Evaluating the Generalizability of Auditory Features for Diagnosing Autism Spectrum Disorder

Drishti Sethia

July 2024

1 Abstract

Voice prosody has emerged as a promising non-invasive biomarker for identifying Autism Spectrum Disorder (ASD). This raises questions about the generalizability of voice-based biomarkers across different languages. Given the strong genetic basis of ASD, we predict that voice-based attributes should remain consistent across languages. We further hypothesize that cross-lingual training can help prevent machine learning models from overfitting to the characteristics of any specific language.

Using data from three datasets (two in Danish and one in English) that include 74 autistic children and 58 neurotypical children, and leveraging the eGeMAPS feature set, we demonstrate the benefits of cross-lingual training. We utilize Extreme Gradient Boosted Decision Trees (XGBoost) algorithms, which have shown high accuracy in classifying medical datasets, to compare the performance of monolingual training (Danish only and English only) to cross-lingual training.

Our findings reveal that cross-lingual learning outperforms monolingual training by 5 percentage points in F1 score and 4 percentage points in AUC-ROC.

Moreover, our model trained and tested using both languages achieved an F1 score of 0.83 and an AUC-ROC score of 0.86. These results highlight the potential of using vocal non-invasive biomarkers to improve the efficiency of ASD diagnosis.

2 Introduction

Autism Spectrum Disorder (ASD) is a rapidly growing developmental disorder affecting the Central Nervous System, leading to motor impairments and cognitive decline [1]. Early diagnosis is crucial as it enables timely access to interventions, significantly improving developmental outcomes [2,3]. Vocal abnormalities have emerged as potential automatic biomarkers for ASD, with machine learning studies achieving high accuracy in identifying patients, with reported accuracies ranging from 57% to 96% across different languages [4–6]. However, questions remain about the generalizability of these models across diverse neurodivergent populations due to limited data [7]. The focus of this study examines the role of the generalizability of ML models through different languages in hopes to create a non-invasive tool for diagnosing ASD.

Currently, the gold standard for ASD diagnosis involves assessments like the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) [8]. These assessments rely heavily on trained clinicians who interpret speech, facial expressions, and gaze behaviors subjectively. However, their reliance on expert judgment makes them time-consuming and inaccessible in many geographical areas, contributing to a delay in diagnosis despite early symptoms appearing as early as two months of age [9, 10]. Due to this, having an objective automated biomarker could greatly help increase the accessibility and convenience of ASD diagnosis.

ASD is associated with repetitive behaviors, speech delays, and motor impairments [11]. The disorder affects the somatic and autonomic nervous systems, crucial for producing sound. Vocal abnormalities have been identified as potential biomarkers [4,5,12], yet findings vary and require further examination.

Given these inconsistencies, the study focuses on assessing the generalizability of machine learning models using vocal biomarkers across different languages. Reported accuracies in classifying neurotypical and autistic patients vary widely across languages [6], prompting investigation into the benefits of cross-lingual training datasets. Several studies have shown that using a cross-lingual training dataset helps improve detection accuracy [5]. ML models are often subject to overfitting to the characteristics of given datasets (potentially overfitting to the patterns associated with a particular language). By utilizing cross-lingual data, ML models are less susceptible to overfitting and thus be more generalizable and potentially more accurate [5].

The objectivity of this study is to evaluate the performance of:

1) Cross-lingual trained models versus mono-lingual trained models.

2) Cross-lingual trained models compared to existing accuracy benchmarks.

3 Methods

3.1 Model Selection

We utilize Extreme Gradient Boosting (XGBoost) trees to train our model [13]. Decision trees have been notable for their ability to not overfit to small datasets unlike more complex models like neural networks [14, 15]. If the XGBoost faces concerns of generalizability, then it is likely that more complex algorithms will overfit to the training sample. Further past research has showcased that decision trees have had high performance in medical binary classification specifically associated with ASD [16]. XGBoost algorithms enhance this capability by automatically selecting the most important features, which allows for transparent model interpretation and eliminates the need for external feature extraction [13].

3.2 Model Overview

This study trains XGBoosting algorithms using [1] US data [2] Denmark Data and [3] both databases. To evaluate our first objective we will compare the

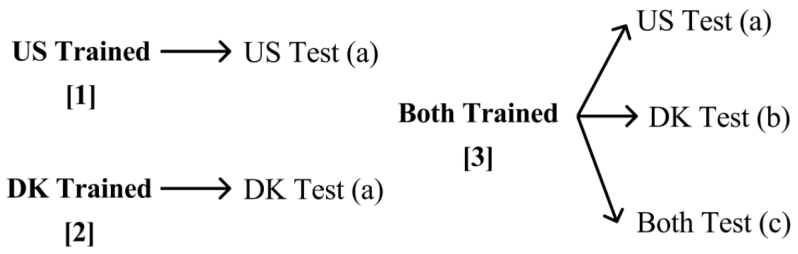

Figure 1: Overview of all models trained.

performance of model [3a] and [3b] when testing on US and Denmark data points respectively, to the performance of model [1a] and model [2a]. In evaluating the second objective, we will train a model trained on both datasets and test on both datasets [3c] to evaluate the performance of cross-lingual learning. Finally we hope to show that XGboosting algorithms have a promising basis in medical diagnosis by comparing the performance of our models to previous models on these datasets.

3.3 Data Source

The data source, as referenced in Rybner et al. (2022) [7], comprises 74 autistic children (475 recordings) and 58 neurotypical (NT) participants (444 recordings). In the Danish dataset, participants describe the content of short videos, whereas in the English dataset, participants retell stories. Both datasets are approximately matched by autism severity level, calculated through ADOS assessments. Specific details are provided in Figure 2. The dataset is publicly available and contains 87 voice features collected through the standard eGeMAPS (emotional Gamut of paralinguistic features) [17]. The eGeMAPS framework captures a comprehensive range of vocal characteristics such as pitch, energy, and spectral features [17].

DK	Participants	Age(Months)	ADOS Com.	ADOS SI	IQ Verbal	IQ Non-Verbal
ASD	24	133	2.71	6.95	99.2	101
NT	27	133	–	–	110	102
Total	51	133	–	–	105	102

Table 1: Mean Analysis of Denmark Participants.

US	Participants	Age(Months)	ADOS Com.	ADOS SI	IQ Verbal	IQ Non-Verbal
ASD	50	153	3.46	8.9	105	106
NT	31	163	–	–	115	114
Total	81	157	–	–	109	109

Table 2: Mean Analysis of US Participants.

3.4 Validation Mechanism

To evaluate the performance of the model, the F1 score is utilized. This score is calculated based on True Positives (TP), False Positives (FP), and False Negatives (FN). The F1 score represents the harmonic mean of precision, defined as $\frac{TP}{TP+FP}$ and recall, defined as $\frac{TP}{TP+FN}$. This metric effectively identifies the model's ability to detect true positives in comparison to false negatives and false positives [18]. Confusion matrices are generated for each classification task, with labels expressed as percentages to facilitate the comparison of different classification accuracies. Additionally, the area under the receiver operating characteristic curve (AUC-ROC), which plots the true positive rate against the false positive rate, is included to provide a second method of validation [19].

3.5 Pipeline

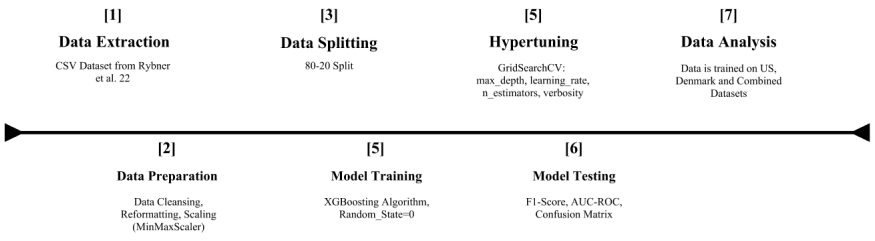

Figure 2: Model pipeline

The dataset was first parsed to sort autistic positive and NT participants. We created a map, associating different ID's with a diagnosis of 0 (NT) or 1 (ASD) using the 'DemoData.csv' file. Note we did discover some inconsistencies, such as ID's having an extra 'A' added to the numerical portion; we only utilized the numerical proportion to identify participants. We then added an extra column, labeled 'diagnosis' to each recording. 3 training datasets were created based on language (English only, Danish only, both languages). Min Max Scaler was applied to all 3 groups which then went into an 80-20 train-test split. To ensure that the dataset containing both languages was not tested on trained samples, we split both the US and Denmark datasets separately and then concatenated both samples to create a cross-lingual dataset. To prevent sampling bias, we ensure that the training dataset containing both languages is created from the combination of the training samples of the US only dataset and the DK only dataset. We further run and test all models, under Random_State=0.

All of the three models are trained using an XGBoost algorithm and are independently hypertuned using Grid Search CV. The algorithm is hypertuned for max_depth (maximum depth of tree), n_estimators (amount of independent trees), and learning rate. To evaluate the performance of each model we extract the F1 score, the AUC-ROC graph and generate a confusion matrix to visualize the results.

4 Results

F1-Score	US-Trained	DK-Trained	Both-Trained
US-Tested	0.82	–	0.87
DK-Tested	–	0.76	0.81
Both-Tested	–	–	0.83

Table 3: F1 scores of all models.

3 and 4 above summarizes the performance across all models using F1 scores and AUC-ROC. In both DK and US datapoints, using cross-lingual learning

AUC-ROC	US-Trained	DK-Trained	Both-Trained
US-Tested	0.80	–	0.84
DK-Tested	–	0.81	0.85
Both-Tested	–	–	0.86

Table 4: AUC-ROC scores of all models.

outperforms monolingual learning by 5 percentage points utilizing F1 score and 4 percentage points utilizing AUC-ROC.

The confusion matrices (4) showcase that training utilizing both languages helps most improve the accuracy of True Positives, by 8.82 and 6.61 percentage points for US and DK datasets respectively.

The model trained on both the languages displays a greater F1 score and AUC-ROC score than any other models trained on this dataset, showing promising results for the usage of cross-lingual training and XGBoosting algorithms [7, 20].

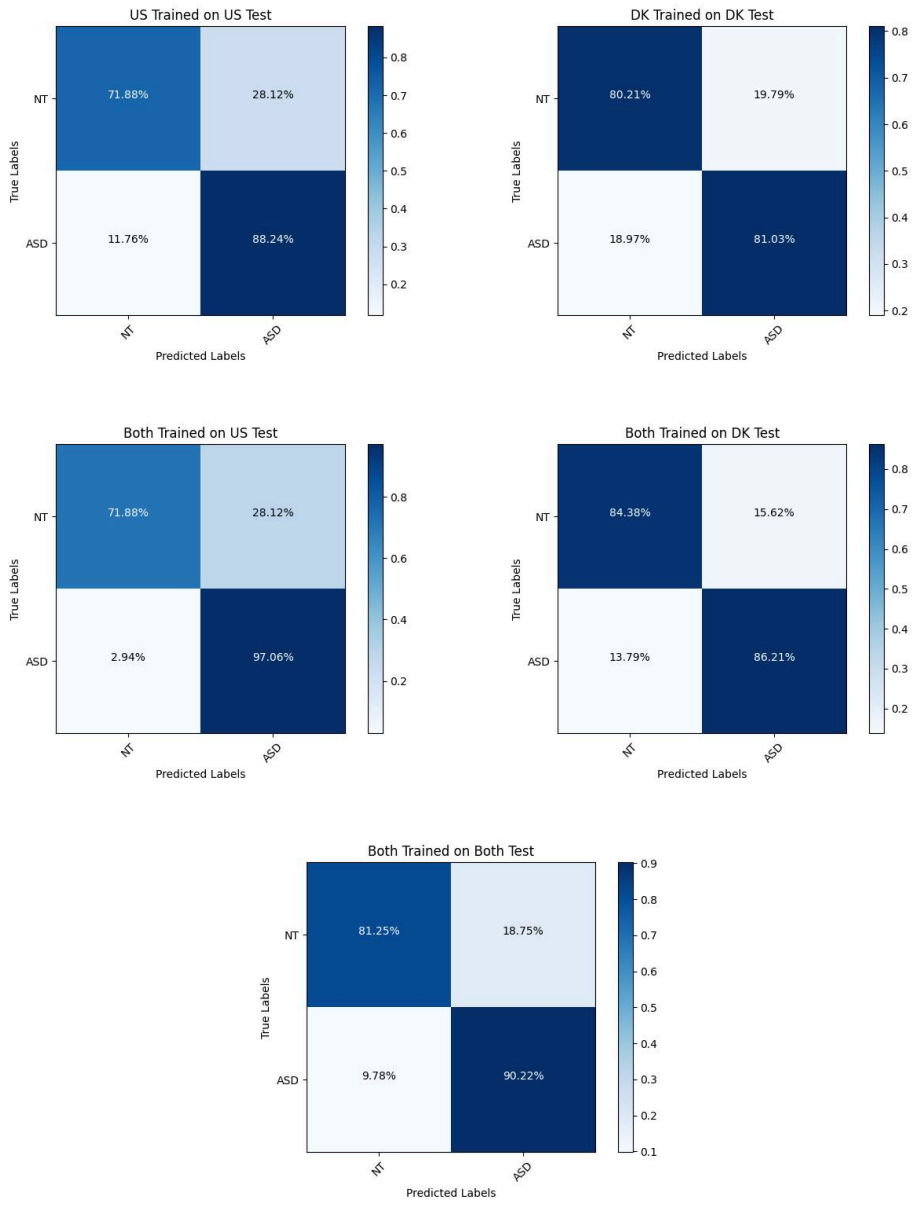

Figure 3: Confusion Matrices for all models

5 Discussion

The results of this study are promising. Coss-lingual training outperformed monolingual training by statistically significant margins utilizing all 3 mechanisms of validation (F1 score, AUC-ROC Curve, Confusion Matrices). This serves to prove our hypothesis that cross-lingual training is less likely to fit to the characteristics of one language and rather identify more generic features. Further, this research serves as validation that characteristics are likely genetic and not dependent on learnt attributes like language.

5.1 Impacts

The impact of this research will have broad implications in [1] scope, [2] accuracy and [3] understanding the neurodivergent brain.

[1] Scope. Cross-lingual training can be applied globally across numerous languages. Moreover, integrating voice-based biomarkers raises the prospect of enhancing ASD diagnosis via smart devices, thereby improving diagnostic convenience and accessibility.

[2] Accuracy. Cross-lingual training enhances the detection of ASD, particularly reducing the occurrence of False Negatives. In diagnostic contexts, a false negative carries greater consequences than a false positive, emphasizing the critical impact of cross-lingual training.

[3] Neurodivergent Brain. Identifying features specifically correlated with ASD across diverse languages aids in distinguishing genetic traits from learned behaviors. This understanding contributes to understanding how the neurodivergent brain influences motor development.

5.2 Limitations

Our study aims to demonstrate the potential of cross-lingual training in ASD diagnosis. We acknowledge the study's limitation of focusing only on two Germanic-based languages, Danish and English, which may share more simi-

larities compared to other languages. Additionally, our participants from the US and Denmark were young children, raising questions about how language development over time could influence diagnosis later in life. Despite these constraints, we hope our research will inspire further exploration of vocal-based biomarkers for ASD diagnosis.

5.3 Next Steps

The next steps of this research include

1) Applying cross-lingual training across diverse languages
2) Finding non-language specific features related to ASD diagnosis
3) Building User Interface for virtual diagnosis

6 Source Code

The source code for this application is linked at https://github.com/dsethia1/ASD_Diagnosis. The dataset for this application is taken from the eGeMAPs generated by Rybner et al. (2022), linked at https://osf.io/9mtpk/?view_only=fa4497a6478b48118e6d15b31cc07567 [7].

References

[1] Karthikeyan Ardhanareeswaran and Fred Volkmar. Introduction. focus: autism spectrum disorders. *The Yale journal of biology and medicine*, 88(1):3—4, March 2015.

[2] Chiugo Okoye, Chidi Obialo-Ibeawuchi, Omobolanle Obajeun, Sarosh Sarwar, Christine Tawfik, Madeeha Subhan Waleed, Asad Wasim, Iman Mohamoud, Adebola Afolayan, and Rheiner Mbaezue. Early diagnosis of autism spectrum disorder: A review and analysis of the risks and benefits. *Cureus*, 15, 08 2023.

[3] Zoe Vinen, Megan Clark, and Cheryl Dissanayake. Social and behavioural outcomes of school aged autistic children who received community-based early interventions. *Journal of Autism and Developmental Disorders*, 53(5):1809–1820, 2023.

[4] Kayleigh K Hyde, Marlena N Novack, Nicholas LaHaye, Chelsea Parlett-Pelleriti, Raymond Anden, Dennis R Dixon, and Erik Linstead. Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders*, 6:128–146, 2019.

[5] Joseph CY Lau, Shivani Patel, Xin Kang, Kritika Nayar, Gary E Martin, Jason Choy, Patrick CM Wong, and Molly Losh. Cross-linguistic patterns of speech prosodic differences in autism: A machine learning study. *PloS one*, 17(6):e0269637, 2022.

[6] Riccardo Fusaroli, Anna Lambrechts, Dan Bang, Dermot M Bowler, and Sebastian B Gaigg. Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis. *Autism Research*, 10(3):384–407, 2017.

[7] Astrid Rybner, Emil Trenckner Jessen, Marie Damsgaard Mortensen, Stine Nyhus Larsen, Ruth Grossman, Niels Bilenberg, Cathriona Cantio, Jens Richardt Møllegaard Jepsen, Ethan Weed, Arndis Simonsen, et al. Vocal markers of autism: Assessing the generalizability of machine learning models. *Autism Research*, 15(6):1018–1030, 2022.

[8] Megan M Stoll, Nicole Bergamo, and Kristina G Rossetti. Analyzing modes of assessment for children with autism spectrum disorder (asd) using a culturally sensitive lens. *Advances in Neurodevelopmental Disorders*, 5:233–244, 2021.

[9] Inge Kamp-Becker, Johannes Tauscher, Nicole Wolff, Charlotte Küpper, Luise Poustka, Stefan Roepke, Veit Roessner, Dominik Heider, and Sanna Stroth. Is the combination of ados and adi-r necessary to classify asd?

rethinking the "gold standard" in diagnosing asd. *Frontiers in Psychiatry*, 12:727308, 2021.

[10] Hanna Drimalla, Tobias Scheffer, Niels Landwehr, Irina Baskow, Stefan Roepke, Behnoush Behnia, and Isabel Dziobek. Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (sit). *npj Digital Medicine*, 3:25, 02 2020.

[11] Melika Kangarani-Farahani, Myrah Anum Malik, and Jill G Zwicker. Motor impairments in children with autism spectrum disorder: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders*, pages 1–21, 2023.

[12] Frédéric Briend, Céline David, Silvia Silleresi, Joëlle Malvy, Sandrine Ferré, and Marianne Latinus. Voice acoustics allow classifying autism spectrum disorder with high accuracy. *Translational Psychiatry*, 13(1):250, 2023.

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

[14] Hiroe Seto, Asuka Oyama, Shuji Kitora, Hiroshi Toki, Ryohei Yamamoto, Jun'ichi Kotoku, Akihiro Haga, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Scientific reports*, 12(1):15889, 2022.

[15] Siripuri Kiran, Ganta Raghotham Reddy, SP Girija, S Venkatramulu, Kumar Dorthi, et al. A gradient boosted decision tree with binary spotted hyena optimizer for cardiovascular disease detection and classification. *Healthcare Analytics*, 3:100173, 2023.

[16] Razieh Asgarnezhad, Karrar Ali Mohsin Alhameedawi, and Hani Akram Mahfoud. An effective model of autism spectrum disorder using machine

learning. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 11(2):389–401, 2023.

[17] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[18] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[19] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[20] Riccardo Fusaroli, Ruth Grossman, Niels Bilenberg, Cathriona Cantio, Jens Richardt Møllegaard Jepsen, and Ethan Weed. Toward a cumulative science of vocal markers of autism: A cross-linguistic meta-analysis-based investigation of acoustic markers in american and danish autistic children. *Autism Research*, 15(4):653–664, 2022.