TILBURG UNIVERSITY

INVESTIGATING LANGUAGE FEATURES AS DIAGNOSTIC MARKERS FOR AUTISM SPECTRUM DISORDER IN CHILDREN

A RANDOM FOREST APPROACH

ÂNGELO, MARTA; KOHNEN, HENRIETTE; VERDUYN, YMKE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

615136, 299066, 138393

COMMITTEE

Dr. Afra Alishahi

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

November 23, 2023

INVESTIGATING LANGUAGE FEATURES AS DIAGNOSTIC MARKERS FOR AUTISM SPECTRUM DISORDER IN CHILDREN

A RANDOM FOREST APPROACH

ÂNGELO, MARTA; KOHNEN, HENRIETTE; VERDUYN, YMKE

1 INTRODUCTION

Autism spectrum disorder (ASD) is a complex, lifelong, neurodevelopmental condition that is, among others, characterized by deficits in social skills and behavior. ASD is a condition with a significant incidence rate estimated around 1% globally (Lord et al., 2020). However, a hallmark of the disorder is heterogeneity (Masi, DeMayo, Glozier, & Guastella, 2017), making it sometimes difficult to diagnose. One of the core symptoms that is thought to be related to ASD and is quite well established is language impairment. Multiple studies showed clear deficiencies in language in multiple domains (I.-M. Eigsti & Bennetto, 2009; Groen, Zwiers, van der Gaag, & Buitelaar, 2008; Kostyuk et al., 2010). In order to improve prognosis for people with ASD, early diagnosis is of utmost importance in order for interventions and therapies to be implemented in early stages and to have the best possible effects. However, thus far, ASD diagnosis usually involves clinical assessments, which are resource intensive and still somewhat subjective. Therefore, with this project, we aimed to contribute to early diagnosis by looking at language as a diagnostic feature. More specifically, we wanted to check if some language features are more telling for ASD than others and if these language features can be used to make a distinction between children with ASD and typically developing children. Following the approaches suggested by Espinoza-Cuadros et al. (2014); Ramesh and Assaf (2021); Slogrove and van der Haar (2020), we chose a Random Forest model implemented in Python that takes the language features whose expressions might be telling of ASD as input. With this, we aspired to build an explainable machine learning model that should not only help with the early detection of ASD but also give an explanation of the most telling features for this process. In particular, we thus aimed to answer the following research questions: *'Is a random forest model able to identify which language features are most telling of ASD in children compared to TD children?'*. Moreover, *'Can a random forest model make a distinction between children with ASD and typically developing children based on their language?'*.

2 DATASET

In order to answer our research question we used the CHILDES database, which consists of a compilation of children's conversations and interactions in transcripts that capture their natural language. We chose language samples of normal developing children (group type TD - typically developing) and children with diagnosed ASD (ASDbank in the Clinical Banks). Specifically, we used transcriptions from experiments conducted by I. Eigsti, Bennetto, and Dadlani (2007), Bang and Nadig (2015) and Tager-Flusberg et al. (1990). Adding the datasets together into one seemed necessary to us in order to get a reasonable amount of training data for the model. However, we made sure that the datasets were sufficiently similar to each other to make this addition possible. All these transcriptions were interview toyplay activities, obtained in a clinical environment, and over the three datasets the transcripts were of similar lengths. From these transcriptions, we extracted language features which we believe could be impaired in children with ASD. Our final data consisted of 38 neurotypical children and 90 typically developing children aged 1-9 years. All of them were English-speaking.

3 FEATURES

Following features were extracted from the transcriptions to use in the Random Forest model. Most of these features were selected based on previous literature that indicated that these might be impaired in children with ASD. Our goal was not necessarily to replicate these findings, but to see which language features are indicative of ASD in the context of a computational model. All features were extracted using CLAN. See the appendix for a more elaborate description of how the features were extracted.

3.1 *Linguistic Metrics*

I. Eigsti et al. (2007) showed a trend for shorter Mean Length of Utterances (MLU) in children with ASD compared to typically developing children. Therefore, we extracted MLU from the transcriptions. More specifically, we extracted the ratio of tokens/utterance, the ratio of morphemes over utterances, number of tokens and number of morphemes used.

3.2 *Lexical Diversity*

Kostyuk et al. (2010) suggested that children with ASD show difficulties in semantic processing, particularly with more abstract or emotion related words. Therefore, we thought it could be interesting to look at the nature of the vocabulary that these children use. Concretely, we chose to look at the lexical diversity by looking at the Type-Token Ratio (TTR; overall lexical diversity), the total number of different item types used, total number of items (tokens) and Lemma-based TTR.

3.3 *Syntactic Features*

According to Groen et al. (2008) the most widely accepted linguistic deficiencies are syntactic deficits. Therefore, we aimed to look at syntactic complexity with the following features: different grammatic types, different grammatic tokens, TTR (syntactic complexity or specific grammatical structures), Clausal Modifier (CMOD), Complementizer (COMP), Clausal Subject (CSUBJ), a non-finite clause that is a nominal modifier or complement (XMOD) and Part of Speech (POS). For the latter, we used the following features: % of adjectives, % of adverbs, % of conjunctions, % of determiners, % of infinitives, % of nouns, % negotiation, % of prepositions, % of pronouns, % of quantatives, % of (modal) auxiliary verbs, % of cop (words used to link the subject to a subject complement), % of verbs, % of past tense, % of past participle tense, % present participle tense and % of plural nouns.

3.4 *Pragmatics/ Interaction metrics*

In their review, Kostyuk et al. (2010) also proposed that individuals with ASD show impairments in effectively using the language. For example, especially joint attention is thought to be a hallmark for ASD. By looking at Mean Length of Turn (MLT) we were able to implement turn-taking features in the Random Forest model. More specifically we used the ratio of words over turns, ratio of utterances over turns, ratio of words over utterances and the ratio of the child's MLT over the adult's MLT to get a share of the conversational load that the child has done. For this ratio, we used words over turns of the child on words over turns of the adults. In order to compute these features we extracted the numbers of turns, number of words and number of utterances.

3.5 *Repetitive language use*

Echolalia, or the repetition of words, is often seen in children with ASD (Kostyuk et al., 2010). Therefore, echolalia of the child and echolalia of the child + CSR (child self repetition) are extracted out of the transcriptions. We thus looked at the repetitions that the child uses of himself/herself but also of the adults present.

4 MODEL

The model we chose for our approach is the Random Forest as it was introduced by Breiman (2001). After trying out some different configurations of hyperparameters, we decided on a Random Forest Classifier with the default hyperparameters provided by the scikit-klearn library (scikit-learn developers (BSD License), 2007-2023) that seemed to work best overall. In order to train the model we started by doing some data preprocessing and cleaning during which we first excluded some auxiliary features we obtained using CLAN. Additionally, we decided to exclude those examples from the data in which the children did not speak at least 20 utterances. This was necessary, as the low number of utterances for these children resulted in missing values for many of the other features that could not be computed by CLAN with that few utterances available overall. Since we aimed to find a distinction between typically developing children and children with ASD based on their overall language, and also wanted to find out which of the features of their language were most important for this classification task, too many missing values in these specific examples were likely rendering them useless for our model. Some more technical preparation of the data can be observed in our code. For reasons of brevity it will not be explained further in this report.

After preprocessing, we thereby ended up with a dataframe with the language data for 128 children and 40 different features for all of them. Since one of the datasets we used for feature extraction (Tager-Flusberg et al., 1990) contained only data from children with ASD, our overall dataset is noticeably unbalanced with 90 children with ASD and 38 typically developing children as mentioned above. Given the rather small size of the overall dataset, we decided to use only 15% of the data as a test set while using the rest for training. We then evaluated the model in multiple different ways in order to get the best possible evaluation of a model on such few datapoints. These results will be presented, analyzed, and discussed in the following section.

5 RESULTS AND INTERPRETATION

When there is only little training data available, k-fold cross-validation is a good approach to make sure a model performs equally well over the different examples in the training data (Yadav & Shukla, 2016), which made it highly relevant for us. To perform 5-fold cross-validation on our training dataset, we used scikit-learn's *cross_val_score*. We obtained both the accuracy and the f1-score for all five folds and got an average over all of them, the results can be found in Table 1. The accuracy values were overall higher, but due to the imbalance of ASD and TD labels in the dataset, the f1-measure is likely a better reflection of the model's performance. Importantly, the f1-values were not very stable over the folds, which indicates that the model's performance is not equally good for all the examples in the data and should remind us to interpret the following results on the test set with care.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Accuracy	0.73	0.77	0.73	0.71	0.86	0.76
F1-score	0.57	0.55	0.57	0.4	0.73	0.56

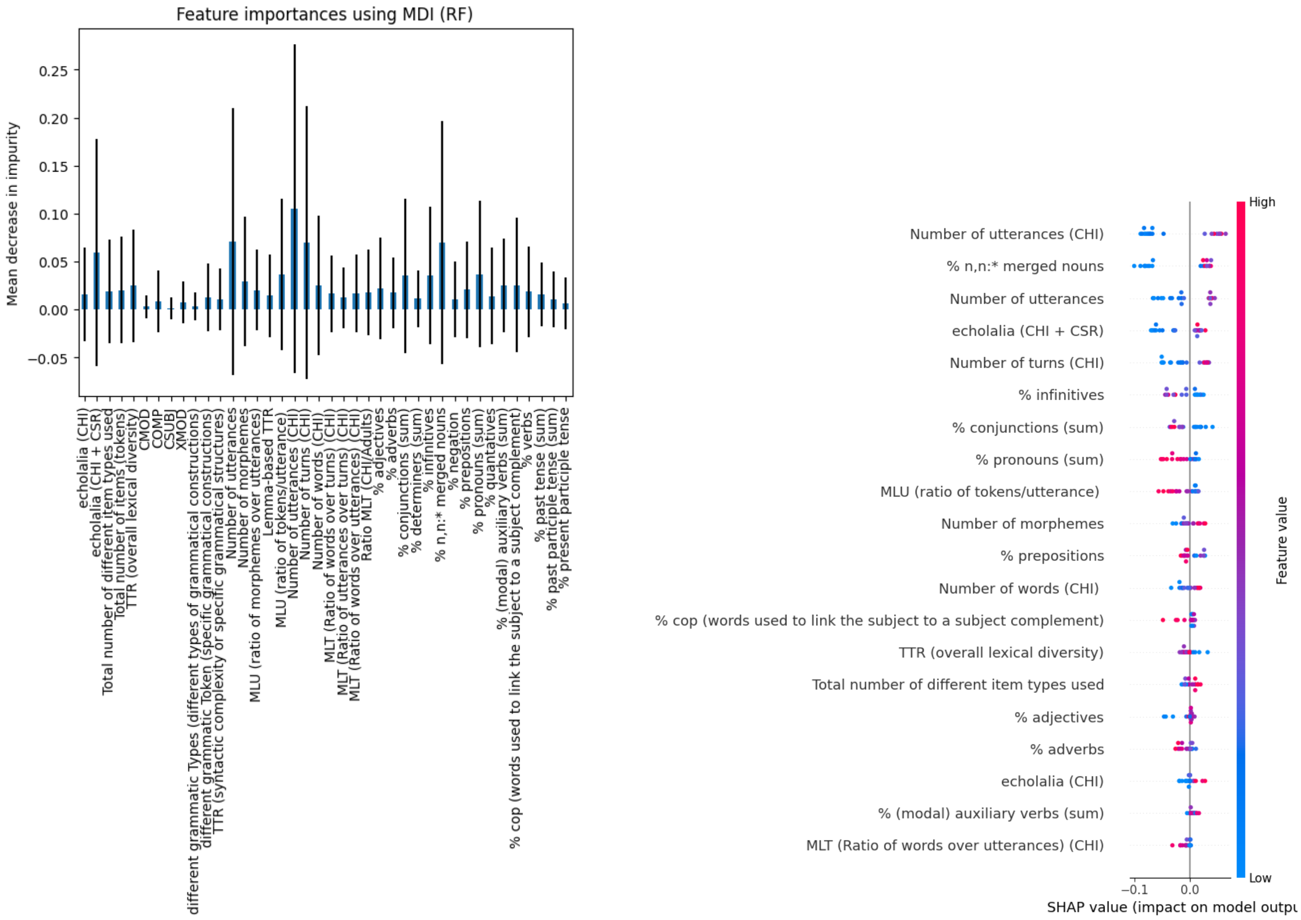
Table 1: Accuracy and F1-Score over 5-fold Cross-Validation

After the cross-validation approach on the training set, we then moved on to applying the model to the test set which included 20 examples from our original dataset. The AUC value for the ROC curve (Figure 2 in the appendix) was 0.85, pointing towards a performance significantly above chance. The confusion matrix (Figure 3 in the appendix) and the classification report with precision, recall, and f1 score for the (Table 2) supported this claim. Importantly, and likely in parts due to the imbalance of the data, our model wrongly classified some children with ASD as typically developing (lower recall of 69%for class ASD), while no typically developing child was wrongly classified to have ASD (high precision of 100% for class ASD). Accordingly, the precision for the TD class was lower (at 64%), while the recall was at 100% here. For clinical application, this model would at least need to get a significantly better recall for the ASD class, so that as few cases as possible are missed.

	Precision	Recall	F1-Score	Support
ASD	1	0.69	0.82	13
TD	0.64	1	0.78	7
Accuracy			0.8	20
Macro Avg	0.82	0.85	0.8	20
Weighted Avg	0.87	0.8	0.8	20

Table 2: The Classification Report for the Random Forest Model

In order to get an idea about which language features are most important for the classification decision of our model, we looked at the mean decrease in impurity (MDI) per feature (Figure 1a). Additionally, we used the Python library SHAP and created a summary plot that shows the impact of the values of the 20 most important features (ordered top to bottom) on the output class (Figure 1b). For a very well-performing model, the important features might be interpreted to be the most telling of ASD in children as well, though an interpretation should always keep in mind that the model reveals correlations rather than causations as is also emphasized by the creators of SHAP (Lundberg, 2018).



(a) MDI for Feature Importance in the Random Forest Model

(b) SHAP Analysis Summary Plot for Feature Impact

Figure 1: Evaluation of Feature Importance and Impact

Given the performance of our model, the most important features here do not necessarily have to be the ones that are also most distinctive of ASD and TD in reality, but the tendencies might still be interesting to look at. Over both analyses, the number of utterances (including as well as excluding non-transcribable utterances), the percentage of merged nouns, the number of turns, and the occurrences of echolalia (including self-repetition) were the five most important features for the model. SHAP additionally identified the percentages of infinitives, conjunctions, and pronouns as well as the MLU of tokens over utterances to be slightly more important than many other features (Figure 4 in the appendix) which is also reflected in the MDI plot.

On the grounds of our literature search, this order of feature importance for the discrimination between ASD and TD seemed reasonable to us. Looking closer at the SHAP summary plot in Figure 1b, for the top five features, we saw that lower feature values made the example more likely to be classified as ASD (coded as value 0) instead of TD (coded as value 1). While this did not surprise us for the number of utterances, the percentage of merged nouns and the number of turns, the fact that the model interprets lower counts of echolalia and self-repetition as indicative of ASD was rather unexpected for us after what we extracted from the literature (I.-M. Eigsti & Bennetto, 2009; Kostyuk et al., 2010). According to Kostyuk et al. (2010), echolalia and repetitive speech patterns are characteristic of ASD in children, while here, the model seems to have used these features the other way around. We paid less attention to the feature values capturing the percentages of infinitives, conjunctions, and pronouns as well as the MLU of tokens over utterances in the SHAP summary plot, as their impact on the model was overall marginal. Nevertheless, we still wanted to point out that for these features, the summary plot also holds partially surprising results. While a higher percentage of infinitives as an indication of ASD seemed in line with the literature, the other three features seemed to have influenced the model in a way different from what is found in the literature (I.-M. Eigsti & Bennetto, 2009; Kostyuk et al., 2010).

6 CONCLUSION AND OUTLOOK

Overall, we would conclude by saying that a model as simple as the Random Forest was able to distinguish between children with ASD and typically developing children surprisingly well given the fact that we only had a quite limited amount of data to train and test it. Provided the tendencies we saw for our model, we would therefore hypothesize that a similar model trained on more and more balanced data could perform reasonably well in distinguishing between children with ASD and typically developing children based on their language and identifying the most telling features.

In addition to the features we used for our present model, we would also suggest including phonological features, which we sadly did not manage to do with the current model, see the appendix for an elaboration on that. The same is true for emotion-related words that play an important role according to the literature (Kostyuk et al., 2010). We tried extracting those with a custom-made Python script (see appendix) but did not get any meaningful results for the words we chose over the transcriptions available to us. Another factor that should be monitored more closely in the future is the age of the children. Our model was trained on data from children over a quite large age range that should be more limited in the future to make the outcomes more reliable, especially given the importance of the number of utterances spoken by the children. Additionally, more feature analysis will be necessary to optimize the model. Excluding some features that seem to be less telling of ASD in children could be excluded in order to improve model performance. As an example, when training the model, we tried excluding the feature representing the percentage of plural nouns in the children's speech, which led to improved overall model performance.

In conclusion, this project indicates that using a computational model with language features, and more specific, a Random Forest approach, could potentially be useful as a diagnostic tool for autism spectrum disorder in children. Also, the differences in language between children with ASD and typically developing children could be captured reasonably well by such a model. However, there are still some challenges that need to be overcome before implementing it in real life.

REFERENCES

Bang, J., & Nadig, A. (2015). Language learning in autism: Maternal linguistic input contributes to later vocabulary. *Autism Research*, *8*(2), 214–233. doi: 10.1002/aur.1440

Bank, T. (2002). *TalkBank Software Programs.* Retrieved from https://talkbank.org/software/

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Cohly, Kostyuk, N., Isokpehi, Rajnarayanan, R. V., Oyeleye, T. O., & Bell, T. P. (2010, 8). Areas of language impairment in autism. *Autism insights*, 31. Retrieved from https://doi.org/10.4137/aui.s5556 doi: 10.4137/aui.s5556

Documentation, N. (2023a). *NLTK - nltk.sentiment.vader module.* Retrieved from https://www.nltk.org/api/nltk.sentiment.vader.html

Documentation, N. (2023b). *NLTK - nltk.tokenize package.* Retrieved from https://www.nltk.org/api/nltk.tokenize.html

Eigsti, I., Bennetto, L., & Dadlani, M. (2007). Beyond pragmatics: morphosyntactic development in autism. *Journal of autism and developmental disorders*, *37*(6), 1007–1023. Retrieved from https://doi.org/10.1007/s10803-006-0239-2 doi: 10.1007/s10803-006-0239-2

Eigsti, I.-M., & Bennetto, L. (2009). Grammaticality judgments in autism: Deviance or delay. *Journal of child language*, *36*(5), 999–1021.

Espinoza-Cuadros, F., Garcia-Zamora, M. A., Torres-Boza, D., Ferrer-Riesgo, C. A., Montero-Benavides, A., Gonzalez-Moreira, E., & Hernandez-Gómez, L. A. (2014). A spoken language database for research on moderate cognitive impairment: design and preliminary analysis. In *Advances in speech and language technologies for iberian languages: Second international conference, iberspeech 2014, las palmas de gran canaria, spain, november 19-21, 2014. proceedings* (pp. 219–228).

Groen, W., Zwiers, M., van der Gaag, R., & Buitelaar, J. (2008). The phenotype and neural correlates of language in autism: an integrative review. *Neuroscience and biobehavioral reviews*, *32*(8), 1416–1425. Retrieved from https://doi.org/10.1016/j.neubiorev.2008.05.008 doi: 10.1016/j.neubiorev.2008.05.008

HuggingFace. (2016). *OpenAI GPT2.* Retrieved from https://huggingface.co/docs/transformers/model_doc/gpt2

Kostyuk, N., Isokpehi, R. D., Rajnarayanan, R. V., Oyeleye, T. O., Bell, T. P., & Cohly, H. H. (2010). Areas of language impairment in autism. *Autism Insights*(2).

Lab, D. D. (2015). *What is Sklearn? | Domino Data Science Dictionary.* Retrieved from https://domino.ai/data-science-dictionary/sklearn

Lord, C., Brugha, T., Charman, T., Cusack, J., Dumas, G., Frazier, T., . . . Veenstra-VanderWeele, J. (2020). Autism spectrum disorder. *Nature reviews. Disease primers*, *6*(1), 5. Retrieved from https://doi.org/10.1038/s41572-019-0138-4 doi: 10.1038/s41572-019-0138-4

Lundberg, S. (2018). *Be careful when interpreting predictive models in search of causal insights — SHAP latest documentation.* Retrieved 2023-11-21, from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%C2%A0insights.html

MacWhinney, B. (2010). *Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4257135/

MacWhinney, B. (2018, 12). Understanding spoken language through TalkBank. *Behavior Research Methods*, *51*(4), 1919–1927. Retrieved from https://doi.org/10.3758/s13428-018-1174-9 doi: 10.3758/s13428-018-1174-9

MacWhinney, B. (2023). *Tools for Analyzing Talk Part 2: The CLAN Program.* Retrieved from https://talkbank.org/manuals/CLAN.pdf

Masi, A., DeMayo, M., Glozier, N., & Guastella, A. (2017). An overview of autism spectrum disorder, heterogeneity and treatment options. *Neuroscience bulletin*, *33*(2), 183–193. Retrieved from https://doi.org/10.1007/s12264-017-0100-y doi: 10.1007/s12264-017-0100-y

Ramesh, V., & Assaf, R. (2021). Detecting autism spectrum disorders with machine learning models using speech transcripts. *arXiv preprint arXiv:2110.03281*.

scikit-learn developers (BSD License). (2007-2023). *RandomForestClassifier.* Retrieved 2023-11-21, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Slogrove, K. J., & van der Haar, D. (2020). Specific language impairment detection through voice analysis. In *Business information systems: 23rd international conference, bis 2020, colorado springs, co, usa, june 8–10, 2020, proceedings 23* (pp. 130–141).

Tager-Flusberg, H., Calkins, S., Nolin, T., Baumberger, T., Anderson, M., & Chadwick-Dias, A. (1990). A longitudinal study of language acquisition in autistic and down syndrome children. *Journal of autism and developmental disorders*, *20*(1), 1–21.

Van Goozen, S., & Frijda, N. H. (1993, 1). Emotion words used in six European countries. *European Journal of Social Psychology*, *23*(1), 89–95. Retrieved from https://doi.org/10.1002/ejsp.2420230108 doi: 10.1002/ejsp.2420230108

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 ieee 6th international conference on advanced computing (iacc)* (pp. 78–83).

APPENDIX A

Clan

CLAN (Computerized Language Analysis) is currently the most general comprehensive tool for the transcription, coding, and analysis of language data (Bank, 2002) used by researchers and clinicians alike. Written by Leonid Spektor and running on Windows, Unix, and OSX platforms, it offers functionalities such as transcription of audio and video data, corpus analysis, keyword concordances, and co-occurrence computations (MacWhinney, 2018). Essential for linguistic research, CLAN provides calculations of metrics, such as Mean Length of Utterance (MLU) and Type-Token Ratio (TTR). Its integration with the TalkBank database makes it a powerful resource for studying conversational interaction, language learning, and language disorders.

CLAN's capabilities extend beyond data transcription to include linking transcripts with media and facilitating acoustic analysis (MacWhinney, 2010). It supports automated tagging for part of speech and grammatical structures in multiple languages, making it a versatile tool in linguistic and conversational analysis (MacWhinney, 2023). CLAN is distinguished by its comprehensive documentation, which can be found in an extensive manual (MacWhinney, 2023) that gathers all its available commands, what they return and extra tools to be used beyond CLAN's functions to enhance the feature extraction. This makes CLAN an essential, free available tool for anyone engaged in the detailed study of language and communication.

Linguistic Feature	**CLAN Command**
Overall Lexical Diversity (TTR)	`freq +t*CHI`
Grammar Construction (TTR)	`freq +s"g|CSUBJ" +s"g|COMP" +s"g|CMOD" +s"g|XMOD" +d2 +t*CHI *.cha`
Lemma-based TTR	`freq +t*CHI +sm;*,o% sample.mor.pst`
MLU (Morphemes/Utterances)	`mlu +t*CHI`
MLT (Words/Turns)	`mlt *.cha`
Echolalia	`chip +bINV +cCHI -nb file.cha`
Part of Speech	`mortable +t*CHI +leng *file.cha`

Table 3: Linguistic features and corresponding CLAN commands

Custom Script

Additionally, a custom script was developed for this project to explore specific aspects of language use in children, leveraging various Python packages to handle the complexity and nuances of natural language processing. Key packages included 'nltk' for tokenizing and tagging parts of speech (Documentation, 2023b), and 'SentimentIntensityAnalyzer' for preliminary sentiment analysis, which was particularly useful in identifying emotional word usage (Documentation, 2023a). The 'sklearn' package intended to assess the topic consistency, offering a way to quantify how consistently children stayed on a topic during a conversation (Lab, 2015).

The transformers library, specifically using the 'GPT2LMHeadModel' and 'GPT2Tokenizer' from this package, aimed at computing anomaly scores for each utterance (HuggingFace, 2016) to detect semantic irregularities or idiosyncrasies in the speech of children with ASD.

Initially, the script faced challenges in identifying a significant number of emotional and abstract words, possibly due to the young age of the children and their developing vocabulary. To address this, we expanded the word list based on established research (Van Goozen & Frijda, 1993). However, the frequency of these words remained relatively low in our data set, and therefore, it was decided not to include these features. Another challenge was representing grammatical repetitions and specific words, which initially appeared more frequently as symbols or other non-numerical data that was challenging to quantify and interpret rather than meaningful linguistic data. While these issues were partly resolved, the efficiency of the CLAN software for these tasks was notably superior, especially regarding a variety of grammatical constructions, so it wasn't justified to use the results from the custom script.

One of the script's strengths was its ability to measure "Topic Consistency" in children's speech, providing insights into their conversational engagement. Despite the value of this information, the labor-intensive process of extracting this data for each transcript made it impractical for large-scale analysis.

PHON

PHON is a sophisticated Java application designed to perform detailed phonological analyses (MacWhinney, 2023). It goes beyond the capabilities of CLAN by providing in-depth examinations of sound patterns, such as phoneme variability and repetition patterns, which is critical as children with ASD often process sounds differently than typically developing peers (Cohly et al., 2010). PHON provides valuable insights into the unique auditory processing and sensitivities present in ASD, contributing to a better understanding of these individuals' linguistic profiles.

However, it was impossible to utilize PHON in this research project due to significant obstacles. Datasets from the Eigsti and Nadig projects were incompatible, with several .cha files failing to import. Even for the data that was successfully imported, PHON's analysis instruments did not work as anticipated, leading to inconclusive reports. These limitations prevented in using this tool, but they highlight an area for future investigation. With more time and computational resources to navigate these technical complexities, using phonological analysis through PHON could significantly enhance the understanding of how children use language taking into account the role the auditory system takes in this process.

APPENDIX B

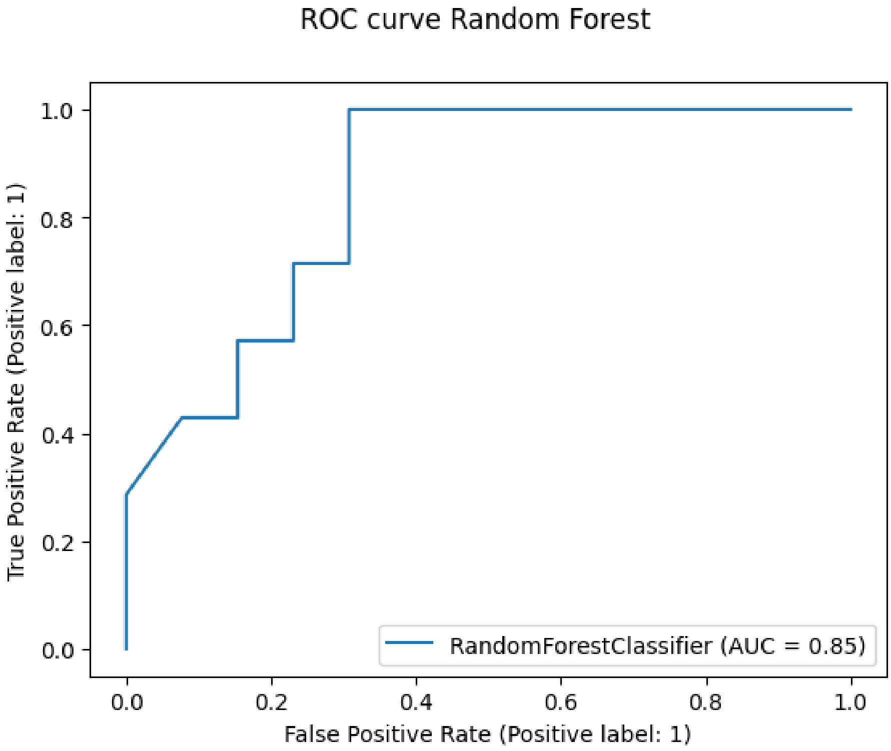


Figure 2: ROC curve for the Random Forest Model

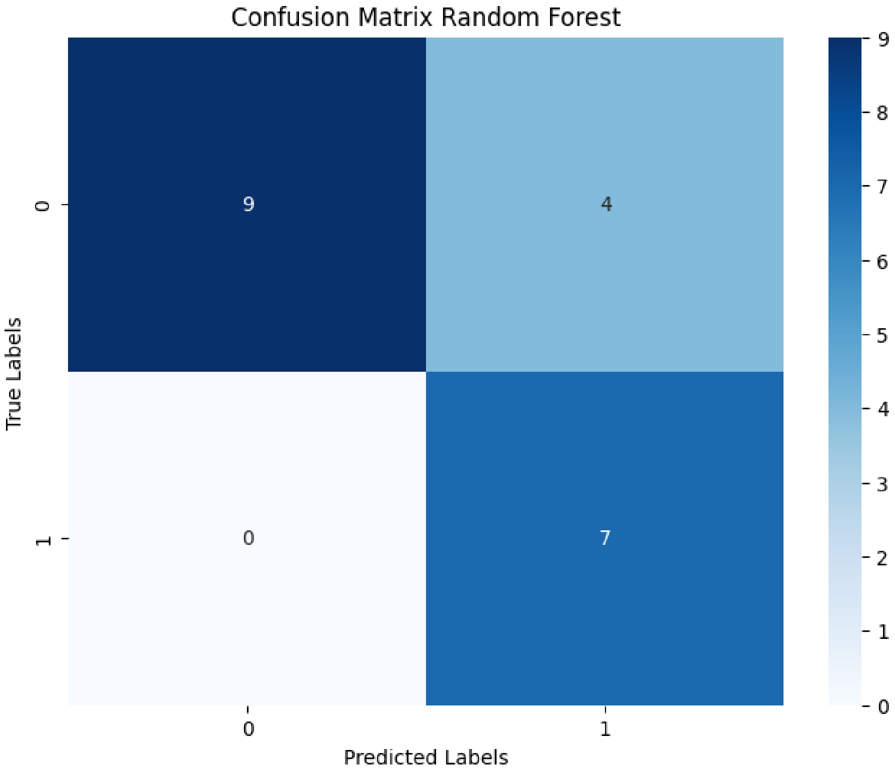


Figure 3: Confusion Matrix for the Random Forest Model

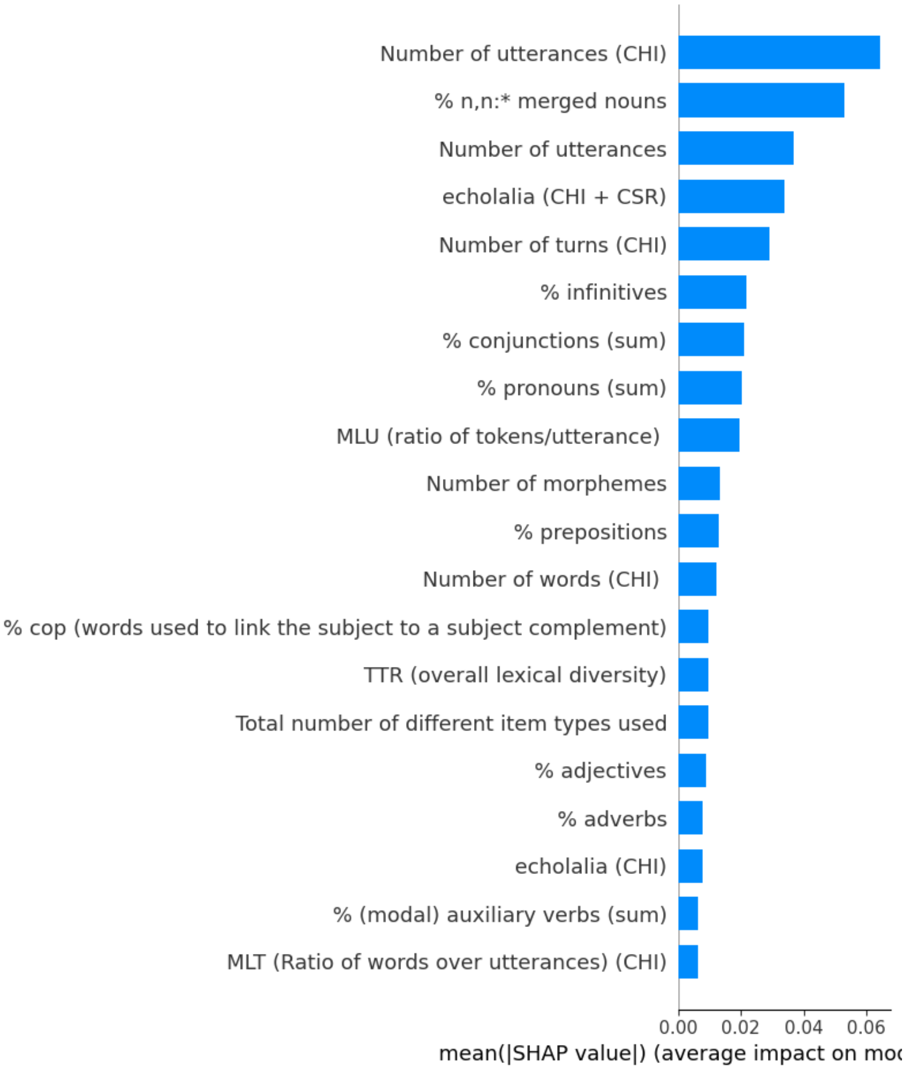


Figure 4: Feature Importance Ranking with SHAP values