Machine Learning Supervised Classification Methodology for Autism Spectrum Disorder Based on Resting-State Electroencephalography (EEG) Signals

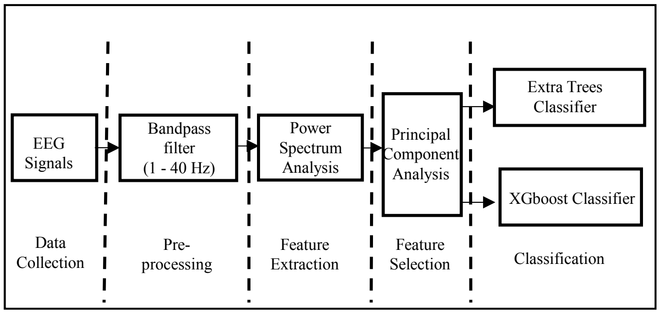
C. Bhaskarachary, A. Jahanian Najafabadi and B. Godde

Neuroscience Group, Jacobs University Bremen, Bremen, Germany
chaitra22bgnhc@gmail.com, {a.jahaniannajafabadi, b.godde}@jacobs-university.de

Autism Spectrum Disorder is a neurological and developmental disorder that starts early in adolescence and lasts throughout a person's life affecting information flow in the brain leading to secondary problems for the patient [1],[2]. Current diagnostic approaches for autism are time-consuming and are mainly based on clinical interviews, to accelerate this process of diagnosing the disease as early as possible with fewer efforts and better accuracy machine learning methods have been proposed recently [3],[4]. Early detection of ASD is vital in enhancing the efficiency of the treatment [5]. The motivation behind this study is the absence of well-defined automated diagnostic procedures for ASD. The objective of this study is to explore and analyze the techniques for EEG pre-processing, feature extraction, classification and identify the abnormal activity for the diagnosis of ASD based on the power spectral density of EEG signals applying machine learning models.

The Figure 1, describes how the brain signal decoding can be viewed as a supervised classification problem. The EEG classification pipeline consists of five major phases: (1) EEG Data Collection (2) EEG data pre-processing (3) Feature Extraction (4) Feature Selection (5) Classification.

Figure 1:EEG Classification Pipeline



The outcome of the classifier was based on 88 normal EEG signals and 100 autistic EEG signals. The signals were sampled at 250Hz, and each signal was segmented into 20 second trails extracting only the eyes-closed part of the signal. Furthermore, the EEG signals were decomposed into four frequency bands: delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz) and beta (13-30 Hz) [3],[4],[6] and power spectral density was calculated. Consequently, Principal Component Analysis (PCA) was implemented reducing the dimensionality of the features to 46 from 440 channels or features and the resulting optimized feature extracted EEG signals were finally fed into the Extra Trees Classifier as well as XGBoost Classifier. XGBoost is used because the framework is more efficient and easier to use algorithm which delivers high performance and accuracy with minimal tuning due to its learning property through parallel and distributed computing [7]. Extra Trees classifier gives higher performance and prevents overfitting of the data. ExtraTrees are capable of feature selection which reduces redundancy, and significance is given to important features which leads to higher accuracy and faster training [8]. Hyperparameter tuning is performed on both the classifiers to improve the accuracy using validation dataset. The performance of the models is evaluated using metrics such as area under the curve (AUC), accuracy, recall, precision, and specificity obtained from confusion matrix [9] on the test data. The evaluation metrics are calculated using the following formulas:

$$Sensitivity\ (Recall) = (TP/(TP+FN)) \quad (1)$$

$$Specificity\ (Selectivity) = (TN/(TN+FP)) \quad (2)$$

$$Accuracy = ((TP+TN)/(TP+FN+TN+FP)) \quad (3)$$

$$Precision = (TP/(TP+FP)) \quad (4)$$

where,

True Positives (TP) – are the number of correctly predicted ASD
False Negative (FN) – are the number of ASD that are incorrectly predicted as not ASD
True Negative (TN) – are the number of correctly predicted as normal signals
False Negative (FN) – are the number of normal signals predicted as ASD.

Table 1, summarizes the classification results on the test data using Extra Trees and XGBoost classifiers and it can be inferred that Extra Trees performs best with Recall-83.3% of the data being correctly identified to have ASD upon the total number of data constituting ASD, Specificity- 54.3% of the data were identified correctly as normal, Precision – 61% of the data is correctly predicted as ASD and classifier predicts with an accuracy of 67.7% and a fair AUC score with 0.74 value. The results of the confusion matrix obtained using the classifiers on the test data can be interpreted as shown in Table 2, and Table 3 for XGBoost and Extra Trees respectively.

Furthermore, EEG signals were analyzed using the extracted feature set. We calculated the distribution of power in an EEG signal as a function of frequency called power spectral density among delta, theta, alpha, and beta over all electrodes and found significant differences amongst them between autistic and normal children shown in Figure 2, which are consistent with the previous studies [2],[3].

Comparing the ASD and healthy EEG signals, results show that enhanced power is found in delta, theta, and beta bands in ASD in comparison with normal EEG signals. However, alpha power is reduced in ASD subjects, which confirms the similar findings of the previous study [2],[3]. The primary goals of this research are accomplished, addressing the challenge to classify ASD and healthy children using resting-state eyes-closed segments of EEG signals. The differences between autistic and healthy children are compared and analyzed using the power spectrum analysis. This research is centered around using an automated methodology for ASD classification using nominal EEG channels to simplify and enhance the efficiency of the diagnosis process. Based on the machine learning models created in this study, it is

Table 1. Summary of test results obtained using extra trees and XGBoost learning algorithms

Metrics	Extra Trees	XGBoost
AUC	0.74	0.72
Accuracy	67.7%	60%
Recall	83.3%	76.7%
Specificity	54.3%	45.7%
Precision	61%	54.8%

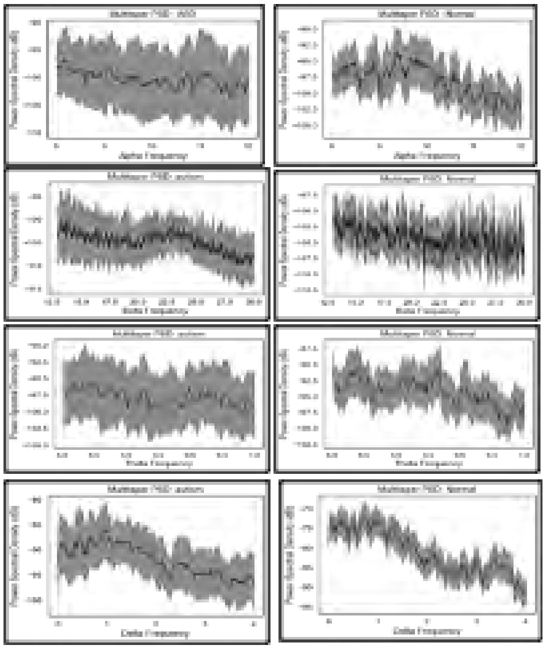
Table 2: Confusion Matrix of XGBoost

Extra Tree Classifier	Predicted Normal	Predicted ASD
True Normal	20	15
True ASD	6	24

Table 3: Confusion Matrix - XGBoost

XGBoost Classifier	Predicted Normal	Predicted ASD
True Normal	16	19
True ASD	7	23

Figure 2: Power spectral density differences between Normal and Autistic EEG signals across theta, delta, alpha and beta frequency bands



observed that the performance of the Extra trees classifier promotes a higher accuracy compared to XGBoost with an accuracy of 67.7%, 83.3% sensitivity, 54.3% specificity, 61% precision and 0.74 AUC. While for XGBoost, the model scored 60% accuracy, 76.7% sensitivity, 45.7% specificity, 54.8% precision and 0.72 AUC.

ACKNOWLEDGMENTS

Hereby, researchers wish to acknowledge and express gratitude for the Child Mind Institute for providing the database under the agreement with Jacobs University Bremen. “This manuscript was prepared using a limited access dataset obtained from the Child Mind Institute Biobank, both healthy and patients with neurodevelopmental disorder particularly ASD. This manuscript reflects the views of the authors and does not necessarily reflect the opinions or views of the Child Mind Institute.” The authors also appreciate all the participants that agreed to voluntarily support the Child Mind Institute and collaborators.

REFERENCES

[1] Jayawardana, Yasith & Jaime, Mark & Jayarathna, Sampath. (2019). “Analysis of Temporal Relationships between ASD and Brain Activity through EEG and Machine Learning”. 151-158. 10.1109/IRI.2019.00035.

[2] G. Brihadiswaran, D. Haputhanthri, S. Gunathilaka, D. Meedeniya, and S. Jayarathna, “EEG-based Processing and Classification Methodologies for Autism Spectrum Disorder: A Review,” Journal of Computer Science, vol. 15, no. 8, pp. 1161–1183, 2019

[3] Kang, J., Han, X., Song, J., Niu, Z., & Li, X. (2020). “The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data,”: Computers in Biology and Medicine, 103722. doi:10.1016/j.compbiomed.2020.103722

[4] T. Pistorius, C. Aldrich, L. Auret and J. Pineda, "Early Detection of risk of autism spectrum disorder based on recurrence quantification analysis of electroencephalographic signals," 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), San Diego, CA, 2013, pp. 198-201, doi: 10.1109/NER.2013.6695906

[5] Ibrahim, Sutrisno & Djemal, Ridha & Alsuwailem, Abdullah. (2017). “Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis,” Biocybernetics and Biomedical Engineering. vol. 38, no. 1, pp. 16-26, 2017, 10.1016/j.bbe.2018.08.006

[6] T. C. U. of Washington, T. Chen, and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 01-Aug-2016. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

[7] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Machine Learning, vol. 63, no. 1, pp. 3–42, 2006.

[8] L. Igual and Seguí Santi, Introduction to Data Science A Python Approach to Concepts, Techniques and Applications, vol. 1. Cham: Springer International Publishing, 2017

Machine Learning Supervised Classification Methodology for Autism Spectrum Disorder Based on Resting-State Electroencephalography (EEG) Signals.

Jacobs University Bremen

Chaitra Bhaskarachary, Amir Jahanian Najafabadi, Benjamin Godde
Neuroscience Group, Jacobs University, Bremen, Bremen, Germany
chaitra22bgnhc@gmail.com, {a.jahaniannajafabadi, b.godde}@jacobs-university.de

Abstract

- Autism Spectrum Disorder (ASD) is a multifactorial neurodevelopmental disorder that affects behavioural impairments in social interaction and communication, along with restricted and repetitive behaviours.
- Current diagnostic approaches for autism are time-consuming and are mainly based on clinical interviews, to accelerate this process of diagnosing the disease as early as possible with fewer efforts and better accuracy machine learning methods have been proposed recently.
- Early detection of ASD is vital in enhancing the efficiency of the treatment and the motivation behind this study is the absence of well-defined automated diagnostic procedures for ASD.
- The main objective of this study is to explore and analyse the techniques for EEG pre-processing, feature extraction, classification and identify the abnormal activity for the diagnosis of ASD based on the power spectral density of EEG signals applying machine learning models.

Methodology

The Figure [1] below illustrates how the brain signal decoding can be viewed as a supervised classification problem. The EEG classification pipeline consists of five major phases: (1) EEG Data Collection (2) EEG data pre-processing (3) Feature Extraction (4) Feature Selection (5) Classification.

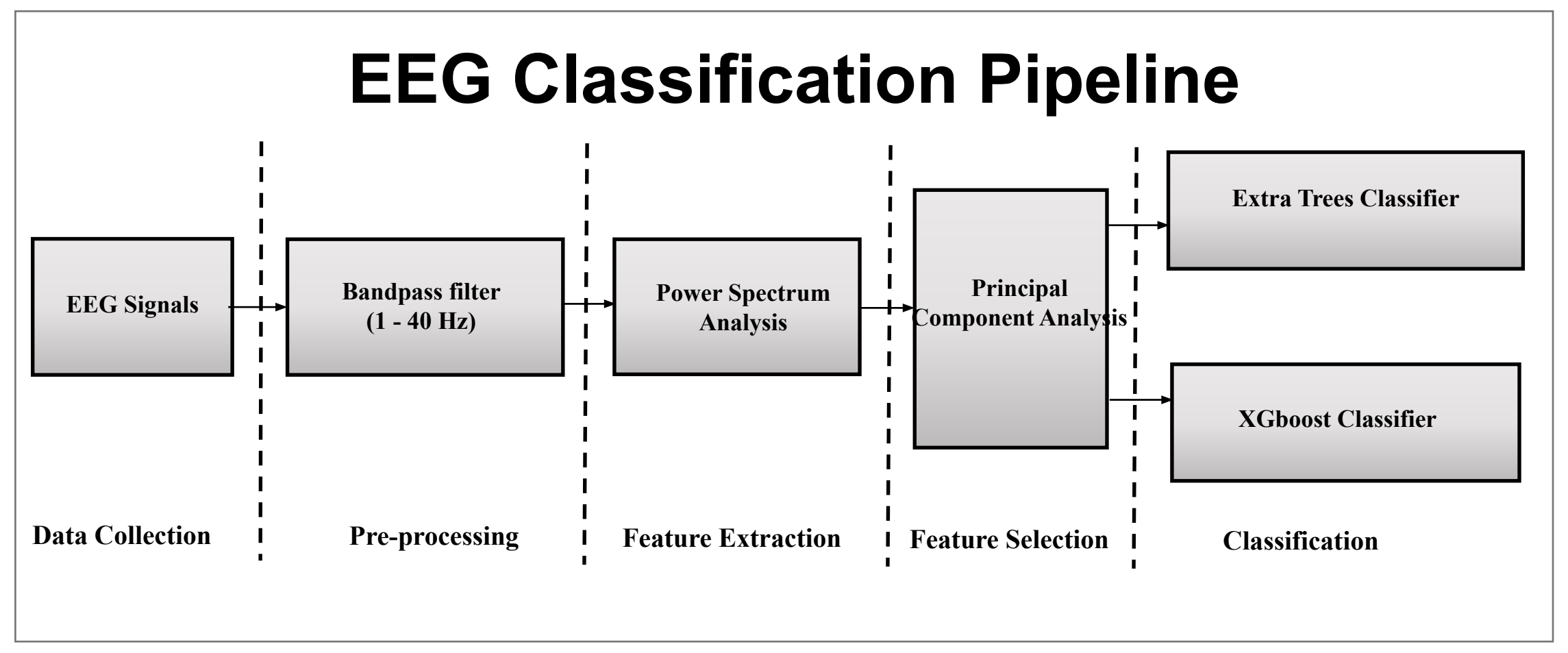


Figure 1. EEG Classification Pipeline

1. EEG Data Collection

- The artifact-free resting-state EEG dataset consists of 100 autistic patients where majority of them were males (82 males, 18 females; mean age: 10.4, SD 3.68) and 88 normal healthy developing subjects (44 males, and 44 females; mean age: 9.8, SD 3.0) within the same age group between 5-19 years old.
- The EEG signals were acquired with HydroCel Geodesic Net with 128 electrodes + Cz at a sampling rate of 500 Hz.

2. EEG Data Pre-processing

- The continuous EEG data were segmented into 20 seconds regular intervals, commonly referred to as epochs.
- Bandpass filtered to restrict signal within a specific frequency range (1- 40Hz).
- Eyes-closed part of the signal was extracted.
- Down-sample the data to 250Hz to improve the computing speed. For ex: by reducing the sampling rate by half say M, the workload is reducing by a factor of four i.e. by (1/M) ^2
- EEG signal voltage readings are converted into millivolts for numerical stability.
- Exponential running standardisation was employed to scale the data to a standard range of values. The final pre-processed data obtained is in the form of a matrix such as (number of trails x channels x timesteps).

3. Feature Extraction

- EEG spectrum contains certain characteristic waveforms that essentially fall within four frequency bands: delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz) and beta (13-30 Hz).
- Additionally, earlier studies have shown that the resting state EEG of ASD showed a "U" shaped profile with excessive power in low-frequency and high-frequency bands and reduced power in the alpha band as illustrated in the figure [2] below.

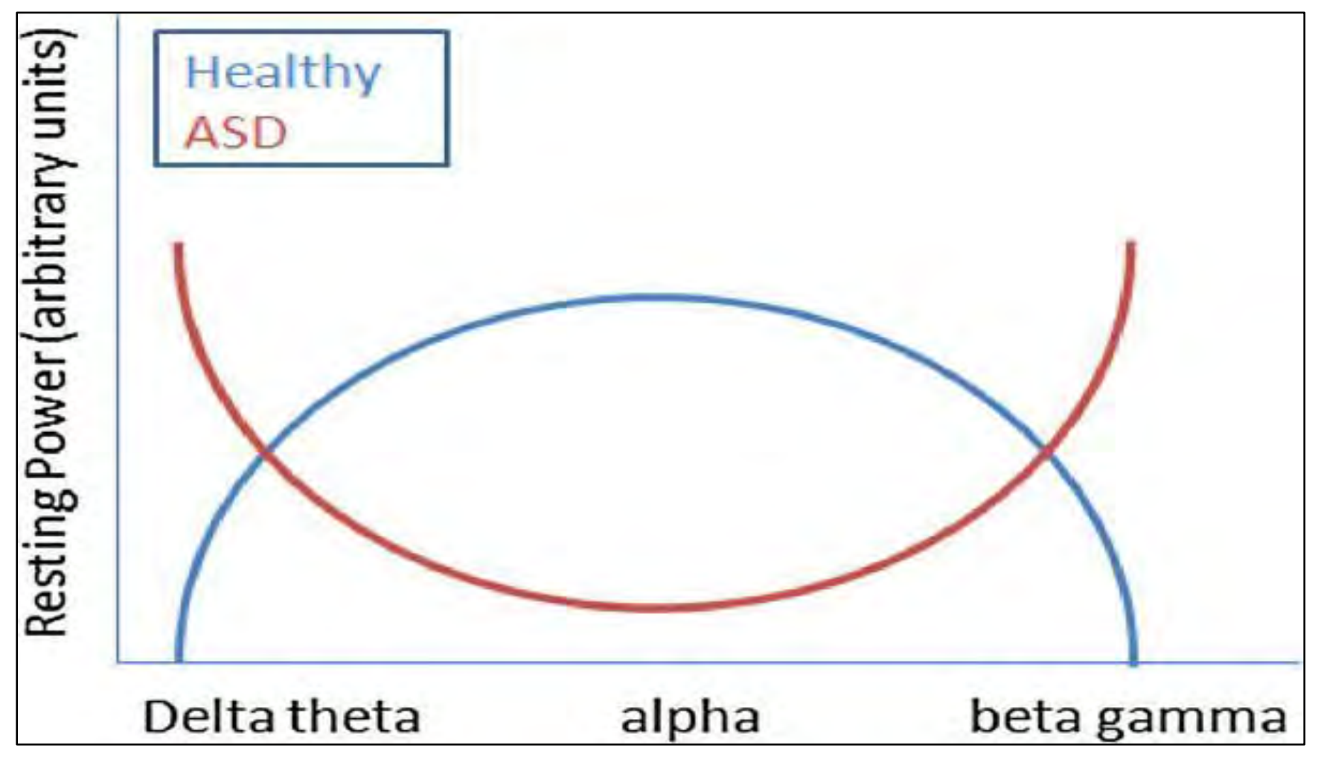


Figure 2. Illustration of U-shaped profile of electrophysiological power

4. Feature Selection

- Principal Component Analysis (PCA) is used to select optimal features for the classification process as it reduces overfitting, enhances accuracy, reduces training time and minimizes dimensionality curse.
- The first principal component has the largest possible variance and consecutive components have the next highest variances and are orthogonal to the previous components.
- The maximum number of principal components required to account for 95% of the variance in the system is considered in this study

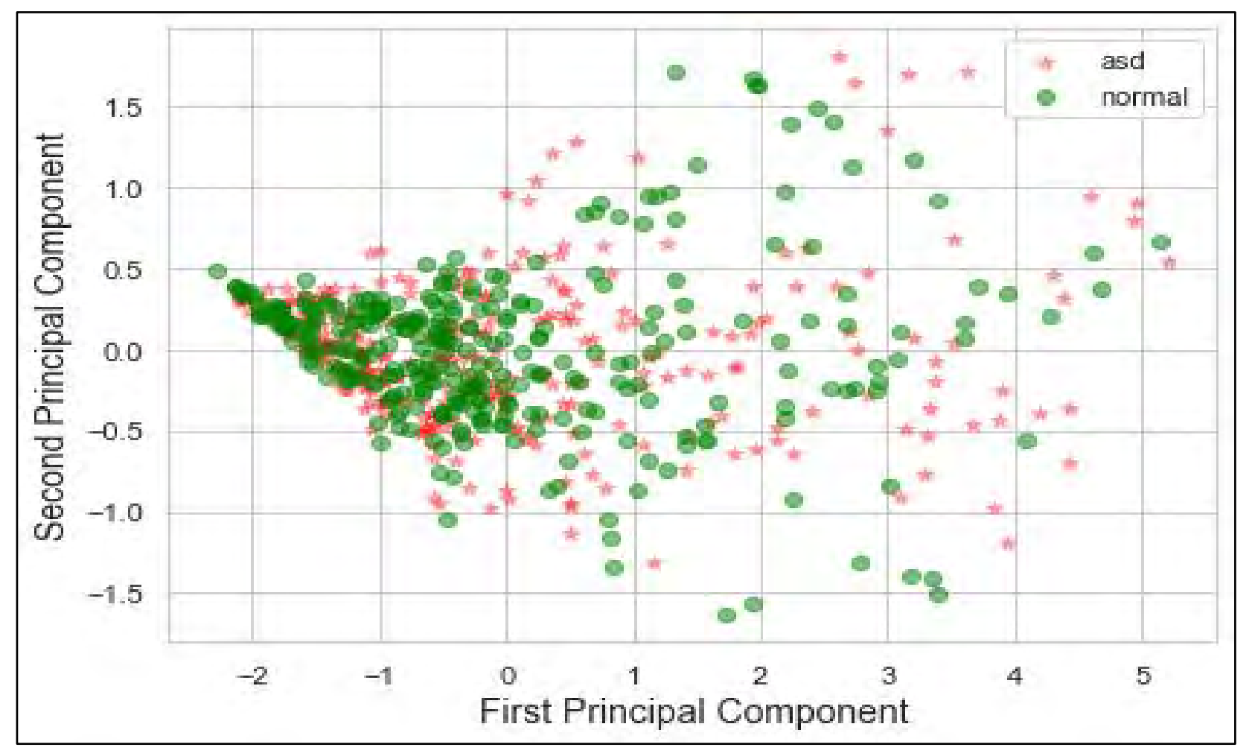


Figure 3. PCA with 95% variance

5. Classification

- Diagnosis of ASD is a binary classification problem, which consists of trying to predict whether the EEG signal corresponds to ASD or a healthy individual.
- The optimized feature extraction EEG signals are finally fed into Extra Trees Classifier, Extreme Gradient Boosting (XGBoost).
- Some of the advantages of using XGBoost over other classification algorithms are parallel processing, regularization prevents overfitting the model and effective tree pruning.
- The Extra Tree Classifier works faster in terms of computational cost, provider higher performance, prevents overfitting of the data. They are capable of feature selection which leads to higher accuracy and faster training.
- For the classification task, the entire dataset is divided into three sets, training, validation and test data. Hyperparameter tuning is performed on both the classifiers to improve the accuracy using validation dataset.

Results and Discussion

The evaluation of classifiers used is based on the conventional metrics called prediction accuracy, precision, recall, specificity and AUC. The evaluation metrics are calculated using the following formulas:

1. $Sensitivity\,(Recall) = \; (TP/(TP + FN))$
2. $Specificity\,(Selectivity) = \; (TN/(TN + FP))$
3. $Accuracy = \; ((TP + TN)/(TP + FN + TN + FP))$
4. $Precision = \; (TP/(TP + FP))$

Where, True Positives (TP) – are the number of correctly predicted ASD. False Negative (FN) – are the number of ASD that are incorrectly predicted as not ASD. True Negative (TN) – are the number of correctly predicted as normal signals. False Negative (FN) – are the number of normal signals predicted as ASD.

Metrics	Extra Trees	XGBoost
AUC	0.74	0.72
Accuracy	67.7%	60%
Recall	83.3%	76.7%
Specificity	54.3%	75.7%
Precision	61%	54.8%

Table 1. Summary of test results obtained using extra trees and XGBoost learning algorithms

From Table 1, it can be inferred that Extra Trees performs best with Recall- 83.3% of the data being correctly identified to have ASD upon the total number of data constituting ASD, Specificity- 54.3% of the data were identified correctly as normal, Precision – 61% of the data is correctly predicted as ASD and the Classifier predicts with an accuracy of 67.7% and a fair AUC score with 0.74 value.

- Besides performing EEG classification, EEG signals were analyzed using the extracted features. We calculated the distribution of power in an EEG signal as a function of frequency called power spectral density among delta, theta, alpha, and beta over all electrodes and found significant differences amongst them between autistic and normal children shown in Figure 4, which are consistent with the previous studies.
- Usually, the low-frequency spectrum has a higher power density than the high-frequency spectrum.
- The differences between autistic and healthy children are compared and analyzed using the power spectrum analysis.
- The results show that enhanced power is found in delta, theta, and beta bands in ASD in comparison with normal EEG signals.

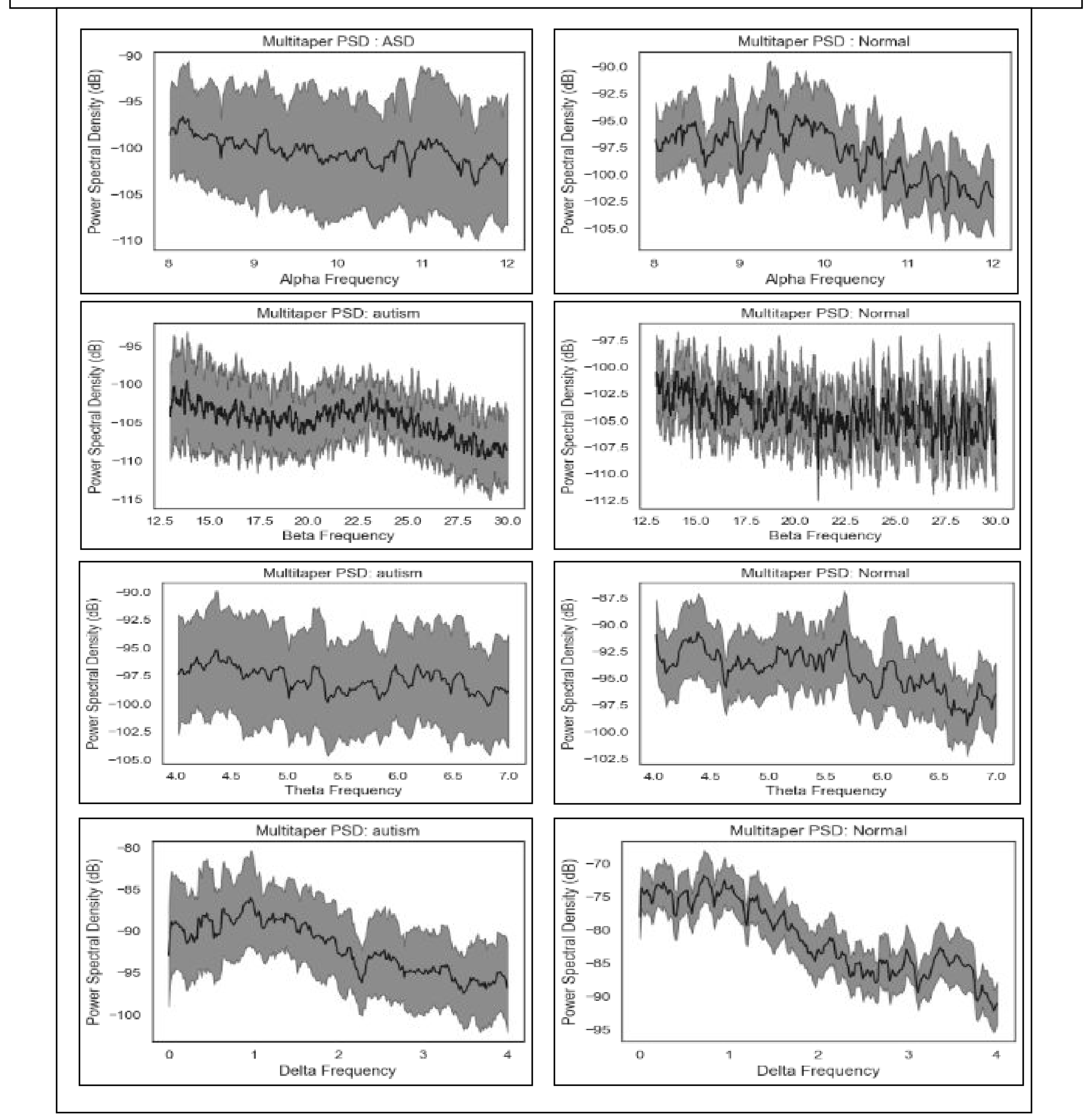


Figure 4: Power Spectral Density differences between Normal and Autistic EEG Signals across delta, alpha, beta and theta frequency bands

Acknowledgements

This research work reported was supported by Jacobs University Bremen. Hereby, researchers wish to acknowledge and express gratitude for the Child Mind Institute for providing the database under the agreement with Jacobs University Bremen. "This manuscript was prepared using a limited access dataset obtained from the Child Mind Institute Biobank, both healthy and patients with neurodevelopmental disorder particularly ASD. This manuscript reflects the views of the authors and does not necessarily reflect the opinions or views of the Child Mind Institute."

v1.0: December 5, 2020