, RFE was only implemented once per task on the entire dataset. It is possible that this caused some indirect overfitting, so future work (with more compute resources) should implement RFE within the resampling loop to avoid this potential issue. Furthermore, we selected the decision tree algorithm as the base estimator for RFE, which may have inadvertently selected features that favoured the tree-based models and biased our results. A more extensive grid search would be required to resolve this.

Thirdly, signal averaging was performed across all events with the assumption that each set of events was equally predictive for ASD. This assumption may not hold. A extensive grid search is required to determine the optimal combination of features that results in the best predictive performance.

6 CONCLUSIONS

Our results provide evidence that machine learning algorithms can distinguish signatures from EEG data of adults performing visual tasks, allowing them to accurately detect ASD. A classification model with an intermediate regression step – using AQ-Short scores as a target – resulted in the best predictive models, providing a novel approach on how to implement machine learning to detect ASD. Tree-based algorithms performed better than support vector machines and linear models, suggesting that a tree-based architecture may be more conducive for this type of problem. As the most robust model, random forest warrants special mention and once again highlights the value of ensemble techniques in real-world problems. Overall, the achievement of a strong classification accuracy shows promise for future applications of machine learning in the assistance of ASD diagnosis.

7 ACKNOWLEDGEMENTS

The author wishes to thank Dr. Dirk Smit and Ricarda Weiland for supervising this project and giving me the freedom to explore a research field with no previous experience in. Thanks are also extended to Gianluca Truda for his constant support and encouragement.

REFERENCES

[1] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. 2013.

[2] Craig J. Newschaffer, Lisa A. Croen, Julie Daniels, Ellen Giarelli, Judith K. Grether, Susan E. Levy, David S. Mandell, Lisa A. Miller, Jennifer Pinto-Martin, Judy Reaven, Ann M. Reynolds, Catherine E. Rice, Diana Schendel, and Gayle C. Windham. The epidemiology of autism spectrum disorders. In *Annual Review of Public Health*, volume 28, pages 235–258, 2007.

[3] Robert E. Nickel and Lark Huang-Storms. Early Identification of Young Children with Autism Spectrum Disorder, jan 2017.

[4] Christian O'Reilly, John D. Lewis, and Mayada Elsabbagh. Is functional brain connectivity atypical in autism? A systematic review of EEG and MEG studies. *PLoS ONE*, 12(5):e0175870, may 2017.

[5] Jared A. Nielsen, Brandon A. Zielinski, P. Thomas Fletcher, Andrew L. Alexander, Nicholas Lange, Erin D. Bigler, Janet E. Lainhart, and Jeffrey S. Anderson. Multisite functional connectivity MRI classification of autism: ABIDE results. *Frontiers in Human Neuroscience*, 7(SEP):599, sep 2013.

[6] Jason J. Wolff, Hongbin Gu, Guido Gerig, Jed T. Elison, Martin Styner, Sylvain Gouttard, Kelly N. Botteron, Stephen R. Dager, Geraldine Dawson, Annette M. Estes, Alan C. Evans, Heather C. Hazlett, Penelope Kostopoulos, Robert C. McKinstry, Sarah J. Paterson, Robert T. Schultz, Lonnie Zwaigenbaum, and Joseph Piven. Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. *American Journal of Psychiatry*, 169(6):589–600, jun 2012.

[7] Yan Jin, Chong Yaw Wee, Feng Shi, Kim Han Thung, Pew Thian Yap, and Dinggang Shen. Identification of infants at risk for autism using multi-parameter hierarchical white matter connectomes. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9352, pages 170–177. Springer Verlag, 2015.

[8] William J. Bosl, Tobias Loddenkemper, and Charles A. Nelson. Nonlinear EEG biomarker profiles for autism and absence epilepsy. *Neuropsychiatric Electrophysiology*, 3(1):1, dec 2017.

[9] Zhongke Gao and Ningde Jin. Complex network from time series based on phase space reconstruction. *Chaos*, 19(3), 2009.

[10] ABR Shatte, DM Hutchinson, SJ Teague Psychological Medicine, and undefined 2019. Machine learning in mental health: a scoping review of methods and applications. *cambridge.org*.

[11] Taban Eslami, Vahid Mirjalili, Alvis Fong, Angela R. Laird, and Fahad Saeed. ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. *Frontiers in Neuroinformatics*, 13:70, nov 2019.

[12] Nicha C. Dvornek, Pamela Ventola, Kevin A. Pelphrey, and James S. Duncan. Identifying autism from resting-state fMRI using long short-term memory networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10541 LNCS, pages 362–370. Springer Verlag, 2017.

[13] Xinyu Guo, Kelli C. Dominick, Ali A. Minai, Hailong Li, Craig A. Erickson, and Long J. Lu. Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Frontiers in Neuroscience*, 11(AUG), aug 2017.

[14] Xia An Bi, Yingchao Liu, Qin Jiang, Qing Shu, Qi Sun, and Jianhua Dai. The diagnosis of autism spectrum disorder based on the random neural network cluster. *Frontiers in Human Neuroscience*, 12, jun 2018.

[15] Colin J Brown, Jeremy Kawahara, and Ghassan Hamarneh. Connectome Priors in Deep Neural Networks to Predict Autism. Technical report.

[16] Zeinab Sherkatghanad, Mohammadsadegh Akhondzadeh, Soorena Salari, Mariam Zomorodi-Moghadam, Moloud Abdar, U. Rajendra Acharya, Reza Khosrowabadi, and Vahid Salari. Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network. *Frontiers in Neuroscience*, 13:1325, jan 2020.

[17] Xia An Bi, Yang Wang, Qing Shu, Qi Sun, and Qian Xu. Classification of autism spectrum disorder using random support vector machine cluster. *Frontiers in Genetics*, 9(FEB), feb 2018.

[18] Jack Fredo, Afrooz Jahedi, Maya Anne Reiter, Ralph-Axel Müller, A R Jac Fredo, and Maya Reiter. Diagnostic Classification of Autism using Resting-State fMRI Data and Conditional Random Forest. *IEEE Engineering in Medicine and Biology Society*, 2018.

[19] David Hairston, Keith W Whitaker, and Jean Vettel. Usability of four commercially-oriented EEG systems. *Article in Journal of Neural Engineering*, 2014.

[20] Andreas Keil, Stefan Debener, Gabriele Gratton, Markus Junghöfer, Emily S. Kappenman, Steven J. Luck, Phan Luu, Gregory A. Miller, and Cindy M. Yee. Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, 51(1):1–21, jan 2014.

[21] William J. Bosl, Helen Tager-Flusberg, and Charles A. Nelson. EEG Analytics for Early Detection of Autism Spectrum Disorder: A data-driven approach. *Scientific Reports*, 8(1), dec 2018.

[22] Dilantha Haputhanthri, Gunavaran Brihadiswaran, Sahan Gunathilaka, Dulani Meedeniya, Yasith Jayawardena, Sampath Jayarathna, and Mark Jaime. *An EEG based Channel Optimized Classification Approach for Autism Spectrum Disorder*.

[23] The Hanh Pham, Jahmunah Vicnesh, Joel Koh En Wei, Shu Lih Oh, N. Arunkumar, Enas W. Abdulhay, Edward J. Ciaccio, and U. Rajendra Acharya. Autism spectrum disorder diagnostic system using HOS bispectrum with EEG signals. *International Journal of Environmental Research and Public Health*, 17(3), feb 2020.

[24] Mark H. Johnson, Emily J.H. Jones, and Teodora Gliga. Brain adaptation and alternative developmental trajectories. *Development and Psychopathology*, 27(2):425–442, may 2015.

[25] T. Gliga, E.J.H. Jones, R. Bedford, T. Charman, and M.H. Johnson. From early markers to neuro-developmental mechanisms of autism. *Developmental Review*,

34(3):189–207, sep 2014.
[26] TM Snijders, B Milivojevic, C Kemner NeuroImage: Clinical, and Undefined 2013. Atypical excitation–inhibition balance in autism captured by the gamma response to contextual modulation. *Elsevier*, 2013.
[27] Luc Kéta, Laurent Mottron, Michelle Dawson, and Armando Bertone. Atypical lateral connectivity: A neural basis for altered visuospatial processing in autism. *Biological Psychiatry*, 70(9):806–811, nov 2011.
[28] MWG Vandenbroucke, HS Scholte, H van Engeland Brain, and undefined 2008. A neural substrate for atypical low-level visual processing in autism spectrum disorder. *academic.oup.com*.
[29] Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, 31(1):5–17, 2001.
[30] Rosa A. Hoekstra, Anna A.E. Vinkhuyzen, Sally Wheelwright, Meike Bartels, Dorret I. Boomsma, Simon Baron-Cohen, Danielle Posthuma, and Sophie Van Der Sluis. The construction and validation of an abridged version of the autism-spectrum quotient (AQ-short). *Journal of Autism and Developmental Disorders*, 41(5):589–596, may 2011.
[31] D Skuse, R Warrington, D Bishop, U Chowdhury Journal of the American ..., and Undefined 2004. The developmental, dimensional and diagnostic interview (3di): a novel computerized assessment for autism spectrum disorders. *Elsevier*, 2004.
[32] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. OpenSesame: An open-source, graphical experiment builder for the social sciences, jun 2012.
[33] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, (7 DEC), 2013.
[34] Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, oct 2017.
[35] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster independent component analysis by preconditioning with Hessian approximations. *IEEE Transactions on Signal Processing*, 66(15):4040–4049, jun 2017.
[36] Yoshua Bengio and Yann LeCun. Scaling Learning Algorithms towards AI. Technical report, 2007.
[37] Wim Van Drongelen. *Signal Processing for Neuroscientists*. Elsevier Inc., 2007.
[38] M Christ, N Braun, J Neuffer, AW Kempa-Liehr Neurocomputing, and undefined 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Elsevier*.
[39] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
[40] L Breiman, J Friedman, CJ Stone, and RA Olshen. Classification and regression trees. 1984.
[41] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001.
[42] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, pages 785–794. Association for Computing Machinery, aug 2016.
[43] John C Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Technical report, 1999.
[44] Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. Classification with class imbalance problem: A Review. *Int. J. Advance Soft Compu. Appl*, 7(3), 2015.
[45] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, nov 2003.

Table 1: Comparison of mean accuracies of visual tasks and learning approaches across algorithms.

Algorithm	BD Accuracy (Class)	BD Accuracy (Reg)	Statistic (Z)	P-value
DT	0.793	0.814	19.0	0.00223
LR	0.696	0.689	65.0	0.22733
RF	0.758	0.789	1.0	0.00016
SVM	0.697	0.575	0.0	0.00013
XG	0.772	0.728	0.0	0.00013
Algorithm	OD Accuracy (Class)	OD Accuracy (Reg)	Statistic (Z)	P-value
DT	0.722	0.698	3.0	0.00021
LR	0.701	0.634	11.0	0.00072
RF	0.751	0.727	24.0	0.00427
SVM	0.688	0.529	0.0	0.00013
XG	0.740	0.661	1.0	0.00016
Algorithm	BD Accuracy (Reg)	OD Accuracy (Reg)	Statistic (Z)	P-value
DT	0.814	0.698	0.0	0.00013
LR	0.689	0.634	12.0	0.00084
RF	0.789	0.727	0.0	0.00013
SVM	0.575	0.529	0.0	0.00013
XG	0.728	0.661	0.0	0.00013
Algorithm	BD Accuracy (Class)	OD Accuracy (Class)	Statistic (Z)	P-value
DT	0.793	0.722	0.0	0.00013
LR	0.696	0.701	66.0	0.24320
RF	0.758	0.751	59.0	0.14742
SVM	0.697	0.688	36.0	0.01758
XG	0.772	0.740	0.0	0.00013