Journal of Autonomous Intelligence (2023) Volume 6 Issue 1
doi:10.32629/jai.v6i1.553

ORIGINAL RESEARCH ARTICLE

Machine learning approach to analyze the impact of demographic and linguistic features of children on their stuttering

Shaikh Abdul Waheed[1*], Mohammed Abdul Matheen[2], Syed Hussain[3], Amairullah Khan Lodhi[4], G.S. Maboobatcha[2]

[1] *Department of Information Technology, G H Raisoni College of Engineering and Management, Pune 412207, India. E-mail: mail.2.abdulwaheed@gmail.com*

[2] *King Saud University, Riyadh 12467, Saudi Arabia.*

[3] *B.S. Abdur Rahman Crescent Institute of Science & Technology, Chennai 400048, India.*

[4] *Electronics & Communication Engineering, Shadan College of Engineering & Technology, Peerancheru, Hyderabad 500086, India.*

ABSTRACT

This study aims at analyzing the impact of gender and race on the linguistic abilities and stuttering of children. The current article also seeks to check whether children with stuttering disorder and normal children differ in linguistic skills. Parametric methods like t-tests and Analysis of Variance (ANOVA) have been applied to test hypotheses. The p-values that were generated in the parametric tests signify that the gender of the child has an impact on the onset of stuttering. However, the race of children did not affect the onset of stuttering. The regression results of the machine learning part have indicated many findings. The results indicated that a child's race does not impact the onset of stuttering. Hence, the null hypothesis about race was accepted by signifying that children of any race can adopt stuttering. This finding also suggests that children can face linguistic difficulties irrespective of their race. Another finding is that children with stuttering (CWS) repeat more words than children with not stuttering (CWNS). In addition, CWS repeat more syllables than CWNS. It indicates that the null hypothesis can be accepted by stating that children can suffer from linguistic difficulties irrespective of their race. Another key finding is that there can be a significant difference in the linguistic abilities of male and female children. Another inference is that the p-values indicate a significant difference between linguistic skills among CWS and CWNS. In other words, CWS are more prone to repeat syllables than normal children.

Keywords: Machine Learning Approach; Regression; Stuttering; Demographic and Linguistic Features

ARTICLE INFO

Received: 29 February 2023
Accepted: 20 April 2023
Available online: 24 May 2023

1. Introduction

Data analytics made it possible to capture hidden patterns from historical data. The data analytics approach is rigorously applied for predicting various diseases and events[1–5]. Data analytics methods can be parametric and non-parametric. This study aimed to utilize parametric methods for analyzing linguistic and stuttering-like disfluencies in children. This study has made use of historical data on stuttering from Vanderbilt University. This data was created based on five-year-long persistent research. Stuttering is a disorder that is seen in people when they communicate. Individuals who stutter struggle to maintain the flow of speech. Commonly, childhood stuttering starts in children at the stage of 3.5 years[6]. Literature suggests that there are plenty of reasons in the background of the beginning of stuttering. Gender disparity is also seen in stuttering. The male population is seen to be more affected by stuttering than the female population[7]. On the other hand, it

was seen that the linguistic abilities of children with stuttering (CWS) were found to be poor in comparison to children with not stuttering (CWNS). Both males and females can be affected by stuttering. However, it was reported that males are more at risk of stuttering than females[8]. Like men who stutter, women who stutter also have the impact of stuttering on their lifestyle and career[9].

The contributions of this paper are as follows:

- Analyzing the effect of children's demographics features like gender and race on the onset of stuttering.
- Analyzing children's linguistic abilities on the onset of stuttering.
- Also, analyzing the impact of parents' economic status on children's onset of stuttering.
- The outcomes of this paper would help Speech Language Pathologists (SLPs) to understand the directionalities of features like demographics, gender, race, and social-economic status of parents on the onset of stuttering.

2. Related work

A research study has been conducted on 32 children to assess their linguistic features. Children were divided into two groups of 16 each. It was found that CWNS showed non-stuttering-like disfluencies (nonSLD), while CWS exhibited stuttering-like disfluencies (SLD)[10]. To assess speech disfluency in children with attention deficit/hyperactivity disorder (ADHD), a total of 15 children have been involved in this study. It was seen that such children are disfluent in speech production[11]. Another study was conducted on 31 children to evaluate fluency. It was noticed that CWS showed delays in motor speech development[12]. Previous studies report that males are more prone to stuttering than females. Males are prone to stuttering may be due to work stress, career tension, and other stressful situations. For example, the ratio of stuttering between males and females is between 3:1 and 5:1[7]. Regarding the relationship between stuttering and family history, many studies have been carried out. For instance, a study reveals that linguistic and attentional deficits can be associated with a family history of stuttering. However, the authors said these findings need to be confirmed with rigorous research[13]. Many studies were carried out to investigate the emotional relation with stuttering[4].

The linguistic features of 41 Jordian children with stuttering issues were evaluated[14]. This study considered phonological and morphological variables of participants' speech during experiments. Stuttering instances were found to occur more in the initial position of the word than in the medial or final position of a sentence. Another similar study was conducted on Lebanese children who stutter[15]. This investigation evaluated the linguistic disfluencies of the subjects through video and audio recordings of their speech. Anderson and Wagovich[16] conducted a research study on CWS and CWNS to assess the speech and language performance of the participants. Abdul Waheed and Abdul Khader[1] investigated the link between temperaments and stuttering to predict childhood stuttering. Modeling was created using logistic regression by taking the temperaments of normal children and children with stuttering as independent variables. In their study, it was identified that temperaments could act as a predictor to forecast stuttering in children. In another study[17], the association between emotions in adults and stuttering was analyzed by considering four types of emotions: boredom, frustration, excitement, and meditation. This investigation revealed that there is a relativity between frustration and boredom and the severity of adulthood stuttering. To find out interconnections among temperaments with respect to childhood stuttering, structural equation modeling was applied[4]. The findings of this research indicate that temperaments have complex interconnections among them in relation to stuttering. However, regression analysis is needed to understand the type of relationship between demographic and linguistic features of children and stuttering. Still, a concrete study is needed to understand linguistic differences between CWS and CWNS with the dataset containing a diverse population of children. Therefore, this research study sets several objectives such as 1) analyzing the correlation between races of children and their linguistic abilities, 2) analyzing the correlation between races of children and the onset of stuttering, 3) analyzing the correlation between gender and onset of stuttering, 4) analyzing whether races of children have an impact on their linguistic abilities, 5) analyzing whether the races of children have an effect on the onset of stuttering, 6) ana-

lyzing whether the gender factor has an impact on the onset of stuttering, and 7) analyzing whether CWS and CWNS differ in linguistic abilities. Along with these objectives, we hypothesized that gender does not differ in the linguistic abilities of children. Also, it was hypothesized that race has no impact on their linguistic abilities. On another side, it was hypothesized that the gender factor has no impact on the onset of stuttering. Similarly, it was assumed that the race of children does not have any effect on stuttering.

3. Proposed work

3.1 Selection of dataset

This study did not experiment directly with the children to collect stuttering data. This study has utilized existing freely accessible datasets on childhood stuttering. Data preprocessing was carried out to handle missing values. It was found that the race variable has three missing values at row numbers 77, 98, and 134. These missing values cannot be filled up using missing value handling functions because it is a categorical variable. Missing values of a categorical variable cannot be replaced with synthetic values. Hence, we planned to delete such rows. The IBM SPSS software tool was utilized to carry out involved experiments. The publicly available dataset (called Developmental Stuttering Project) from Vanderbilt University was taken into consideration for performing the intended analysis. In this dataset, data points about CWS and CWNS were recorded. The present study has utilized information like gender, race, and linguistics from the selected dataset. This dataset has data points of about 138 children (average age: 48 months). Out of 138 data points, race-related data about three children was missing; three rows (77, 98, 134) were deleted. Here, seven races (1–7) of children have been mentioned.

3.2 Parametric machine learning approach: Linear regression

From the machine learning approach, the linear regression was applied to find the correlation between 1) races of children and their linguistic abilities, 2) races of children and onset of stuttering, and 3) gender and onset of stuttering. This regression analysis finds out the correlation between the outcome variable and independent variables. In the first case, the outcome variable was disfluency per 100 words, and the independent variable was the race of the children. While in the second case, the SSI_total was considered as the outcome variable, and the races of children were considered as independent variables. Here, the outcome variable represents a stuttering score of children against the Stuttering Severity Instrument (SSI) scale. In the third case, the SSI_total was considered as the outcome variable, and the genders of children were considered as independent variables.

3.3 Non-parametric machine learning approach: T-test and ANOVA

In this approach, T-test and ANOVA were utilized to find within groups differences as mentioned in the below sub-sections.

3.3.1 Analyzing whether races of children have an impact on children's linguistic abilities

To confirm whether the races of children affect their linguistic abilities, we applied "ANOVA" (Analysis of Variance). While applying this test, we consider two cases of linguistic abilities of children: per 100 words and 100 syllables. The null hypothesis was set as the races of children have no impact on children's linguistic abilities.

1) Based on disfluency per 100 words

In the first case, a variable like "Disfluency_SLD per 100 words" was considered as the dependent variable against the categorical independent variable "race". Race has seven categorical values, as shown in **Table 1**.

Table 1. Demographic information of children: Races

Label	Race
1	Asian
2	Black or African American
3	American Indian/Alaska Native
4	Hawaiian
5	White
6	More than one race
7	Unknown/not reported

2) Based on disfluency per 100 syllables

For the second case, a feature such as "Disfluency_SLD per 100 syllables" was taken as the dependent variable, while the categorical variable

"race" was considered as an independent variable. In both cases, the null hypothesis was that the races of children do not have any impact on their linguistic skills.

3.3.2 Analyzing whether the races of children have an impact on onset of stuttering

To confirm this fact, ANOVA was applied between groups of races and the dependent variable SSI_total. As shown in **Table 1**, race is a categorical variable with seven values. Here, the null hypothesis was set as the races of children would not impact the onset of stuttering.

3.3.3 Analyzing whether gender factor has an impact on onset of stuttering

To analyze this fact, an "independent samples t-test" was performed by taking gender (M & F) as an independent variable against the dependent variable "SSI_total", representing the SSI-based total of stuttering. This dependent variable is the sum of several factors, such as the frequency and duration of dysfluency and the behavior of children. The null hypothesis was considered as the gender of children would not have an impact on stuttering.

3.3.4 Analyzing whether CWS and CWNS differ in linguistic abilities

To analyze whether CWS and CWNS differ in linguistic abilities, we applied an "independent samples t-test" between linguistic features and the dependent variable "talkergroup_SSI", which represents the stuttering score. Here, CWS and CWNS are two groups of children that are independent of each other. While applying this test, we consider two cases of linguistic abilities of children: per 100 words and 100 syllables. The null hypothesis was set as linguistically no significant difference between CWS and CWNS.

1) Based on disfluency per 100 words

In the first case, a variable like "Disfluency_SLD per 100 words" was considered as the dependent variable against the categorical independent variable "talkergroup_SSI".

2) Based on disfluency per 100 syllables

For the second case, features such as "Disfluency_SLD per 100 syllables" were taken as a dependent variable, while the categorical variable "talkergroup_SSI" was considered as an independent variable. In both cases, the null hypothesis was set as CWS and CWNS do not differ in linguistic skills.

3.3.5 Applied data analytics methods

Data analytics can be performed using parametric and non-parametric methods. The present study adopted a parametric tests-based data analytics approach. In this approach, data is assumed to be in normal distribution form. We adopted parametric tests like "independent samples t-test" and ANOVA. The t-test analyzes variance between two groups, while ANOVA is efficient in capturing differences between more than two groups. The formula of the t-test is,

$$t = \frac{\text{Variance between Groups}}{\text{Variance within Groups}}$$

The null hypothesis is stated as no significant difference between the means (μ) of the two groups. This can be represented as $\mu 1 = \mu 2$.

The alternative hypothesis is stated as at least one mean is different.

1) The formula of ANOVA

AVONA coefficient is calculated using the below formula,

$$F = \frac{Variance\ between\ samples}{Variance\ within\ samples}$$

In this technique, the null hypothesis is stated as no significant difference between the means (μ) of groups of samples. This can be represented as:

$$\mu 1 = \mu 2 = \mu 3 = \ldots = \mu k.$$

The alternative hypothesis is stated as at least one mean is different.

4. Result analysis

Table 2 indicates that gender has a 32% correlation with stuttering score (i.e., SSI_total). The correlation between stuttering scores and races is found to be void.

Figure 1 gives the clear impact of gender on the onset of stuttering. This figure exhibits that gender 1 is more prone to stuttering than gender 0. While performing linear regression, the numbers 0 and 1 were assigned to males and females, respectively. **Figure 2** indicates that a child's race does not impact the onset of stuttering. This figure exhibits the onset of stuttering across the races.

Table 2. Correlation table

		SSI_total	Gender	Race
Pearson correlation	SSI_total	1.000	.320	.000
	Gender	.320	1.000	0.045
	Race	.000	.045	1.000
Sig.(1-tailed)	SSI_total	.	.000	.500
	Gender	.000	.	.301
	Race	.500	.301	.
N	SSI_total	135	135	135
	Gender	135	135	135
	Race	135	135	135

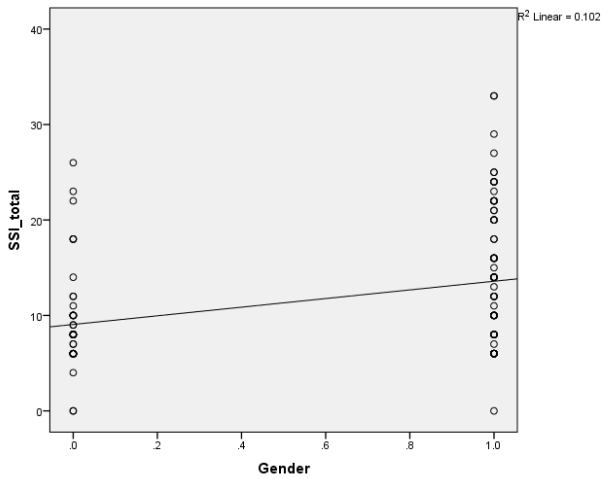

Figure 1. Correlation between gender and onset of stuttering.

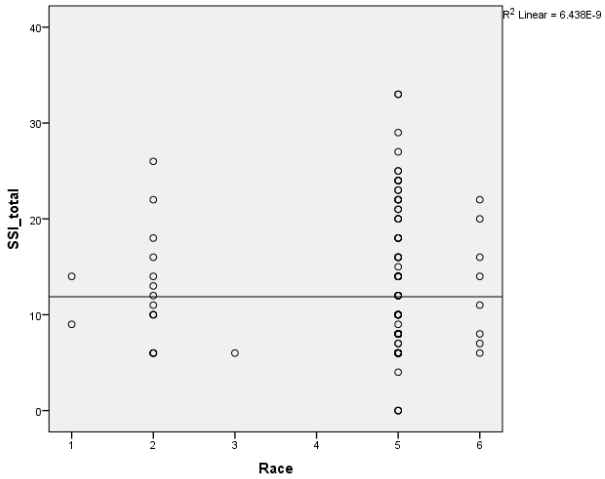

Figure 2. Correlation between race and onset of stuttering.

Figure 3 provides a clear picture of the impact of gender on the rate of disfluency. This figure also shows that gender 1 is prone to disfluency. So, **Figure 1** and **Figure 3** imply that gender 1 is prone to disfluency. In turn, gender 1 may lead to the onset of stuttering. Unlike the results displayed in **Figure 2**, **Figure 4** shows that across the races, children can stutter because there is no significant impact on the onset of childhood stuttering.

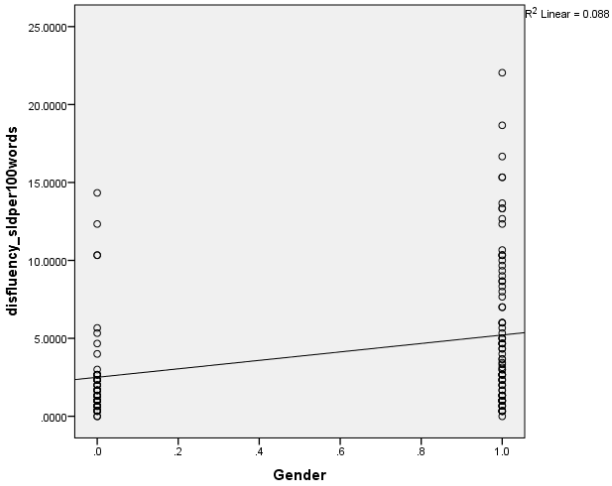

Figure 3. Correlation between race and onset of stuttering.

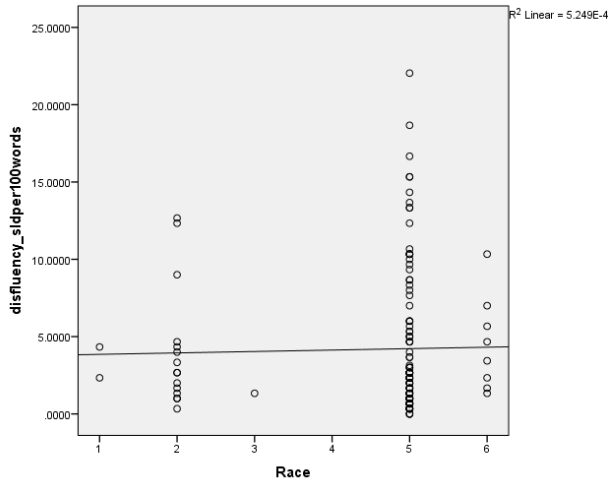

Figure 4. Correlation between race and onset of stuttering.

4.1 Difference of linguistic skills in terms of word in CWS and CWNS

The statistical results like mean, std. deviation, std. errors and p-value that were generated during the t-test between children and linguistic features in terms of per 100 words have been shown in **Table 3**. The p-values (0.000) show that CWS and CWNS differ significantly in linguistic features in terms of 100 words. Hence, the null hypothesis was rejected as CWS repeats more words than CWNS.

4.2 Difference of linguistic skills in terms of syllables in CWS and CWNS

Another key result about linguistic features in terms of syllables has been presented in **Table 4**. The p-value signifies that there can be a significant difference between linguistic abilities among CWS and CWNS. In other words, CWS are more prone to repeat syllables than normal children. Hence, the null hypothesis was also rejected by CWS repeating

Table 3. Group statistics between linguistic skills of CWS and CWNS in terms of per 100 words

	Levene's test for equality of variances		T-test for equality of means						
Disfluency_SLD per 100 words	**F**	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std. error difference**	**95% confidence interval of the difference**	
								Lower	**Upper**
Equal variances assumed	106.65	.000	−13.24	133	.000	−6.77	.51	−7.79	−5.76
Equal variances not assumed			−11.26	57.65	.000	−6.77	.60	−7.98	−5.76

Table 4. Group statistics between linguistic skills of CWS and CWNS in terms of per 100 syllables

	Levene's test for equality of variances		T-test for equality of means						
Disfluency_ SLD per 100 syllables	**F**	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std. error difference**	**95% confidence interval of the difference**	
								Lower	**Upper**
Equal variances assumed	106.70	.000	−13.25	133	.000	−5.89	.44	−6.77	−5.01
Equal variances not assumed			−11.27	57.65	.000	−5.89	.52	−6.94	−4.84

more syllables than CWNS.

4.3 Relation between races and linguistic ability of children

Table 5 represents ANOVA results between children's races and their linguistic abilities. Here, the p-value (0.970) indicates no significant relation between the races of children and their linguistic ability. This finding accepts the null hypothesis by stating that children can suffer from linguistic difficulties regardless of race.

Table 5. Measuring impact of races over linguistic abilities using ANOVA

	Sum of squares	**df**	**Mean square**	**F**	**Sig.**
Between groups	10.71	4	2.68	.132	.970
Within groups	2,633.72	130	20.26		
Total	2,644.43	134			

4.4 Impact of races of children on onset of stuttering

Table 6 depicts one-way ANOVA results between children's races and the onset of stuttering. The generated p-value is greater than 0.05. This p-value reflects no significant association between the races of children and the onset of stuttering. Hence, here null hypothesis was accepted by signifying that children of any race can adopt stuttering.

Table 6. Measuring impact of races on stuttering using ANOVA

	Sum of squares	**df**	**Mean square**	**F**	**Sig.**
Between groups	50.04	4	12.51	.259	.904
Within groups	6,290.82	130	48.39		
Total	6,340.86	134			

4.5 Impact of gender (M & F) on onset of stuttering

The results that were produced during that t-test between gender (M and F) and the onset of stutter-

Table 7. Group statistics of impact of gender (M & F) on stuttering

	Levene's test for equality of variances		T-test for equality of means						
SSI_total	**F**	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std. error difference**	**95% confidence interval of the difference**	
								Lower	**Upper**
Equal variances assumed	12.94	.000	3.90	133	.000	4.52	1.16	2.23	6.82
Equal variances not assumed			4.23	130.001	.000	4.52	1.07	2.41	6.64

ing are listed in **Table 7**. The resulting p-value is less than 0.05. The null hypothesis was rejected by stating that there is gender disparity in the adaptation of stuttering. Previous literature also indicates that male candidates are more prone to stuttering than females[8].

4.6 Linguistic differences between gender (M & F)

4.6.1 Linguistic differences per 100 words

The "independent samples t-test" was performed among data points of gender (M & F) and their linguistic features in 100 words. The results of this test are listed in **Table 8**. The obtained p-value (0.000) indicates that the null hypothesis was rejected by saying there can be a significant difference in the linguistic abilities of males and females.

Table 8. Group statistics of impact of gender (M & F) on linguistic skills per 100 words

	Levene's test for equality of variances		T-test for equality of means						
Disfluency_SLD per 100 words	**F**	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std. error difference**	**95% confidence interval of the difference**	
								Lower	**Upper**
Equal variances assumed	15.96	.000	3.59	133	.000	2.71	.76	1.22	4.21
Equal variances not assumed			3.99	132.78	.000	2.71	.68	1.37	4.05

4.6.2 Linguistic differences per 100 syllables

Table 9 represents the results of the "independent samples t-test" that was performed between data samples of gender (M & F) and their linguistic score in terms of per 100 syllables. The obtained p-value (0.000) rejects the null hypothesis by stating that there can be a significant difference in the linguistic abilities of males and females. Thus, we applied the data analytics approach through parametric methods using t-test and ANOVA. Hypotheses were tested against p-values.

Table 9. Group statistics of impact of gender (M & F) on linguistic skills per 100 syllables

	Levene's test for equality of variances		T-test for equality of means						
Disfluency_SLD per 100 syllables	**F**	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std. error difference**	**95% confidence interval of the difference**	
								Lower	**Upper**
Equal variances assumed	15.96	.000	3.59	133	.000	2.36	.66	1.06	3.66
Equal variances not assumed			3.99	132.78	.000	2.36	.59	1.19	3.53

5. Conclusions

This study analyzed whether CWS and CWNS differ in linguistic skills through historical data. The present article also analyzed the impact of gender and race on children over childhood stuttering. The findings of this reveal that gender has an impact on the onset of stuttering, but race did not show any significant impact on stuttering. This finding shows that children across races can be prone to stuttering. In the future, non-parametric methods like machine learning algorithms can be applied to predict the odds of stuttering[18].

Conflict of interest

All authors declare no conflict of interest.

References

1. Abdul Waheed S, Abdul Khader PS. A machine learning approach for managing the potential risk of odds of developmental stuttering. International Journal of System Assurance Engineering and Management 2021. doi: 10.1007/s13198-021-01151-6.
2. Waheed SA, Revathi S, Matheen MA, *et al*. Processing of human motions using cost effective EEG sensor and machine learning approach. In: 2021 1st International Conference on Artificial Intelligence

and Data Analytics (CAIDA); 2021 Apr 6–7; Riyadh, Saudi Arabia. New York: IEEE; 2021. p. 138–143. doi: 10.1109/CAIDA51941.2021.9425088.
3. Waheed SA, Khader PSA. A novel approach for smart and cost effective IoT based elderly fall detection system using Pi camera. In: 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC); 2017 Dec 14–16; Coimbatore. New York: IEEE; 2018. p. 1–4. doi: 10.1109/ICCIC.2017.8524486.
4. Waheed SA, Khader PSA. Healthcare solutions for children who stutter through the structural equation modeling and predictive modeling by utilizing historical data of stuttering. SAGE Open 2021; 1–23. doi: 10.1177/21582440211058195.
5. Waheed SA, Khader PSA. IoT based approach for detection of dominating emotions in persons who stutter. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC); 2020 Oct 7–9; Palladam, India. New York: IEEE; 2020. p. 14–18. doi: 10.1109/I-SMAC49090.2020.9243392.
6. Yairi E, Seery CH. Stuttering: Foundations and clinical applications. Upper Saddle River, NJ: Pearson; 2015.
7. Yairi E, Ambrose N, Cox N. Genetics of stuttering: A critical review. Journal of Speech Language and Hearing Research 1996; 39(4): 771–784. doi: 10.1044/jshr.3904.771.
8. Bloodstein O, Ratner NB. A handbook on stuttering. 6th ed. Clifton Park, NY: Thomson/Delmar Learning; 2008.
9. Nang C, Hersh D, Milton K, *et al*. The impact of stuttering on development of self-identity, relationships, and quality of life in women who stutter. American Journal of Speech-Language Pathology 2018; 27(3S): 1244–1258. doi: 10.1044/2018_AJSLP-ODC11-17-0201.
10. Zackheim CT, Conture EG. Childhood stuttering and speech disfluencies in relation to children's mean length of utterance: A preliminary study. Journal of Fluency Disorders 2003; 28(2): 115–142. doi: 10.1016/S0094-730X(03)00007-X.
11. Lee H, Sim H, Lee E, *et al*. Disfluency characteristics of children with attention-deficit/hyperactivity disorder symptoms. Journal of Communication Disorders 2017; 65: 54–64. doi: 10.1016/j.jcomdis.2016.12.001.
12. Smith A, Goffman L, Sasisekaran J, *et al*. Language and motor abilities of preschool children who stutter: Evidence from behavioral and kinematic indices of nonword repetition performance. Journal of Fluency Disorders 2012; 37(4): 344–358. doi: 10.1016/j.jfludis.2012.06.001.
13. Choi D, Conture EG, Tumanova V, *et al*. Young children's family history of stuttering and their articulation, language and attentional abilities: An exploratory study. Journal of Communication Disorders 2018; 71: 22–36. doi: 10.1016/j.jcomdis.2017.11.002.
14. Alqhazo M, Al-Dennawi S. The linguistic aspects of the speech of Jordanian children who stutter. International Journal of Pediatric Otorhinolaryngology 2018; 109: 174–179. doi: 10.1016/j.ijporl.2018.04.003.
15. Merouwe SS, Bertram R, Richa S, *et al*. Identification of stuttering in bilingual Lebanese children across two presentation modes. Journal of Fluency Disorders 2023; 76: 105970. doi: 10.1016/j.jfludis.2023.105970.
16. Anderson JD, Wagovich SA. Relationships among linguistic processing speed, phonological working memory, and attention in children who stutter. Journal of Fluency Disorders 2010; 35(3): 216–234. doi: 10.1016/j.jfludis.2010.04.003.
17. Abdul Waheed S, Abdul Khader PS. IoT based intelligent healthcare monitoring system for measuring emotional intelligence in adults having speech disorder. In: Network modeling analysis in health informatics and bioinformatics. Berlin, Germany: Springer Nature; 2021.
18. Bhat SA, Dar MA, Elalfy H, *et al*. A novel framework for modelling wheelchairs under the realm of Internet-of-Things. International Journal of Advanced Computer Science and Applications (IJACSA) 2021; 12(2): 745–751. doi: 10.14569/IJACSA.2021.0120293.