INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Towards an Automated Screening Tool for Developmental Speech and Language Impairments

Jen J. Gong[1], *Maryann Gong*[1], *Dina Levy-Lambert*[1], *Jordan R. Green*[2], *Tiffany P. Hogan*[2], *John V. Guttag*[1]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2] MGH Institute of Health Professions, Charlestown, MA, USA

jengong@mit.edu, mmgong@mit.edu, dinalevy@mit.edu,
jgreen2@mghihp.edu, thogan@mghihp.edu, guttag@mit.edu

Abstract

Approximately 60% of children with speech and language impairments do not receive the intervention they need because their impairment was missed by parents and professionals who lack specialized training. Diagnoses of these disorders require a time-intensive battery of assessments, and these are often only administered after parents, doctors, or teachers show concern.

An automated test could enable more widespread screening for speech and language impairments. To build classification models to distinguish children with speech or language impairments from typically developing children, we use acoustic features describing speech and pause events in story retell tasks. We developed and evaluated our method using two datasets. The smaller dataset contains many children with severe speech or language impairments and few typically developing children. The larger dataset contains primarily typically developing children. In three out of five classification tasks, even after accounting for age, gender, and dataset differences, our models achieve good discrimination performance (AUC > 0.70).

Index Terms: developmental speech and language impairment classification, speech-pause characteristics, machine learning

1. Introduction

In 2012, 44.7% of children ages 3-5 who received services under the U.S. Department of Education Individuals with Disabilities Education Act (IDEA) had "speech or language impairments" listed as one of their disability categories. This incidence exceeded that of developmental delay (37.2 %) and autism (7.8 %) [1]. Difficulty with speech and language in childhood has a negative impact on self-concept, social interactions, academic achievement, and vocational potential [2, 3]. Current identification of children with speech and language impairments relies on parental reporting, primary care physician evaluation, or, after the start of school, the input of teachers. Epidemiologic studies show that approximately 60% of children with speech and language impairments do not receive the intervention they need because their impairment was missed by parents and professionals who lack specialized training [4].

Automated screening tools based on characteristics extracted from speech do not rely on subjective evaluation and are more easily scalable for use by professionals, parents, and experts. These tools could facilitate earlier diagnosis and more extensive and inexpensive monitoring. Earlier diagnosis of these disorders could allow for earlier targeted intervention, and continued monitoring could help clinicians better understand the nature and progression of these disorders.

Towards this end, we used machine learning to develop classification models to distinguish children with speech and language impairments from typically developing children. We extracted features from acoustic signals of children retelling a story to predict whether or not they have a speech or language impairment. In this work, we focused on the utility of acoustic features based on speech and pause events, building on previous work that has shown that features relating to pauses in speech can be used to effectively predict disorders such as childhood apraxia of speech [5, 6]. While language-based features would almost certainly have predictive utility, the high variability of children's speech makes developing accurate automatic speech recognition tools challenging, and existing adult-based speech recognition algorithms cannot easily be applied to children [7, 8]. In addition, these speech-pause characteristics are simple to compute and are robust to different recording environments and speaker characteristics.

We developed and tested our methods on two sets of speech samples. The first was obtained from a population of children who were typically developing (TD) or were diagnosed with idiopathic speech and language impairments in the following categories: Childhood Apraxia of Speech (CAS), Specific Language Impairment (SLI), Speech Sound Disorder (SSD), or a comorbidity (CAS/SLI). The second dataset was obtained from a much larger population of children who were primarily typically developing, with only a few members identified as having language impairments.

These disorders can be categorized into speech disorders (SSD and CAS) and language disorders (SLI). Children with SSD and CAS have difficulty producing speech sounds correctly [9]. For children with CAS, this is a result of a motor planning impairment, rather than as a result of physical impairment [10]. SLI differs from CAS and SSD because it is a language impairment, rather than a speech disorder. While children with language disorders may not have trouble producing correct speech sounds, they may have difficulty understanding others or expressing themselves [11]. Our dataset also contains children with a comorbid condition (CAS/SLI).

In this paper, we demonstrate that even after adjusting for the effects of age, gender, and dataset, features capturing speech and pause characteristics of a story retell task have predictive value in distinguishing children with impairments from children who are typically developing.

http://dx.doi.org/10.21437/Interspeech.2016-549

2. Related Work

Developing an automated screening tool for developmental speech and language impairments is a challenging task because speech in children is highly variable; children's voices differ through development, and their language also differs extensively across age groups. Lastly, children's pronunciation of words and speaking rate are not the same as those of adults; thus, many existing adult-based speech recognition algorithms may not transfer well to children [7, 8]. The high variability in the typical development of children's speech means that it can be much more difficult to differentiate children with impairments from those who are typically developing. Systems such as PEAKS [12] and LENA [13] have been developed to further these ends. [14] presents an automated tool that detects speech errors that commonly occur in children with childhood apraxia of speech (CAS).

Participants in the Autism Sub-Challenge of the Interspeech 2013 Computational Paralinguistic Challenge developed models for two tasks: 1) distinguishing typically developing children from atypically developing children, and 2) diagnosing the disorder type, where the disorders were Pervasive Developmental Disorder (PDD), Pervasive Developmental Disorder Non-Otherwise Specified (PDD-NOS), and specific language impairment (Dysphasia). The dataset had 2,542 speech recordings from 99 children, where 12 had diagnoses of PDD, 10 had diagnoses of PDD-NOS, and 13 had specific language impairment. The children were asked to imitate sentences of different modalities and intonations. These examples were stratified by age and gender for model development and testing. On the task of classifying typical versus disordered samples, the baseline performance (reported unweighted average recall) using a linear SVM was 90.7% on the test set [15]. The metric of unweighted average recall obscures performance on the individual classes, and neither the baseline nor the challenge winner [16] accounted for the potential confounding effects of age and gender. Accounting for these factors is especially important since speech is widely variable across ages even in children who are typically developing.

We develop and test classifiers for distinguishing several speech- and language-impaired subgroups from typically developing. We use linear regression models to account for age, gender, and dataset confounders in each of the features, and utilize the residuals in our classification model. Our audio recordings were obtained on a story retell task, rather than a sentence repetition task. This task may enable us to capture more language-related characteristics, even using acoustic features.

3. Data

3.1. D1: Speech and language impairments

Speech recordings were collected from 53 children who were typically developing (TD) or had diagnoses of childhood apraxia of speech (CAS), specific language impairment (SLI), speech sound disorder (SSD), dyslexia, or a comorbidity (CAS/SLI). Samples from children with dyslexia were removed from the analysis because of the small sample size (4 children). Details on the children used in the analysis are shown in Table 1. Speech samples were segmented to remove dialogue from the interviewer and then concatenated together. The children ranged in age from 4 years, 7 months to 17 years, 8 months. The average length of the audio samples was 30.1 seconds (standard deviation = 10.6 seconds).

All children had normal nonverbal IQ (no cognitive impairments). Diagnoses for the children were determined based on a battery of tests administered by speech and language professionals with years of experience and specialized training. Emphasis in selecting this subset of children was placed on purity of the diagnosis, rather than on sample size. Frequently, speech and language impairments are comorbid with other developmental conditions. In the rest of the paper, we will refer to this dataset as D1. We defined language impairment in D1 to be a diagnosis of SLI, and a speech impairment to be a diagnosis of SSD, CAS, or CAS/SLI. CAS/SLI was grouped with the speech impairments in this analysis because their primary disorder was identified as CAS.

Table 1: Statistics on children in each diagnosis group in D1.

Diagnosis	Number	Mean Age in Months (Std)	Gender (Male)
Typical	10	137.1 (47.0)	9
CAS	9	123.7 (45.8)	9
CAS/SLI	11	124.2 (26.1)	11
SLI	10	119.0 (14.5)	4
SSD	9	100.3(13.7)	4

3.2. D2: Typically developing children

Speech recordings were collected from children ages 4 years, 1 month to 9 years, 10 months. The story told to the children in this dataset was longer and more detailed than the one told to the children in D1. The average length of the speech samples in this dataset was 50.4 seconds, with a standard deviation of 41.8 seconds. Statistics on this dataset are shown in Table 2. Very few children were impaired in this sample; nineteen of the 201 children were classified as having a language impairment, and none of the children were diagnosed with a speech impairment. We will refer to this dataset as D2.

Table 2: Statistics on children in each diagnosis group in D2.

Diagnosis	Number	Mean Age in Months (Std)	Gender (Male)
Typical	182	79.1 (18.9)	90
Language impairment	19	74.6 (8.3)	7

3.3. Differences between datasets

In this work, we chose to pool the two datasets, despite their differences. While both sets of children were asked to retell a story, the story itself was different in the two datasets. Children in D1 were asked to tell a much shorter story than those in D2 (average length of 30.1 seconds in D1 vs. 50.4 seconds in D2). In addition, the children in D2 were younger than those in D1 and had a smaller variation in age distribution. The gender make up of the two datasets also differed: most of the children in D1 were male, while D2 was more evenly distributed between genders (97 male, 104 female). Finally, D1 primarily contained children with severe speech or language impairments (only 10 TD children), while D2 primarily contained TD children and only 19 children with language impairments. The children with language impairments in D2 also had much less severe conditions than those in D1.

We chose to pool these datasets rather than analyzing each one separately because there are too few TD children in D1, and too few well-characterized impairments in D2. However, in order to do a meaningful analysis of whether speech-pause features are predictive of impairments (rather than the confounding factors of age, gender, and dataset), we had to account for these

confounders in each of the features. This process is described in Section 4.2.

4. Methods

4.1. Signal Processing and Feature Extraction

We preprocessed the speech samples ($F_s = 44.1$ KHz) by manually segmenting out portions of interaction with the examiner, silence at the beginning and ends of the sample, and speech irrelevant to the story retell task in both datasets. The signals were de-trended by removing the mean, and then low-pass filtered ($F_c = 5$ KHz) to remove high-frequency noise. We extracted the amplitude envelope from the signals. The envelope was calculated by squaring the signal, decimating the rectified signal by a factor of 20, and then applying a low-pass filter with a cutoff frequency of 30 Hz.

We used a speech-pause thresholding algorithm, similar to the one proposed in [17], to segment the signal into speech events and pause events. These events were determined by specifying three thresholds: 1) minimum speech event time, 2) minimum pause event time, and 3) a minimum amplitude threshold. In addition to these thresholds, we applied a pitch detector (between 100 Hz and 1200 Hz) to segment out noise artifacts (e.g., hitting the microphone, page turns) while still preserving all voiced activity. We also included a separate identifier for "long" pauses vs. "short" pauses with the goal of separating language-related pauses from speech-related pauses.

We used thresholds of 50 ms and 150 ms to determine continuous speech and pause events, respectively. These thresholds were chosen from auditory and visual inspection of the acoustic signals to prevent short sounds from being counted as speech, and unvoiced utterances from being counted as pauses. Segments identified as pauses that were shorter than 150 ms were included in the surrounding speech event. Similarly, segments identified as speech events that were shorter than 50 ms were included in the surrounding pause event. Lastly, we used a minimum amplitude threshold for speech events. This amplitude threshold was based on the mean and standard deviation of a segment of noise from the signal. Finally, pause events that were longer than 1 second were considered long pauses; all others were considered short pauses.

We extracted 13 features from the speech and pause events: the mean (μ), standard deviation (σ), and coefficient of variation (CV) statistics of the speech event durations and short pause event durations, the number of speech events (n_{speech}), number of short pause events (n_{shortp}), and number of long pause events (n_{longp}) (all normalized by the duration of the signal T), the fraction of time corresponding to speech (T_{speech}/T), the mean and standard deviation of long pause event durations (μ_{longp} and σ_{longp}), and the coefficient of variation ratio (CVR) between the short pause durations and speech durations. The CVR was identified in [5] as a diagnostic marker for CAS.

4.2. Correcting for Confounders

As discussed earlier, we pooled data from the two datasets, despite the fact that these datasets differ in several aspects. To determine if the features of interest were predictive of speech or language impairments, we first adjusted each predictor for the confounds of age, gender, and dataset. Otherwise, these features result in misleadingly good results by using this confounding information. In addition, the model weights would capture the predictive effects of age or gender rather than the features themselves, making the weights less informative. We adjusted for confounders using a two-step process. In the first, we fit linear regression models using the dataset as the independent variable and each speech-pause feature as the dependent variable. We then fit a second linear regression separately for each dataset with age and gender as independent variables to the residuals from the first step. These regression models were fit using only the TD samples from the two datasets. We transformed skewed features using the Box-Cox transformation [18] before fitting the regression model. Residuals after the two-step adjustment were used as the new predictors in our classification model. We developed separate regression models for the two data sets between age and each of the predictors because the age distributions are different, and fitting a single linear regression to both of the data sets did not remove significant correlations with age in the smaller dataset (D1).

To evaluate how well this process adjusted for the confounders, we evaluated the Pearson correlation between each feature and age before and after the correction. We also did hypothesis tests of the differences in the means of each feature between males and females, as well as between D1 and D2. We used a Wilcoxon rank-sum test for these hypothesis tests because the number of examples in some categories is small (e.g., 10 TD in D1) and the Wilcoxon rank-sum test is nonparametric [19].

4.3. Model Development

In all of our experiments, we used a binary classifier to distinguish TD children (-1) from children with impairments (+1). We used L2-regularized logistic regression to predict the probability of having an impairment (Equation 1). We used a linear classifier for interpretability, and we used L2-regularization to prevent overfitting.

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \log\left(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}\right). \qquad (1)$$

We used 5-fold cross-validation on the training set to select the best value for C. We used a fixed asymmetric cost parameter equal to the class imbalance so that misclassification of impairments was weighted more heavily than misclassification of typically developing. We searched for C in the range 10^{-3} to 10^{2} in powers of 10. Features were Z-score normalized based on the training set. All models were trained using the LIBLINEAR implementation of L2-regularized logistic regression in Scikit-learn [20, 21].

4.4. Evaluation

We used a leave-one-out (LOO) holdout procedure to evaluate our methods. We chose LOO because the number of positive examples in our tasks ranged from as few as 10 to as many as 58, and LOO maximizes the number of training examples. We then calculated the estimated positive predictive value (PPV), recall, specificity, and area under the Receiver Operating Characteristic curve (AUC) [22].

5. Results

In this section, we present our experimental results for adjusting for confounders and predicting impairments vs. TD (from both datasets). We use the following notation for the subpopulations we consider: 1) language impairments: Imp^L, 2) speech impairments: Imp^S, 3) all impairments: Imp, and 4) typically developing: TD. To specify impairments or TD from a specific dataset, we use a subscript (e.g., TD_{D1} denotes typically developing children from D1).

5.1. Adjusting for confounders

We adjusted for confounders using the two-step process described in Section 4.2. Before taking the residuals, there were four variables (μ_{pause}, σ_{pause}, CV_{pause}, n_{longp}/T) with significant differences (p-value < 0.05) between the TD children in D1 vs. D2 , and three variables (T_{speech}/T, μ_{speech}, n_{speech}/T) with significant differences between the TD gender groups. Two variables (σ_{speech}, $\text{CV}_{\text{speech}}$) were significantly correlated with age in D1 and eight (n_{shortp}/T, n_{speech}/T, μ_{shortp}, μ_{speech}, σ_{shortp}, σ_{speech}, $\text{CV}_{\text{shortp}}$, CVR) were significantly correlated with age in D2. Taking the residuals after adjusting for age, gender, and dataset removed all significant correlations with age (p-value = 1) and all significant differences in the means of the features between gender groups (p-value > 0.4) or between datasets (p-value > 0.5).

We evaluated the coefficient of determination (R^2) of the regression models from the two steps. These values are shown in Figures 1 and 2. Figure 1 shows that dataset explains only a small amount of the variation in the features. However, removing these confounding effects is important in the interpretation of the importance of features in our final model. Figure 2 shows that age and gender explain much more of the variation in the features in D1 compared to D2, perhaps because D1 has only 10 TD children, whereas D2 has 182.

5.2. Predicting Impairments

We evaluated our method on several tasks: TD (from both datasets) vs. 1) Imp_{D1}, 2) Imp, 3) Imp^L_{D1}, 4) Imp^L, and 5) Imp^S_{D1}.The estimated AUC, positive predictive value (PPV), sensitivity, and specificity are shown in Table 3. We considered tasks with and without the impairments from D2 because the impairments from D2 are much less severe than those from D1. All TD samples from both datasets were used for all tasks.

These results show that our performance is better on tasks where impairments from D2 are not included (AUC of 0.79 for Imp_{D1} vs. AUC of 0.68 for Imp). One reason why this might be is that the impairments in D1 are much more severe than those in D2. In addition, our performance on predicting speech impairments (Imp^S_{D1}) exceeded the performance on predicting language impairments (Imp^L_{D1} or Imp^L) in PPV and Sensitivity. This is not surprising, since we are using only speech-pause characteristics and no language-based features.

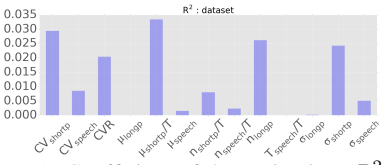

Figure 1: Step 1. Coefficient of determination (R^2) scores for linear regression models with dataset as a predictor.

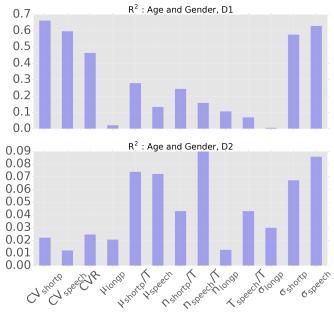

Figure 2: Step 2. Coefficient of determination (R^2) scores for linear regression models with dataset residuals using age and gender as predictors for D1 (top) and D2 (bottom).

Table 3: Estimated AUC, Positive Predictive Value (PPV), Sensitivity, and Specificity results for LOO classification of typically developing vs. different impairments.

Impairment	N	n	AUC	PPV	Sensitivity	Specificity
Imp_{D1}	231	39	0.7879	0.4444	0.7179	0.8177
Imp	250	58	0.6817	0.3563	0.5345	0.7083
Imp^L_{D1}	202	10	0.8625	0.2308	0.6000	0.8958
Imp^L	221	29	0.6212	0.1842	0.4828	0.6771
Imp^S_{D1}	221	29	0.7437	0.2969	0.6552	0.7656

5.3. Feature weights

The average normalized feature weights for the five classification tasks we considered are shown in Figure 3. There are notable differences in the feature weights for TD vs. Imp^L compared to TD vs. Imp^L_{D1}. For example, the number of short pauses normalized by duration was more indicative of Imp^L (language impairments from both datasets) than of Imp^L_{D1} (language impairments from only dataset 1). This is similarly true for the long pauses. This result may hint at the differing natures of the tasks; because the story in D2 was longer and more complex than D1, the children may have paused to think about their responses more often than the children in D1. Finally, the CVR had a higher weight for speech impairments than for language impairments. This supports previous work that has shown that the CVR could be used as a diagnostic marker for CAS [5].

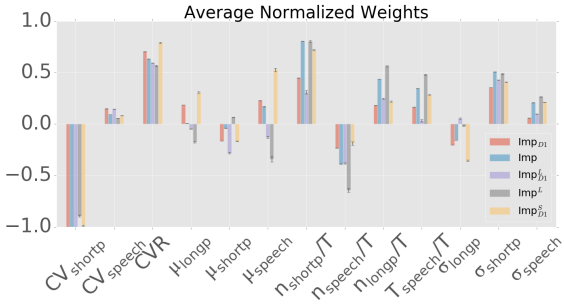

Figure 3: Average normalized feature weights with standard error across LOO models.

6. Conclusions & Discussion

In this paper, we demonstrate the utility of speech-pause features in screening children with speech and language impairments. We adjusted features to remove confounding effects of age, gender, and dataset and demonstrate good discriminative performance on the task of distinguishing speech and language impairments from typically developing. Future work will further investigate the best amplitude and timing thresholds for speech events, short pauses, and long pauses. In addition, although we did adjust for confounding effects, the age and task differences across the datasets may still affect the analysis. Future work will investigate how we can better adjust for these effects by segmenting speech from D2 into shorter utterances like those in D1.

7. Acknowledgements

This research was supported by the University of Nebraska Health Research Consortium (Co-PIs: Hogan & Green), the National Institutes of Health, R03 DC9667 (PI: Hogan), the National Science Foundation Graduate Research Fellowship under Grant No. 1122374, and Quanta Computer Inc.

8. References

[1] U.S. Department of Education, "36th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act, 2014," 2014, http://www2.ed.gov/about/reports/annual/osep/2014/parts-b-c/36th-idea-arc.pdf.

[2] J. Bird, D. V. Bishop, and N. Freeman, "Phonological awareness and literacy development in children with expressive phonological impairments," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 2, pp. 446–462, 1995.

[3] G. Conti-Ramsden and N. Botting, "Social difficulties and victimization in children with SLI at 11 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 1, pp. 145–161, 2004.

[4] J. B. Tomblin, N. L. Records, P. Buckwalter, X. Zhang, E. Smith, and M. O'Brien, "Prevalence of specific language impairment in kindergarten children," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 6, pp. 1245–1260, 1997.

[5] L. D. Shriberg, T. F. Campbell, H. B. Karlsson, R. L. Brown, J. L. McSweeny, and C. J. Nadler, "A diagnostic marker for childhood apraxia of speech: The lexical stress ratio," *Clinical Linguistics & Phonetics*, vol. 17, no. 7, pp. 549–574, 2003.

[6] J.-P. Hosom, L. Shriberg, and J. R. Green, "Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (asr) methods," *Journal of Medical Speech-Language Pathology*, vol. 12, no. 4, p. 167, 2004.

[7] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, Oct. 2007.

[8] A. Hagen, B. Pellom, and K. Hacioglu, "Generating synthetic children's acoustic models from adult models," in *NAACL-Short '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, May 2009.

[9] ASHA. Speech sound disorders-articulation and phonology. [Online]. Available: http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935321&section=Overview

[10] ——. Childhood apraxia of speech. [Online]. Available: http://www.asha.org/public/speech/disorders/ChildhoodApraxia/

[11] ——. Preschool Language Disorders. [Online]. Available: http://www.asha.org/public/speech/disorders/Preschool-Language-Disorders/

[12] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "Peaks–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[13] J. R. Dykstra, M. G. Sabatos-DeVito, D. W. Irvin, B. A. Boyd, K. A. Hume, and S. L. Odom, "Using the language environment analysis (lena) system in preschool classrooms with children with autism spectrum disorders," *Autism*, vol. 17, no. 5, pp. 582–594, 2013.

[14] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "Tabby Talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Communication*, vol. 70, pp. 49–64, Jun. 2015.

[15] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Interspeech*, 2013.

[16] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders." in *Interspeech*, 2013.

[17] J. R. Green, D. R. Beukelman, and L. J. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech," *Journal of Medical Speech-Language Pathology*, vol. 12, no. 4, p. 149, 2004.

[18] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211–252, 1964.

[19] E. Whitley and J. Ball, "Statistics review 6: Nonparametric methods," *Critical Care*, vol. 6, pp. 509–513, 2002.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[22] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.