Applications of Machine Learning for Online Learning Systems towards Children with Speech Disorders

Amr Jadi
Department of Computer Science and Information
College of Computer Science and Engineering,
University of Ha'il, Ha'il, Saudi Arabia

Ali Alzahrani
Computer Science Department
Collage of Computer and Information Systems
Islamic University of Madinah,Madinah, Saudi Arabia

Abstract
Specific Language Impairment is one of the serious disorders that interferes with spontaneous communication skills in children. Children suffering from this disorder may have reading, speaking, or listening impairments, and such type of disorders are also termed Autism Speech Disorder (ASD) in medical terminology. The aim of the article is to define specific language impairment in children and the problems it can cause. The different methods adopted by speech pathologists to diagnose language impairment. Finally implementing machine learning models to automate the process and help speech pathologists and pediatricians/ in diagnosing the specific language impairment.
Keywords:
Specific Language Impairment, ASD, automate Logistic Regression, Random Forests.

I. INTRODUCTION

THE specific language impairment (SLI) is considered as the common type of special educational needs (SENs). This interferes with the development of language skills in children, who otherwise have no other physical or mental disabilities. According to Broomfield and Dodd's survey, 6.4% of normal children in the UK had a speech disorder in the absence of any other impairments [1]. Children with language impairment are the risk of various difficulties such as emotional and social difficulties, reading and writing difficulties which in turn result in lower academic achievements. This language impairment arising in children during their childhood can have an impact through their adolescence and adult life. The symptoms of specific language impairment are generally observed by parents or teachers in pre-school aged children. A child with language impairment often struggles to learn and frame grammatically correct words and sentences using correct syllables and grammar. This kind of ailment ends up with disorganized in speech and writing behavior. The children with such disorder find difficulty in using the right words to express an opinion or thought and is considered late talker.

However, a clear distinction needs to be made between language impairment and language delay. In this linguistically pluralistic world, there is a problem of overidentification as different populations have different linguistic structures [2]. However, the children learn to grasp words and sounds based on what is taught by parents and from the surrounding environment.

As stated, children with language impairment find it hard to express thought, and their speech is mainly characterized by the mispronunciation of words. This inability to express thought through speech causes children with speech disorder to withdraw socially, become frustrated due to their inability to remember and use the words needed to make themselves understood while expressing their thoughts.

According to research, it is observed that 8% of boys and 6% of girls of the same age group are affected with language impairment showing that a greater number of boys are likely to be affected with the language impairment [3]. The same levels of gender discrepancy were also found in another research.

II. BACKGROUND REVIEW

To carry on the work a lot of research needs to be carried out and there is a necessity for analyzing the existing methods which are implemented by the different authors. This section provides the information and the implementation of the algorithms done by the various authors by using the Machine Learning Technique. The author states that dyslexia, according to the DSM-V, is a neurologically based learning impairment. The World Federation of Neurology says that the disorder happens in adolescents who, don't have a traditional education process, do not achieve literacy, typing, or phonological awareness that are consistent with their cognitive prowess. The quality of early identification in the treatment of dyslexia is critical. As clinical diagnosis processes are costly and need expert guidance, pupils are frequently inadequately. They want citizens to realize whether they have dyslexia as quick as practicable. To do this, researchers developed a virtual simulation that captures wide support for web engagement metrics to screen dyslexia in Language. This can be achieved with machine learning algorithms, they are in the Gaussian Support Vector Machine, the binary classifier of the library for support vector machines (LIBSVM) is used (SVM). Therefore, with help of this algorithm, the speech disorder

Manuscript received August 5, 2022
Manuscript revised August 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.8.7

can be identified in a pre-defined manner [4].

The researcher explains that the Individuals with Disabilities Education Act (IDEA) educates almost 7,000,000 kids with disabilities in the country's school systems. The proportion of children is eligible for special education services because they have poor reading comprehension skills, which for several children is a precursor to future educational impairments and/or language disability in primary school. Observational research of young expressive language impairments reveals that they are predisposed to literacy, typing, and arithmetic difficulties in the latter. Local child characteristics have the explicative ability for early diagnosis of specific language impairment in infants, potentially leading to effective treatment in the education at the early levels. To determine variables that may be directly important to physicians' recognition practices, researchers used machine-learning technology to find the most valid, going to define the behaviour of children that distinguished those with medically recognized communication difficulties from their typically developing. The proportion of children diagnosed with the main speech impairment at school attendance will tend to have considerably reduced linguistic competence over age, struggle with kindergarten preparedness, and struggle to read and understand, the latter due partly to its impact on elevated English skills. Therefore, with the help of machine learning children with a speech disorder can be easily identified [5].

The author tries to explain that because of the severe lack of speech and language pathologists (SLPs), computer-assisted speech therapy is gaining popularity (CAST). Several CAST products are designed specifically for kids with speech problems. Speech Training, Assessment, and Remediation (STAR), Vocaliza, Speech Assessment and Interactive Learning System (SAILS), and Phoneme Factory Sound Sorter are just a few examples (PFSS). The majority of such technologies, on the other hand, are focused on the computerized provision of special education cues, allowing psychotherapy to be guided personally by the kid. All of these computerized disrupted voice analysis methods, however, are still not accurate enough to be employed medically. The large spectrum of speech disorders and criteria, as well as the scarcity of suitably large disturbed language datasets, time progress difficult. Furthermore, suitable for older discourse, kid talk has more inter-and intra-speaker variation, making it challenging to comprehend. The absence of taking cognizance of dysfunctional verbal vocabulary, the need for efficacious speaker methods for generating structures to retrieve child portions from audiotapes of the verbal therapeutic relationship, and restrictions in automatic speech recognition of children's speech are all major interests in implementing a robust algorithm to quickly identify adolescence SSDs. Therefore, these are disorders faced by children at an early age and they can be identified as quickly as possible [6].

The researchers explore that the DSM-V divides speech disorders into five categories. Linguistic, voice quality, early life proficiency, verbal interaction, autistic, and other miscommunications are all examples of cognitive disabilities. In the settings of residences, schools, and other socialization, the DSM-5 defines characteristics, frequency, and practical influence in a kid's interaction. The ability to efficiently communicate together is referred to as a conversation. Oral, visual, textual, and other creative methods are used to communicate with people. To excel in any anthropogenic activities, communication skills are important. The three types of cognitive disabilities are used in this project. Normal Children (NC), Phonological Disorder (SSD), and Speech-language Disorder (SLD) are the three categories. There at age of four, NC will struggle to say most vocalizations. NC will pronounce all of the vocalizations accurately by the age of, except for a few more difficult noises. At the age of 9, NC begins to talk with more difficult vocalizations. At the age of five, NC can accurately identify the letters g, k, r, and two consonants. SSD youngsters, on the other hand, intentionally misspelled such symbols and sounds, talk differently, and leave things altogether. SSD starts using improper natural speech when they are 9 old and persists until they receive appropriate services. Speaking smoothness and flowing are the most common issues for SLD youngsters. Pupils learn via their parents, instructors, relatives, and their surroundings as they get older. At the age of five, the majority of youngsters can readily acquire and know the language. SLD children, on the other hand, have trouble pronouncing certain languages correctly. Therefore, there are variations in children suffering from a speech disorder. At a certain age, each disorder has certain symptoms and effects on kids [7].

The concept illustrates that speech problems are broad phrase that encompasses a variety of talents and characteristics resulting from a problem with pronunciation, fluidity, and/or voice. As a result, the quantity, pace, or clarity of the patient's speech is abnormal, possibly affecting the talker's capacity to be comprehended. Apraxia of speech is amongst the most frequent language disorders. It can be inherited from a parent at any age and time of a neurological illness or injury. Dysarthric voice is generally triggered by congenital brain asphyxiation, which prevents proper monologue maturation. Cerebral palsy is one of the most frequent causes of communication disabilities in babies born in Europe. Furthermore, due to the complexity of dysarthria and the participation of numerous components of the language supply chain, there is a great deal of variation between dysarthria speakers. Although typical procedures are inadequate for persons with disabilities, automatic speech recognition (ASR), or the technique by which a system recognizes and acts of the spoken words, is an open call in the accessible software industry. As a consequence, these users encounter miscommunication that can lead to isolation, and they are now further isolated by the latest influx of human-computer interaction that is becoming highly integrated into an ordinary routine and is therefore not resilient to an atypical speech. Therefore, with help of ASR speech disorders can be recognized at an early age [8]. Finally, from the research, it is observed that the various ML algorithms produce different accuracies based on the model building. Most of the authors used Logistic Regression, SVM and Random Forest algorithms to get good results. Due to some reasons, these models provided somewhat score values based on some conditions.

III. PROPOSED WORK

Based on the drawbacks that are observed in the existing methods. This paper tries to provide the issues and try to overcome the drawbacks which are seen in the existing models.

This paperwork discusses the implementation of machine learning models such as Logistic Regression and Random Forests to automate and detect the Specific Language Impairment and Tourette's disorder in children and help speech-language pathologists to diagnose the specific language impairment effectively. The proposed algorithms try to predict the speed disorders of children.

A. Logistic Regression

The logistic regression is a classifier model used to predict the binary outcome variable. For modelling, logistic regression is chosen initially, since the aim is to identify whether a child has a specific language impairment or not, which is binary. The logistic regression is essentially a linear function but with a sigmoid function that produces an output that is either 0 or 1 [9]. This feature of logistic regression is useful for the classification of binary classes.

For logistic regression, the hypothesis can be defined as

$$h_\theta(x) = g(\theta^T x) \tag{1}$$

where 'g' the sigmoid function.

The sigmoid function is represented as

$$g(z) = \frac{1}{1+ e^{-z}} \tag{2}$$

where z is a real number.

Now the hypothesis or the logistic regression written when the sigmoid hypothesis is inserted in the sigmoid function obtained as.

$$h_\theta(x) = \frac{1}{1+ e^{-\theta^T x}} \tag{3}$$

In the case of linear regression, the sum of squared error is minimized,

$$SSE = \sum(y_i - \widehat{y_i})^2 \tag{4}$$

When the above equation is minimized for logistic regression, it turns out that the Eq. iv) is a non-convx function.

To solve this function, the log likelihod action which when maximized is a convex function.

$$Log\ likelihood\ = \sum[y_i \ln \widehat{y_i} + (1 - y_i)\ \ln(1 - \widehat{y_i})] \tag{5}$$

Since the log (1) the value of the function is 0. As 'x' increases greater than 1, the log(x) value is going to be above zero, and when the value of 'x' is below 1 it is going to be negative. Since, between 0 and 1, the maximum value log $(\widehat{y_i})$ can take is 0 since it reaches 0 when the value of y is 1. Therefore, $1 \ln \widehat{y_i}$ will maximize when $\widehat{y_i}$ approaches 1. The logistic regression outputs probability, which can be converted to either 0 or 1. This feature can be useful to determine whether a child has a specific language impairment or not.

Since the number of samples and the information about the classes available in the dataset is not enough to train a machine learning model that can produce a decent output, the data is oversampled using Synthetic Minority Oversampling Technique (SMOTE) to obtain more balanced data. Since the data contains numerical information in the independent variables, the data is standardized using the regular standardization method. The standardization process involves calculating the mean and standard deviation and dividing the difference between the observed sample and the mean with the standard deviation. This ensures that all the values of a feature are set on the same scale before modelling.

A threshold value of 0.05 for probability value is set and the statistically insignificant features that do not contribute to the model are identified and removed using the p-value obtained from the Logit model. This method ensures that only relevant features are included in the model. This is the main advantage of automating the process.

B. Random Forest

The random forest is an ensemble technique that builds different decision trees using different samples to take a majority vote to solve the classification problem.

The random forest algorithm uses the bagging technique to obtain random samples from the original data to build a decision tree. The process is repeated by taking a different number of samples taken from the original data to produce a decision tree each time. The final output is obtained by the majority voting among the decision trees [10].

The main advantage of using random forest is that not all the features are considered to form a decision tree. Features are chosen at random by the algorithm, and each decision tree will be split based on different features leading to a greater ensemble to aggregate.

The same pre-processing and cleaning techniques which were used for logistic regression are adopted for random forests. Since using the default parameters for random forest can cause the model not to perform well, the GridSearchCV technique is used to find the best parameters which can be used in the modelling.

IV. Results and Analysis

This section discusses the results and the analysis part when the child speech dataset is applied. Initially few steps need to be on the data while sending the dataset directly to the model they are discussed below.

*1) **Data acquisition:*** The dataset used for this analysis is obtained from www.kaggle.com, for free. The dataset is prepared by the author using the data derived from the transcripts in the CHILDES project. The author prepared the dataset by selecting relevant information from the Conti 4 ramsden, ENNI, and Gillam datasets. The dataset consists of the narratives of children attempting to narrate a wordless picture.

*2) **Data cleaning:*** After obtaining the data is cleaned. This step involves removing the unnecessary and irrelevant information and replacing the missing values wherever needed in the dataset.

*3) **Exploratory data analysis:*** After ensuring the data is free of missing values and errors, the data is then visualized using charts and graphs to understand the data.

*4) **Feature selection:*** To select only the statistically significant columns from the dataset for modelling a logit model is trained on the dataset to find the probability value (p-value) of all the features. Based on the threshold value the columns which are statistically significant are selected.

The number of boys and girls affected by specific language impairment and Tourette's disorder (the graph is obtained from the dataset chosen for this analysis, about which will be later) is shown in Fig. 1. Similarly analyzing the information of the male and female details present in the corpus are shown in Fig. 2.

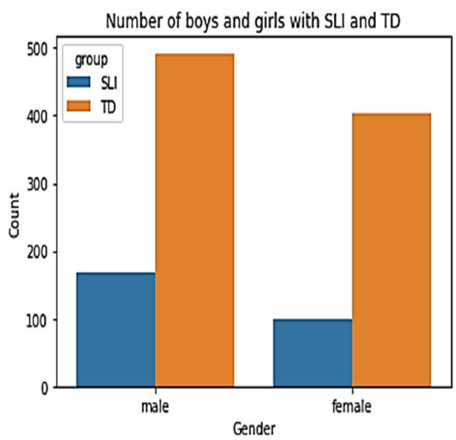

Fig. 1. Analyzing the number of males and females in SLI and TD.

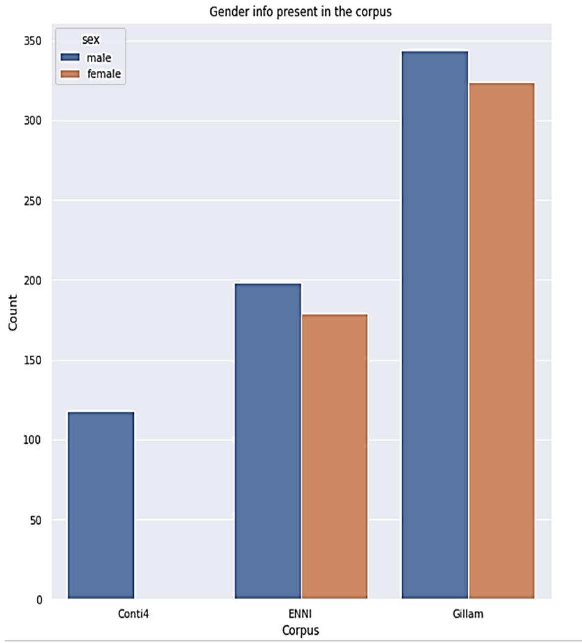

Fig. 2. Gender info presents in the corpus

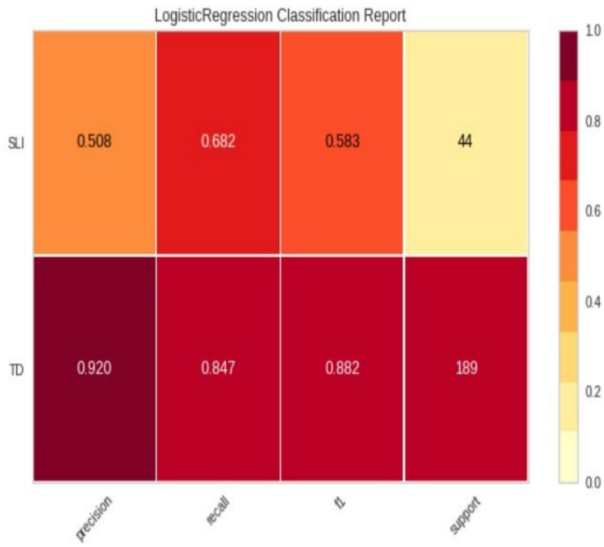

Fig. 3. Gender info presents in the corpus

Now obtaining the classification report for Logistic Regression and analysing the precision, recall, f1, and support scores values for both SLI and TD as shown in Fig. 3. From the heat map, it is observed that the light colour indicates that they are not correlated whereas the dark colour represents that the two variables are they are highly matched. Similarly, the analyzing the random forest classification report in the form of a heat map as shown in Fig. 4.

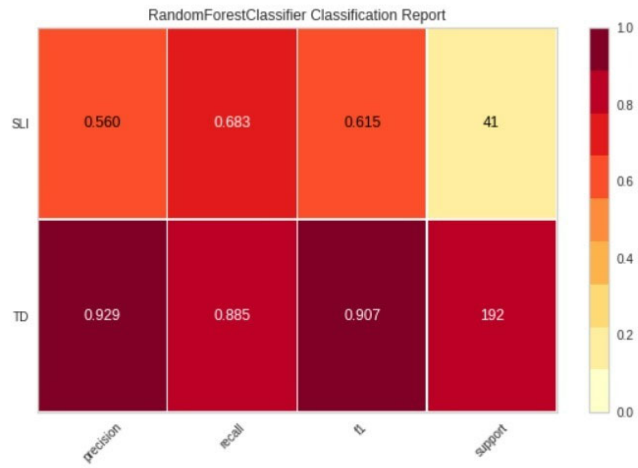

Fig. 4. Analyzing the classification report for SLI and TD using Random Forest.

V. Discussions and Evaluation

This section tries to explore and makes the evaluations on the obtained results. However, the discussions are made in such a way to analyze the results, which are obtained while providing the dataset to the models. In detail, the discussion is described below.

Evaluation methods adopted by speech pathologists:

The speech-language pathologists evaluate the child with Specific Language Impairment through interviews and questionnaires to assess a child's learning ability and through standardized tests. The information such as the number of words spoken by the child to narrate or express a thought to finish a task and the number of words spoken by the examiner as prompts to help the child finish the task is recorded, as children with SLI are more likely to be dependent on prompts given by the examiner for them to finish the task. Then the results are compared with other children of the same age group to identify if the child suffers from specific language impairment. This method of analyzing the transcripts manually can be a tedious task for speech-language pathologists and pediatricians.

The results of Logistic Regression are shown in Fig. 5. This result tries to explain the in-detail information about the data when the dataset is applied to the model along with the score and error values. Here the model is named with Logit and also displays the date, time and time consumed for the model to learn the whole dataset. Here a number of observations show how the model is trained and tried to extract the features from the data for each and every image. These results output also discuss the coefficient correlations and tries to analyze the errors which consist of p values, Z scores for various variables. From the result, it is seen that const have a higher coefficient when compared with others. Similarly, const has a more standard error and p-value while dealing with other parameters.

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	1163
Model:	Logit	**Df Residuals:**	1151
Method:	MLE	**Df Model:**	11
Date:	Sat, 05 Feb 2022	**Pseudo R-squ.:**	0.3261
Time:	05:57:04	**Log-Likelihood:**	-422.25
converged:	True	**LL-Null:**	-626.59
Covariance Type:	nonrobust	**LLR p-value:**	8.850e-81

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	27.3265	5.692	4.801	0.000	16.170	38.483
age	0.3086	0.046	6.694	0.000	0.218	0.399
childtnw	0.0062	0.001	5.535	0.000	0.004	0.008
averagesyl	-19.0915	4.324	-4.416	0.000	-27.565	-10.617
mluwords	-1.5397	0.191	-8.053	0.000	-1.914	-1.165
propositionsin	-0.0845	0.044	-1.927	0.054	-0.171	0.001
irregularpasttense	-0.0223	0.009	-2.365	0.018	-0.041	-0.004
regularpasted	-0.1176	0.023	-5.146	0.000	-0.162	-0.073
regular3rdpersons	-0.0866	0.019	-4.670	0.000	-0.123	-0.050
uncontractibleaux	-0.0416	0.015	-2.814	0.005	-0.071	-0.013
worderrors	0.4135	0.074	5.557	0.000	0.268	0.559
fk	1.5529	0.369	4.204	0.000	0.829	2.277

Fig. 5. Results of Logistic Regression.

The selected features are then sent into the logistic regression model after splitting the data into train and test data using the standard train-test split method available in the sklearn library.

Along with the LR results, the classification report is shown in Fig. 6. Here 0 indicates non-cancer and 1 represents cancer. The data shows the classification report of the various scores. The logistic regression model provides a training accuracy of 79%. Whereas the testing accuracy is seen as 88%.

```
-----------------------------------------
        Logistic regression results
-----------------------------------------
              precision    recall  f1-score   support

           0       0.77      0.81      0.79       707
           1       0.80      0.76      0.78       707

    accuracy                           0.79      1414
   macro avg       0.79      0.79      0.78      1414
weighted avg       0.79      0.79      0.78      1414

              precision    recall  f1-score   support

           0       0.92      0.85      0.88       189
           1       0.51      0.68      0.58        44

    accuracy                           0.82       233
   macro avg       0.71      0.76      0.73       233
weighted avg       0.84      0.82      0.83       233
```

Fig. 6. Classification of Logistic Regression.

The classification report of Random Forest is shown in Fig. 7. Basically, a random forest is a combination of multiple decision trees and this model tries to extract each and every feature based upon the feature the model takes the decision. When the ASD dataset is applied to the training this model provides an accuracy of 96%. The testing data provides an accuracy of 96%. Normally micro averages deal with calculating each individual class. And this tries to add all the true positive, false negatives values together. Then these calculated values will be applying to scores.

```
----------------------------------------
        Random Forests results
----------------------------------------
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       704
           1       0.96      0.96      0.96       704

    accuracy                           0.96      1408
   macro avg       0.96      0.96      0.96      1408
weighted avg       0.96      0.96      0.96      1408

              precision    recall  f1-score   support

           0       0.93      0.89      0.91       192
           1       0.56      0.68      0.62        41

    accuracy                           0.85       233
   macro avg       0.74      0.78      0.76       233
weighted avg       0.86      0.85      0.86       233
```

Fig. 7. Classification report of Random Forest.

VI. CONCLUSION

This paper concludes the work that is done on the detection of speed disorder among children. The collected dataset was pre-processed now implementing logistic regression and random forests on the data and based on the results obtained for each model, the Specific Language Impairment (SLI) in a child can be identified with 88% and 96% accuracy respectively. Since the data available to train the machine learning models is very limited, that is, 267 cases for SLI and 896 cases for TD, and the models were trained on synthetic data to identify the classes, the accuracy could not be improved any further. The accuracy of these machine learning models can be improved if more data with relevant features are available to train and automate the process.

VII. ACKNOWLEDGMENT

The author is grateful to acknowledge the management and staff of Hail University, Hail, Saudi Arabia for their support and academic encouragement to carry out the research at a wide spectrum of research areas in computer science Engineering.

VIII. REFERENCES

[1] B. Dodd. Differential diagnosis and treatment of children with speech disorder. John Wiley & Sons, 2013.

[2] A. R. Lederberg, B. Schick, and P. E. Spencer. "Language and literacy development of deaf and hard-of-hearing children: successes and challenges." Developmental psychology 49, no. 1 (2013): 15.

[3] J.B. Tomblin, L. Nancy. Records, Paula Buckwalter, Xuyang Zhang, Elaine Smith, and Marlea O'Brien. "Prevalence of specific language impairment in kindergarten children." Journal of speech, language, and hearing research 40, no. 6 (1997): 1245-1260.

[4] H. Tager-Flusberg, and C. Kasari. "Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum." Autism research 6, no. 6 (2013): 468-478.

[5] L. Rello, E. Romero, M. Rauschenberger, A. Ali, K. Williams, J. P. Bigham, and N. C. White. "Screening dyslexia for English using HCI measures and machine learning." In Proceedings of the 2018 international conference on digital health, pp. 80-84. 2018.

[6] L.M. Justice, W. Y. Ahn, and J.A.Logan. "Identifying children with clinical language disorder: An application of machine-learning classification." Journal of learning disabilities 52, no. 5 (2019): 351-365.

[7] M. Shahin, U. Zafar, and B. Ahmed. "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results." IEEE Journal of Selected Topics in Signal Processing 14, no. 2 (2019): 400-412.

[8] C.S. Kanimozhiselvi, and S. Santhiya. "Communication Disorder Identification from Recorded Speech using Machine Learning Assisted Mobile Application." In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 789-793. IEEE, 2021.

[9] sklearn, 2022 [online], available at URL: < https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>, accessed on February 4, 2022.

[10] D. Mulfari, G. Meoni, M. Marini, and L. Fanucci. "Machine learning assistive application for users with speech disorders." Applied Soft Computing 103 (2021): 107147.