Identifying the Most Salient Audio and Language Features for Pediatric Specific Language Impairment Classification

Ronak Chadha

Abstract

Specific language impairment, also known as SLI, is a pediatric language disorder that delays the development of typical speech functions without the influence of other developmental delays or neurological disorders. SLI prevents children from clearly communicating their thoughts or desires with others and can persist throughout their lives if left undiagnosed. With the ability to provide scalable diagnostic services in the comfort of one's home, machine learning solutions offer the potential for an accessible screening method for SLI, enabling a parent or guardian to identify potential markers and consult with a speech and language therapist about clinical actions. To address this opportunity, I developed a machine-learning solution to classify SLI based on audio and language features derived from the Talkbank Collection of the CHILDES dataset. I applied feature selection to identify the most salient features using top-ranked gradient-boosting features, logistic regression coefficients, and mutual information scores. The gradient-boosting classifier outperformed the other two methods, achieving 85% average accuracy, 85% average precision, and 83% average recall. The top features across the three feature selection strategies were the z-score of mean utterance length, age, perplexity of 1-gram SLI, word types to word token ratio, number of nouns followed immediately by a verb, flesch–kincaid score, repetitions, possessives, and the z-score of word errors. Of note, the flesch-kincaid score and perplexity of n-gram sequences, while not new, are relatively understudied features in SLI analysis and would benefit from additional research. Interestingly, prior ML studies have found these features appear in the context of other conditions, such as mild cognitive impairment and dementia [1,2].

Keywords

Specific Language Impairment, Machine Learning, Speech Analysis

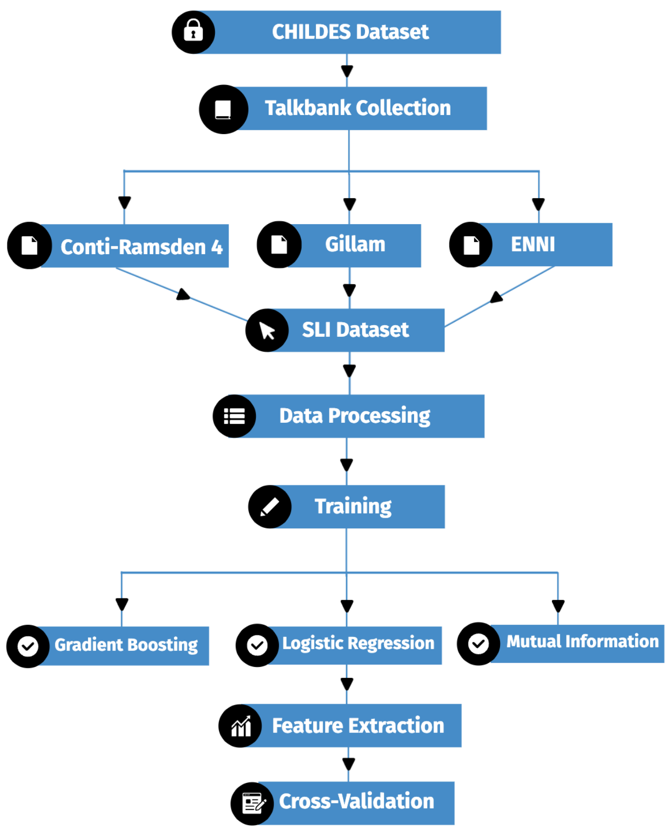
Introduction and Background

Specific Language Impairment, also known as SLI, is a pediatric language disorder that delays the development of typical speech functions without the influence of other developmental delays or neurological disorders [3]. It prevents children from clearly communicating their thoughts or desires with others and can persist throughout their lives if left undiagnosed [4]. Detecting SLI early is essential to help treat and correct it [5].

The current methods of diagnosis require a trained speech-language pathologist to evaluate the child using existing tests, which can include direct observation of the child, interviews, and questionnaires completed by parents, guardians, or teachers, assessments of the child's learning ability, and standardized language tests [6, 7]. These methods are time-consuming, expensive, and often inaccessible, as health insurance does not always cover them. There are currently no machine learning-based solutions available to diagnose specific language impairments. With the ability to provide scalable diagnostic services in the comfort of one's home, machine learning solutions offer a potentially low-stress and accessible screening method for SLI, allowing a guardian to consult with a speech and language therapist about clinical actions.

To address this opportunity, I developed a machine-learning solution to classify SLI based on audio and language features derived from the Talkbank Collection [8] of the CHILDES dataset [9]. The Talkbank Collection consists of various subsets of narrative-based tasks, including Conti-Ramsden 4 [10], Gillam [11], and ENNI [12]. These narrative-based tasks involve children attempting to accomplish a wordless picture task. This choice was partly due to previous research on the subject matter that had indicated its superiority in identifying pediatric SLI [13]. These datasets explore common and unfamiliar aspects of language, such as the number of fillers spoken to the Flesch-Kincaid readability tests [14]. I applied feature selection to identify the most salient features for SLI classification using top-ranked gradient-boosting features, logistic regression coefficients, and mutual information scores. I will discuss the importance of the top-ranked features, which can inform the development of future digital diagnostics and therapeutics for SLI.

Methods



Dataset

The dataset is from the Talkbank Collection [8] of the CHILDES dataset [9]. The Talkbank Collection consists of various subsets of narrative-based tasks, including Conti-Ramsden 4 [10], Gillam [11], and ENNI [12], which involve children attempting to accomplish a wordless picture task. This choice was partly due to previous research on the subject matter that had indicated its superiority in identifying pediatric SLI [13]. These datasets explore both the common and unfamiliar aspects of linguistics, such as the number of fillers spoken to the Flesch-Kincaid readability tests [14].

Data Processing

I pre-processed the data before training and evaluating the model's efficacy. This pre-processing included imputation, upsampling, removing columns or groups of data, and shuffling the dataset. I imputed all of the missing values by calculating the average value of the data points in a column. Additionally, since the number of children with SLI was significantly less than the number of children that were typically developing, I upsampled the data of the children with SLI to match the number of typically developing children. Next, I removed columns that revealed whether the child had SLI or was typically developing and which corpus the data originated from. Lastly, I shuffled the dataset to reduce the chance for the model to develop bias.

Machine Learning Modeling

The data is trained on the gradient boosting classifier classification model. The gradient boosting classifier [15] achieves peak performance by optimizing a model's weights and reducing prediction errors [16]. The classifier builds an initial model based on the training dataset and then builds subsequent models to rectify the errors present in the previous models.

Model Evaluation Metrics

The model was evaluated on three standard metrics [17]: accuracy, precision, and recall. I performed five fold cross-validation, a method in which the data is split evenly into five parts or folds. One fold is used for testing, while the other four folds are used for training. I then calculated each metric's average accuracy, precision, recall, and error.

Feature Selection Strategy

Before feature selection [18], I discovered the classification models that yielded the highest accuracy, precision, and recall values were the gradient-boosting classifier and logistic regression [19]. Using the built-in methods for the feature importance of each classification model and mutual information [20], I trained the gradient-boosting classifier on the top 15 features from each of these classification models and mutual information. I concluded the top 15 features from the gradient boosting classifier model yielded the highest accuracy, precision, and recall.

Results

Table 1a presents the top 15 features of the gradient boosting classifier post-imputation. Table 2a presents the top 15 features of the logistic regression model post-imputation. Table 3a presents the top 15 features of the mutual information quantity post-imputation. Tables 1b, 2b, and 3b for each training method analyze the set of features with the logistic regression raw coefficients. Among the three methods, nine features appear most frequently: z-score of typically developing group's mean length utterance, age, perplexity of 1-gram SLI, word types to word token ratio, number of nouns followed immediately by a verb, flesch–kincaid score, repetitions, possessives, and the z-score of typically developing group's word errors. As determined by the testing on raw data post-imputation, the gradient boosting classifier performed the best without feature selection, cross-validation, and upscaling of data. Once these three methods were implemented, the features selected by the gradient boosting classifier proved to be the best performing as it scored an average accuracy of 85%, average precision of 85%, and average recall of 83% with 15 features, as seen in Figure 1. The model trained with the top 15 features of logistic regression scored around 78% for all three metrics. When the model was trained with the top 15 features for mutual information, it scored around the same, as shown in Figure 1.

Table 1a.

The feature importance of the top 15 features used for prediction according to the gradient boosting classifier.

Feature Name	Feature Importance
Z-score of typically developing group's mean length utterance	0.25872
Word errors	0.14249
age	0.07943
age_years	0.06828
Ratio of raw to inflected verbs	0.05190
Verb utterances	0.04924
Perplexity of 3-gram SLI	0.03323
Perplexity of 1-gram TD	0.02320
Perplexity of 1-gram SLI	0.02255
Perplexity of 2-gram TD	0.02170
Mean Length of Utterance of Morphemes	0.01991
Word Types to Word Token Ratio	0.01991
Perplexity of 2-gram SLI	0.01459
Number of Nouns followed immediately by a verb	0.01388
Flesch-Kincaid Score	0.01385

Table 1b.

According to the logistic regression classifier, the raw coefficients of the top 15 features from the gradient boosting classifier are sorted based on the logistic regression model's coefficient magnitude.

Feature Name	**Raw Coefficients**
Ratio of raw to inflected verbs	1.60436
Mean Length of Utterance of Morphemes	-0.58453
Word errors	0.52501
Verb utterances	-0.42880
Z-score of typically developing group's mean length utterance	-0.15839
Word Types to Word Token Ratio	0.12958
Flesch-Kincaid Score	0.09790
Number of Nouns followed immediately by a verb	0.07145
Perplexity of 1-gram TD	0.07014
age	0.02244
Perplexity of 2-gram TD	0.00401
age_years	0.00187
Perplexity of 2-gram SLI	0.00099
Perplexity of 3-gram SLI	-0.00073
Perplexity of 1-gram SLI	-0.00025

Table 2a.

The importance of the top 15 features used for prediction according to logistic regression.

Feature Name	**Feature Importance**
Sample Z-score using TD group's Number of Verb Utterances	0.81093
Sample Z-score using typically developing group's mean length of utterance	-0.55866
Z-score of typically developing group's word errors	0.42469
Number of Nouns followed immediately by a verb	0.13755
regular_past_ed	-0.13316
regular_3rd_person_s	-0.06930
Number of plurals used	0.05679
3rd. singular nominative pronoun followed by verb	-0.04393
possessive_s	-0.03431
Number of Determinant Pronouns followed by a Noun	0.02248
uncontractible_aux	-0.02237
Mean Length of Utterance of 1st 100 words	-0.01524
Number of “on” prepositions used	0.00649
Repetitions	-0.00194
Index of Productive Syntax Score	-0.00019

Table 2b.

According to the logistic regression classifier, the raw coefficients of the top 15 features from the logistic regression classifier are sorted based on the logistic regression model's coefficient magnitude.

Feature Name	**Raw Coefficients**
Sample Z-score using TD group's Number of Verb Utterances	0.81093
Sample Z-score using typically developing group's mean length of utterance	-0.55866
Z-score of typically developing group's word errors	0.42469
Number of Nouns followed immediately by a verb	0.13755
regular_past_ed	-0.13316
regular_3rd_person_s	-0.06930
Number of plurals used	0.05679
3rd. singular nominative pronoun followed by verb	-0.04393
possessive_s	-0.03431
Number of Determinant Pronouns followed by a Noun	0.02248
uncontractible_aux	-0.02237
Mean Length of Utterance of 1st 100 words	-0.01524
Number of “on” prepositions used	0.00649
Repetitions	-0.00194
Index of Productive Syntax Score	-0.00019

Table 3a.

The feature importance of the top 15 features used for prediction according to the mutual information quantity.

Feature Name	**Feature Importance**
Perplexity of 1-gram SLI	0.03595
Sample Z-score using typically developing group's mean length of utterance	0.02944
Number of “in” prepositions used	0.02810
age	0.02533
Repetitions	0.02444
Number of Determinant Nouns followed by a Personal Pronoun	0.01962
Flesch-Kincaid Score	0.01743
Pronouns followed by Auxillary Verb	0.01567
Number of Do's	0.01495
Average number of syllables per word	0.01284
present_progressive	0.01281
Word Types to Word Token Ratio	0.00943
possessive_s	0.00877
Total Number of Words	0.00863
sex	0.00668

Table 3b.

According to the logistic regression classifier, the raw coefficients of the top 15 features from mutual information are sorted based on the logistic regression model's coefficient magnitude.

Feature Name	**Raw Coefficients**
Sample Z-score using typically developing group's mean length of utterance	0.48281
Flesch-Kincaid Score	-0.31360
Number of Do's	0.24791
Number of “in” prepositions used	-0.12210
Average number of syllables per word	0.05054
sex	-0.03263
Repetitions	0.02738
Word Types to Word Token Ratio	0.02118
Number of Determinant Nouns followed by a Personal Pronoun	-0.02109
possessive_s	-0.01941
Pronouns followed by Auxillary Verb	-0.01693
age	0.00672
Total Number of Words	-0.00133
present_progressive	0.00106
Perplexity of 1-gram SLI	0.00000

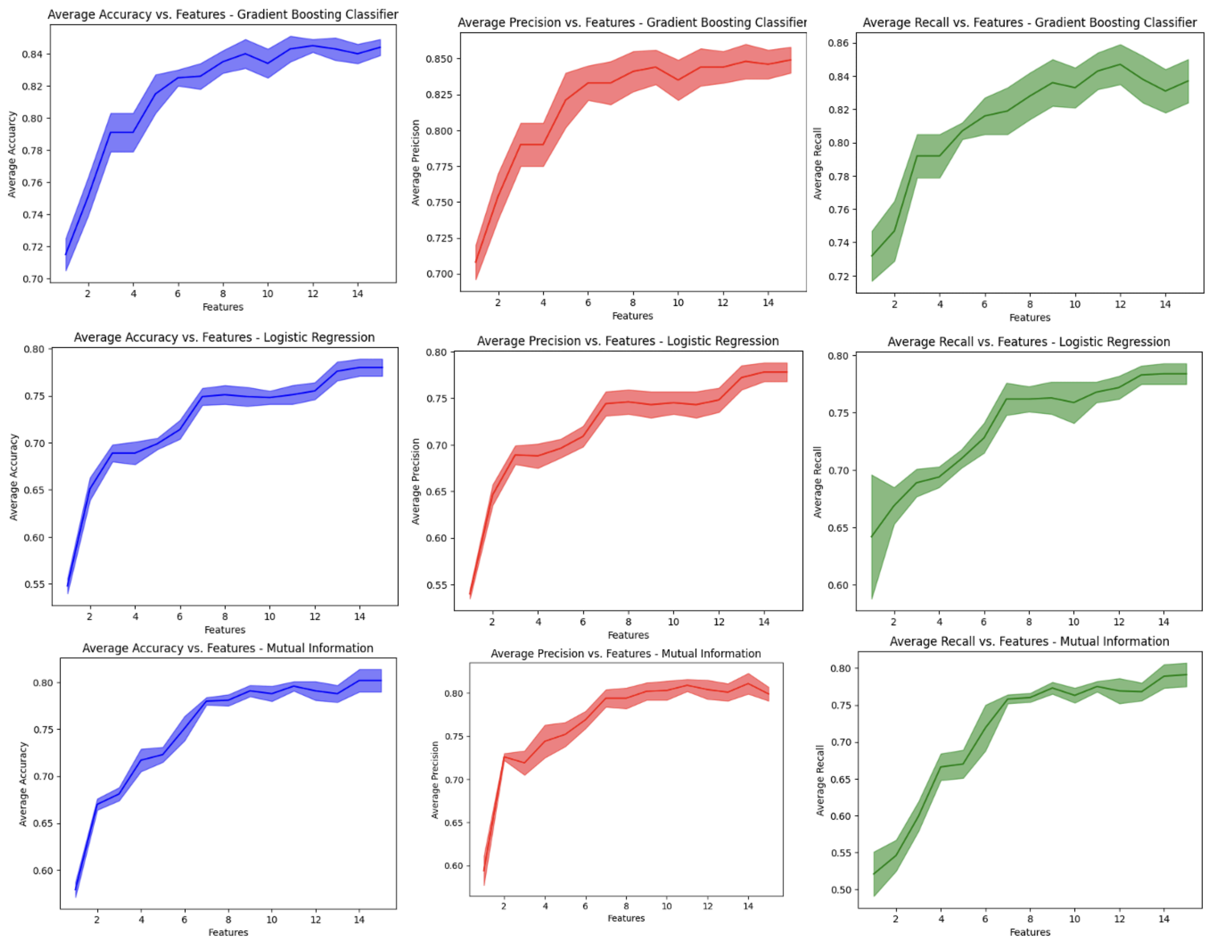


Figure 1. The average accuracy, precision, and recall metrics measured (in %) with the corresponding number of features

Discussion and Conclusion

The top features across the three methods of feature selection were the z-score of typically developing group's mean length utterance, age, perplexity of 1-gram SLI, word types to word token ratio, number of nouns followed immediately by a verb, flesch–kincaid score, repetitions, possessives, and the z-score of typically developing group's word errors. Among the features identified through the three feature selection methods, most are common indicators of SLI, including word errors, verb utterances, and repetitions. [21, 22, 23, 24] The flesch-kincaid score is a common readability test administered to determine the difficulty of a text. The perplexity of n-gram sequences is a measurement used to evaluate the performance of a natural language processing model. It evaluates how well the model predicts the next word in the sequence. The correlation between flesch-kincaid scores and SLI and the perplexity of n-gram sequences and SLI can be further evaluated by researchers, as these are relatively understudied in the context of SLI.

A common feature identified in the gradient boosting classifier model was the perplexity score of n-gram sequences. These n-gram sequences refer to a continuous sequence of words used for language analysis; the perplexity score refers to the model's ability to correctly comprehend a sequence of words [25]. Since the gradient boosting classifier works to continuously iterate on itself to achieve performance gains, it found that analyzing n-gram sequences less than or equal to three words was a strong indicator of whether a pediatric patient had SLI or was typically developing.

A common thread throughout the prevalent features in the logistic regression model was the inclusion of different parts of speech, including plurals, possessives, verbs, and more. The logistic regression model could distinguish between a pediatric patient's potential for SLI or for being typically developing by relying on the usage of each of these parts of speech. Verbs and plurals tended to be used more frequently by typically developing pediatric patients [22, 26].

Limitations

The primary limitation of this project was the limited dataset for both training and testing. I used the Talkbank Collection of the CHILDES dataset, which consists of data from around one thousand unique pediatric patients. The project was also limited in variability, as the dataset consisted of wordless picture tasks and was based on transcripts of children completing these tasks. This limitation prevents the model from analyzing other features, including tongue and lip movements, speech duration, stuttering, and utterance speed.

Future Work

This work can be expanded upon by incorporating a wider range of speaking tasks beyond narrative-based tasks during the data collection phase. Example tasks can include the use of words, conversations, and more. This will open up the possibility of discovering new methods and features that can be used to diagnose SLI. Further, using deep neural networks can enable the model to engineer complex nonlinear features from the dataset. Also, multimodal learning with a multitude of input sources such as audio, text, or video inputs can make this model more accessible by allowing the user to choose an input into the model that is most convenient for them. Finally, using novel text and audio embeddings, as well as the wav2vec and word2vec algorithms, can provide a solution to quantify and process speech.

List of Abbreviations

SLI (Specific Language Impairment), TD (Typically Developing)

References

1. Sporna, A. B. Static versus interactive online resources about dementia: A comparison of readability scores.

2. Cohen, T., & Pakhomov, S. (2020). A tale of two perplexities: sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. *arXiv preprint arXiv:2005.03593*.
3. Bishop, D. V. (2006). What causes specific language impairment in children?. *Current directions in psychological science*, *15*(5), 217-221.
4. Duinmeijer, I. (2013). Persistent problems in SLI: which grammatical problems remain when children grow older. *Linguistics in Amsterdam*, *6*, 28-48.
5. Ebbels, S. H., Van Der Lely, H. K., & Dockrell, J. E. (2007). Intervention for verb argument structure in children with persistent SLI: A randomized control trial.
6. Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences*, *21*(1), 67.
7. Ebbels, S. (2014). Introducing the SLI debate. *International Journal of Language & Communication Disorders*, *49*(4), 377.
8. MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior research methods*, *51*, 1919-1927.
9. MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.
10. Conti-Ramsden, Nicola Botting Zoësimkin, Emma Knox, G. (2001). Follow-up of children attending infant language units: Outcomes at 11 years of age. *International journal of language & communication disorders*, *36*(2), 207-219.
11. Colozzo, P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment.
12. Schneider, P., Hayward, D., & Dubé, R. V. (2006). Storytelling from pictures using the Edmonton narrative norms instrument. *Journal of speech language pathology and audiology*, *30*(4), 224.
13. Rezzonico, S., Chen, X., Cleave, P. L., Greenberg, J., Hipfner-Boucher, K., Johnson, C. J., ... & Girolametto, L. (2015). Oral narratives in monolingual and bilingual preschoolers with SLI. *International Journal of Language & Communication Disorders*, *50*(6), 830-841.
14. Flesch, R. (2007). Flesch-Kincaid readability test. *Retrieved October*, *26*(3), 2007.
15. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937-1967.
16. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.
17. Flach, P. (2019, July). Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9808-9814).
18. Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, *52*(4), 4543-4581.

19. Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, *122*, 56-69.
20. Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, D. (2018, July). Mutual information neural estimation. In *International conference on machine learning* (pp. 531-540). PMLR.
21. Botting, N., & Conti‐Ramsden, G. (2001). Non‐word repetition and language development in children with specific language impairment (SLI). *International Journal of Language & Communication Disorders*, *36*(4), 421-432.
22. Conti-Ramsden, G., & Jones, M. (1997). Verb use in specific language impairment. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1298-1313.
23. Lahey, M., & Edwards, J. (1999). Naming errors of children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(1), 195-205.
24. Bishop, D. V. (1994). Grammatical errors in specific language impairment: Competence or performance limitations?. *Applied Psycholinguistics*, *15*(4), 507-550.
25. Popel, M., & Mareček, D. (2010). Perplexity of n-gram and dependency language models. In *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings 13* (pp. 173-180). Springer Berlin Heidelberg.
26. Oetting, J. B., & Rice, M. L. (1993). Plural acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *36*(6), 1236-1248.
27. O'Keefe, David. (2017, April). Diagnose Specific Language Impairment in Children, Version 6. Retrieved June 23, 2023 from https://www.kaggle.com/datasets/dgokeeffe/specific-language-impairment.