Journal of Learning and Teaching in Digital Age, 2023, **8**(1), 25-31
https://dergipark.org.tr/en/pub/joltida
ISSN: 2458-8350 (online)

Research Paper

A Novel Approach to Identify Difficult Words in English to Promote Vocabulary Acquisition of Children Dually Diagnosed with Autism Spectrum Disorder and Hearing Loss

Ebru Efeoğlu[a], Ayşe Tuna[b]*

[a](0000-0001-5444-6647), Kütahya Dumlupınar University, ebru.efeoglu@dpu.edu.tr
[b](ORCID ID: 0000-0003-2666-1694), Trakya University, aysetuna@trakya.edu.tr
*Corresponding author

ARTICLE INFO

Received: 23 February 2022
Revised: 8 April 2022
Accepted: 6 July 2022

Keywords:
Autism Spectrum Disorder
Hearing Loss
Challenges
Recommendations
Future Research Directions

doi: 10.53850/joltida.1078232

ABSTRACT

Usage of complex words causes significant problems not only in reading but in writing as well and eventually leads to poor academic achievement of students, poorer particularly for hearing impaired children. The dual diagnosis of Autism Spectrum Disorder (ASD) and hearing impairment pose additional challenges mainly due to the difficulties that come with making accurate decisions. Hence, parents must be provided with the information about the signs and symptoms of ASD and deafness or partial hearing loss, as well as appropriate intervention strategies. Although different learning activities can be used to enlarge such children's vocabulary, if the presented words are difficult to learn, it will be very hard to realize this. Identifying difficult words and replacing them with simple ones both make the readability of a text easier and help such children enhance their vocabulary knowledge in a shorter period of time. Therefore, in this study we propose a classification approach that identifies difficult words among a given set of words in English. The lexical and semantic features of the words in the dataset were extracted based on the language rules specific to hearing impaired children. In the classification approach, five popular classification algorithms were used and the algorithms' performance in identifying difficult words was evaluated using various performance metrics. As the results show, the K-Nearest Neighbors algorithm is the most suitable algorithm for identifying difficult words in English for the target group.

INTRODUCTION

Children with Autism Spectrum Disorder (ASD) have impairments in social and communication skills. ASD can be diagnosed using specific tests but the diagnosis can only be confirmed by a clinician or professional (Phillips et al., 2021). The Autism Diagnostic Observation Schedule is one of the tests used in the diagnosis of ASD. For children with hearing impairment, the test can be adapted by using sign language. On the other hand, another ASD diagnostic test, the Autism Diagnostic Interview-Revised (ADI-R), could be adapted with some visual tools to help determine whether a child has ASD, hearing loss, or both (McTee et al., 2019).

Children with ASD often have additional diagnoses (Szymanski & Brice, 2008). It is known that hearing impairment existence among children with ASD is higher compared to typically developing children (Trudeau et al., 2021). Children with hearing impairment isolate themselves socially because of the communication related difficulties they experience (VanDam & Yoshinaga-Itano, 2019). They also have difficulties in communicating with gestures and expressing their feelings through facial expressions (VanDam & Yoshinaga-Itano, 2019). Hearing difficulties and ASD often overlap, which make autism traits worse and diagnoses more complicated (Myck-Wayne, Robinson, & Henson, 2011). Children with ASD have tremendous demands from their families and this may make hearing loss go unnoticed (Myck-Wayne, Robinson, & Henson, 2011). Because not reacting to noises, speech delay and problems, consistently repeated words, behavioral problems, having trouble communicating, lack of attention, lack of eye contact and clumsiness are common signs and symptoms of hearing loss and they may be understood as being symptoms of ASD (VanDam & Yoshinaga-Itano, 2019). Therefore, at an early age, listening and communication skills of children must be evaluated so that signs of hearing loss and ASD can be identified (Trudeau et al., 2021). Then, if needed, appropriate intervention can be initiated as soon as possible in order to relieve parents of a massive challenge (Wiley, Gustafson, & Rozniak, 2014).

Individuals with hearing impairment generally experience a great deal of difficulty in acquiring spoken languages contrary to their natural acquisition of signed languages (Berent, 2001). Due to not being able to have full access to the sounds and intonations of a spoken language, individuals with hearing impairment have a significant difficulty in the spoken language acquisition processes and even if they succeed in acquiring the spoken language, the process occurs at a considerably slower rate compared to normally hearing individuals (Quigley & King, 1980). Therefore, most individuals with hearing impairment do not accomplish full acquisition of a spoken language and have constant difficulties in written expression and reading comprehension (Berent, 2001).

Sentences are made by the combination of simple words, complex words or both. If a text consists of many complex words, it becomes a difficult one which leads to problems not only in writing but in reading as well. Compared to their normal hearing peers, such texts are more difficult for children with hearing impairment to learn and understand. Using different activities and flash cards, speech language therapists, their parents and teachers help children with hearing impairment learn new words and enlarge vocabulary. Nevertheless, if words in a text are difficult to learn, to develop understanding of the word in children's mental lexicon will still be considerably difficult. In English, the grammatical order of words is known as Subject, Verb, and Object. On the other hand, children with hearing impairment as English learners do not follow this order (Berent, 2001). As a consequence, when they are asked to read and write, it becomes a serious challenge for them. Therefore, significant time and effort from parents, teachers and speech language therapists are required to address all these difficulties.

In this paper, we propose a novel approach to identify difficult words among a given set of words in English so that difficult words in a text can be replaced with easier ones and this way that activities to promote vocabulary acquisition can be coordinated better. In the dataset (Ansar, Qamar, Bibi, & Shaheen, 2019) we used for classification study, the lexical and semantic features of the words are based on the language rules specific to children with hearing impairment. The rest of this paper is as follows. The next section provides information about our approach proposed for pre-intervention phase carried out for children with hearing impairment. The third section presents the results of our classification study. Finally, the paper is concluded in the fourth section.

Dataset and Novel Approach for Pre-Intervention Phase

Dataset

The dataset used in the study was taken from the study carried out by Ansar, Qamar, Bibi, and Shaheen (2019). The dataset consists of 1000 words: a total of 600 words as the training set and a total of 400 words as the test set. It was collected from the English textbooks and online sources being taught at elementary and secondary level schools established specifically for children with hearing impairment. Before building the dataset, unstructured data was first reviewed by a group of experts and then preprocessed. This way, the generation of linguistic rules that helps to label a particular word as difficult or not difficult accurately was possible. The rows in the dataset were reviewed by the experts and a word was accepted as difficult if 2 or more experts decided that it was difficult (Ansar, Qamar, Bibi, & Shaheen (2019). The features that make a word difficult are listed in Table 1.

Table 1. Features that make a word difficult (adapted from (Ansar, Qamar, Bibi, & Shaheen, 2019))

Features of Words	Characteristics
Character count	Increasing the number of characters in a word makes the word more difficult.
Syllable count	Increasing the number of syllables in a word makes the word more difficult.
Part of speech tags	Compared to adjectives and verbs, nouns are easier to learn. Adverbs are the most difficult since they describe an abstract idea. Children with hearing impairment can learn concrete words easier than abstract words.
Presence of ch, st, th, f, or sh	If ch, st, th, f, or sh is present in a word, the word is considered difficult. Because ch, st, th, f, and sh can produce high-frequency sounds during pronunciation.
Presence of c or k	Depending on context, pronunciation of c and k are different or the same. This makes it difficult for children with hearing impairment to recognize during writing or reading, what to read.
Presence of g or j	Depending on context, pronunciation of g and j are different or the same. This makes it difficult for children with hearing impairment to recognize during writing or reading, what to read.
Frequency of occurrence	If a word appears in a text less frequently, it is deemed as difficult otherwise easy.

Classification Algorithms

The following list briefly describes classification algorithms used in this study. The reason for preferring these algorithms is that their working principles are different from each other, except for Support Vector Machine and Sequential Minimal Optimization algorithms.

- K-Nearest Neighbors (KNN): It is a supervised learning algorithm and mostly used for classification and sometimes for regression. As its name suggests that it considers *k* nearest neighbors (data points) to predict the class or continuous value for the new data point (Ma, Du, & Cao, 2020). It finds the nearest neighbors in a dataset by using distance metrics in order to realize classification and its success generally depends on the distance metrics preferred in the study and the number of neighbors represented by *k* (Jiang, Pang, Wu, & Kuang, 2012; Xia et al., 2015).
- Naive Bayes (NB): It is a probabilistic classification method based on the well-known Bayes' theorem. It calculates the probability that a new data belongs to any of the existing classes using the classified sample data. The class with the highest probability among the values found is accepted as the class to which the sample belongs (Bermejo, Gámez, & Puerta, 2011).
- Linear Discriminant Analysis (LDA): It was first developed by R. A. Fisher for binary classifications in 1936 (Cohen, Cohen, West, & Aiken, 2002). It was later generalized by Rao (1948). In discriminant analysis, discriminant functions allow distinguishing classes from each other and this way it is decided which class the new sample should be included in (McLachlan, 1992). The LDA finds a linear combination of features that characterizes or separates two or more classes of

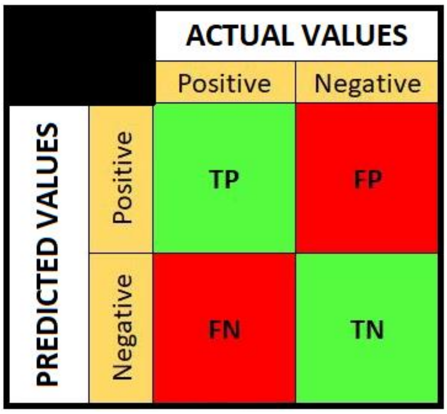
objects or events (Han et al., 2020). Although the resulting combination may be used as a linear classifier, the LDA is typically used for dimensionality reduction before later classification (Li et al., 2021).

- Support Vector Machine (SVM): In this algorithm, each data item is plotted as a point in the n-dimensional space (where *n* is the number of features) with the value of each feature being the value of a particular coordinate. Next, the classification is performed by finding the hyperplane that distinguishes quite well from the two classes. In this regard, support vectors are just the coordinates of the observation. Therefore, the SVM is just a boundary that best separates the two classes (Cortes & Vapnik, 1995).
- Sequential Minimal Optimization (SMO): It is an algorithm for solving the quadratic programming (QP) problem that arises during the training of SVMs without the need for any extra matrix storage and numerical QP optimization steps (Platt, 1998). At each step, the SMO selects two Lagrange multipliers to jointly optimize, finds optimal values for these factors, and updates the SVM to reflect the new optimal values (Platt, 1998).

Metrics

Depending on the success of predictions, related values are presented in a confusion matrix as shown in Figure 1. These values are grouped in four groups. In this study, these four groups have the following meanings.

- True Positive (TP): If the algorithm used predicts an easy word as an easy word.
- True Negative (TN): If the algorithm used predicts a difficult word as a difficult word.
- False Positive (FP): If the algorithm used predicts a difficult word as an easy word.
- False Negative (FN): If the algorithm used predicts an easy word as a difficult word.



		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

Figure 1. Confusion matrix

Basic metrics commonly used in classification studies are listed in Table 2. As explained in Table 2, while accuracy considers all correct and incorrect predictions, other metrics focus on different aspects of classification. While recall is the answer to the question of how many TPs are correctly identified, precision shows how many of the values predicted as positive are actually positive (Tharwat, 2021). Except for the ones listed in Table 2, Cohen's Kappa coefficient (abbreviated as Kappa), Root Mean Square Error (RMSE) and Area Under the Curve (AUC) are used in this study. Receiver Operating Characteristic (ROC) curve is a probability curve, while AUC represents the decomposable measure or degree of parameters. A ROC curve makes it possible to visualize to what extent the classes for which the model results are to be predicted differ (Bradley, 1997; Martínez-Camblor, Pérez-Fernández, & Díaz-Coto, 2021). A high AUC score indicates good separation between classes (easy word, difficult word). The Kappa statistic represents the extent to which the data collected are correct representations of the variables measured (McHugh, 2012). Finally, RMSE is a measure of accuracy and allows comparing prediction errors of different models for a particular dataset (Hyndman & Koehler, 2006).

Table 2. Basic metrics commonly used in classification studies

Metric	Interpretation	Formula
Accuracy	It indicates the overall performance of a classifier.	(TP+TN)/(TP+TN+FP+FN)
Precision	It indicates how accurate the positive predictions are.	TP/(TP+FP)
Recall	It indicates the coverage of actual positive samples.	TP/(TP+FN)
Specificity	It indicates the coverage of actual negative samples.	TN/(TN+FP)
F-measure	It is a hybrid metric particularly suitable for imbalanced datasets.	2TP/(2TP+FP+FN)

Classification Results and Analysis

The flowchart of the proposed approach is shown in Figure 2. The dataset used in this study, in which the training and test sets are found, consists of 1000 words. In this dataset, 520 of the words are easy words and 480 of them are difficult words. In this study, different from the previous study that used the same dataset (Ansar, Qamar, Bibi, & Shaheen (2019), the training and test sets were combined and cross validation was applied. Thus, it was possible to test all the words.

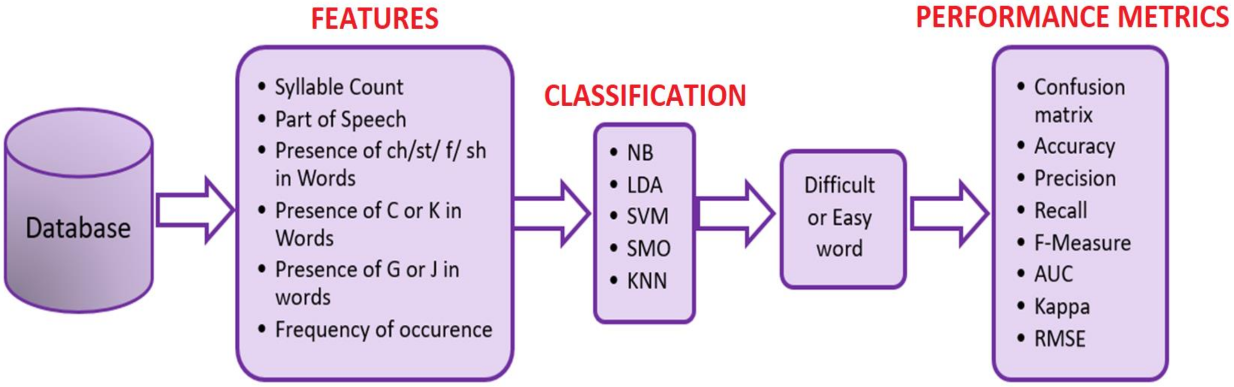
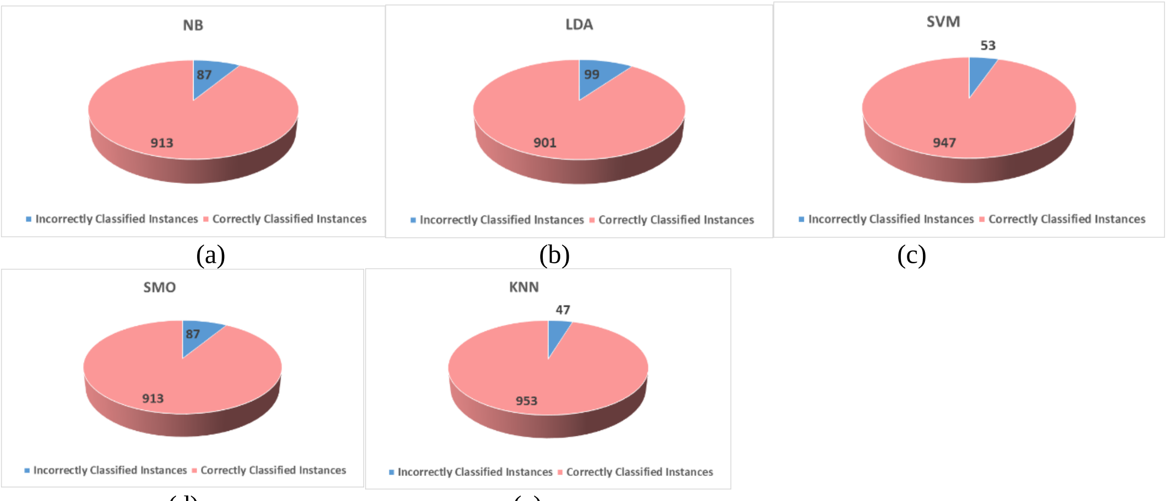


Figure 2. Flowchart of the proposed approach

Confusion matrices consisting of information about the number of words that the classification algorithms predicted correctly and incorrectly are given in Figure 3. In addition, the number of correctly and incorrectly predicted words belonging to each algorithm is given in Figure 4. As it can be seen in Figure 4, the KNN and SVM algorithms were the most successful ones in predicting the words correctly, and the LDA showed the worst performance in classifying the words. In the confusion matrices, the parts indicated in green are TP and TN values. These are the number of words that the algorithms predicted correctly. As it can be seen in the confusion matrices, the NB algorithm predicted 463 of 520 easy words correctly and 57 of them incorrectly. While identifying 480 difficult words, the NB algorithm classified 30 difficult words as easy words. The KNN correctly predicted the highest number of difficult words and the highest number of easy words. Therefore, it had the best performance. Compared to the others, the LDA correctly predicted the least number of difficult words and the least number of easy words. While calculating the performance metrics of the algorithms, precision, recall, F-measure and AUC values were calculated separately in order to examine their performance in predicting both easy words and difficult words. The performance metrics are given in Figure 5.

Naive Bayes				
	Predicted class			
Actual Class		Easy	Difficult	Total
	Easy	TP=463	FN=57	520
	Difficult	FP=30	TN=450	480
	Total	493	507	1000
LDA				
	Predicted class			
Actual Class		Easy	Difficult	Total
	Easy	TP=460	FN=60	520
	Difficult	FP=39	TN=441	480
	Total	499	501	1000
SVM				
	Predicted class			
Actual Class		Easy	Difficult	Total
	Easy	TP=497	FN=23	520
	Difficult	FP=30	TN=450	480
	Total	527	473	1000
SMO				
	Predicted class			
Actual Class		Easy	Difficult	Total
	Easy	TP=469	FN=51	520
	Difficult	FP=36	TN=444	480
	Total	505	495	1000
KNN				
	Predicted class			
Actual Class		Easy	Difficult	Total
	Easy	TP=501	F=19	520
	Difficult	FP=28	TN=452	480
	Total	529	471	1000

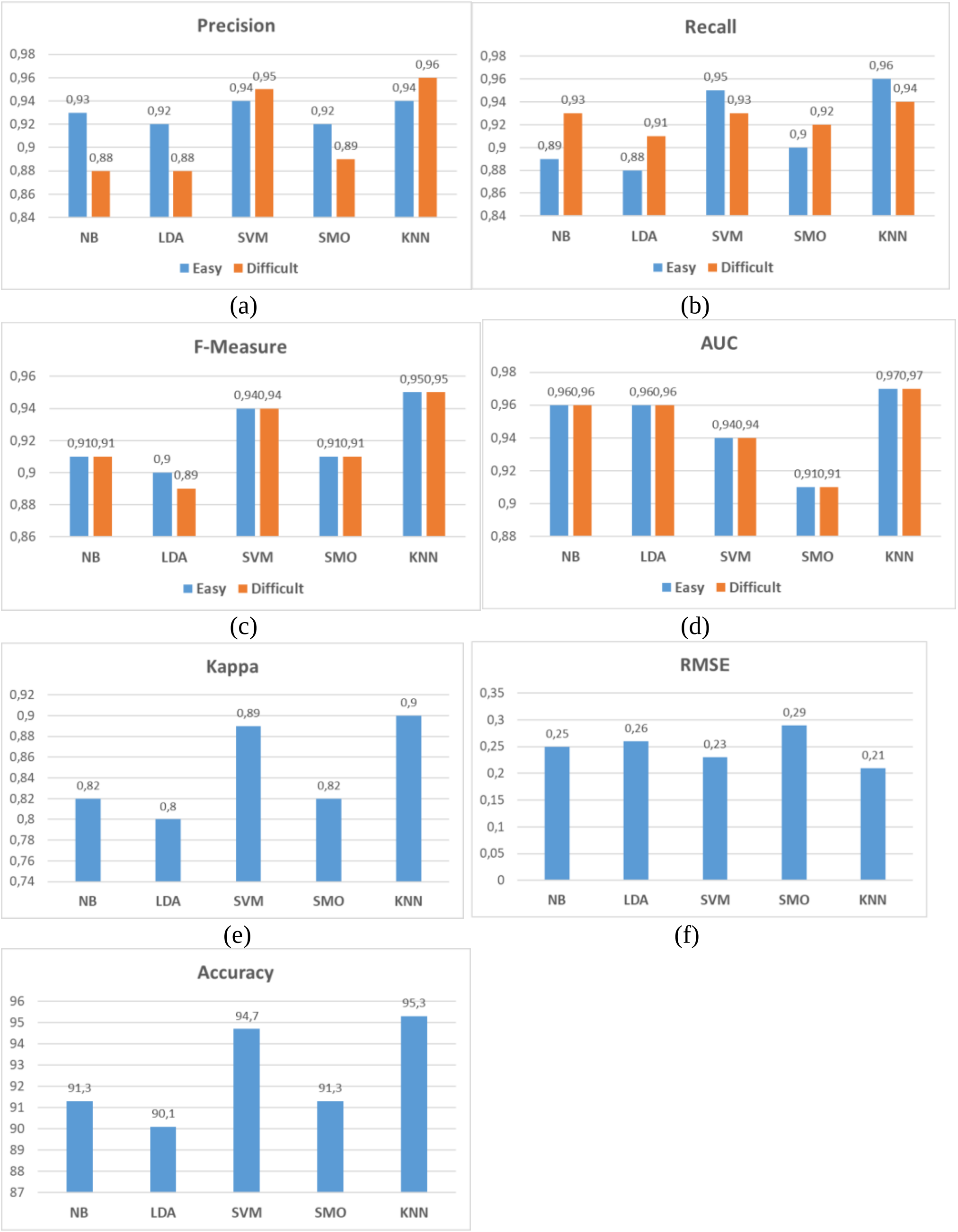
Figure 3. Confusion matrices that show the prediction power of the classification algorithms used in this study



(a) (b) (c)

(d) (e)

Figure 4. Number of correctly or incorrectly classified words: (a) NB, b) LDA, c) SVM, d) SMO, e) KNN



(a) (b)

(c) (d)

(e) (f)

(g)

Figure 5. Performance metrics of the classifiers used in this study: a) Precision, b) Recall, c) F-Measure, d) AUC, e) Kappa, f) RMSE, g) Accuracy (%)

As shown in Figure 5, the highest Precision and Recall values were achieved by the KNN algorithm, followed by the SVM algorithm. On the other hand, the lowest Precision and Recall values were achieved by the LDA algorithm. Since the KNN algorithm achieved the highest AUC value, it can be seen as the best algorithm when the differentiation performances of all the algorithms were considered. When the Kappa values of all the algorithms were considered, it can be seen that the KNN algorithm achieved the highest value and the LDA algorithm achieved the lowest value. RMSE is checked to understand how much error has been generated when a prediction has been realized. The KNN algorithm achieved the lowest RMSE and the SMO algorithm achieved the highest RMSE. The KNN achieved the highest F-Measure values, too. Finally, the highest accuracy was achieved by the KNN algorithm and the lowest accuracy was achieved by the LDA algorithm. The overall results confirm the success of the KNN algorithm over the others. Compared to an accuracy of 92% achieved by the C4.5 algorithm (Quinlan, 1993) in (Ansar, Qamar, Bibi, & Shaheen, 2019), in this study both the KNN and SVM algorithms achieved slightly higher accuracy rates. However, this may be a consequence of the fact that the dataset was used in this study in a different way.

CONCLUSION

ASD interferes with an individual's communication and socialization skills. Although ASD is not curable, various therapies can help children with ASD, especially medium to high-functioning ones, to live independently. As the literature shows, children with hearing impairment have higher chances of having ASD; therefore, hearing impairment needs to be identified at an early age. This way a timely diagnosis of ASD can be realized and then timely intervention can be ensured because hearing impairment can interfere negatively with children's development in terms of social, communication and language skills. It is known that complex words are both difficult to be read and written. The use of complex words makes it harder for hearing impaired children to read and understand. Therefore, such difficult words must be identified and be replaced with simple ones so that such children can read and understand more easily and enhance their vocabulary knowledge in a shorter period of time.

In this study a classification approach was proposed to identify difficult words among a given set of words in English. The classification approach is based on the language rules specific to hearing impaired children. A dataset consisting of 1000 words, 520 easy words and 480 difficult words, was used and different from the previous studies, the training and test sets were combined and cross validation was applied. In this way all the words were tested. To make a fair comparison five different classification algorithms were chosen and their performances were analyzed based on the well-known performance metrics. As the results proved, the KNN algorithm had the best performance in identifying difficult words in English. Future work of this study consists of developing a mobile application that analyzes a document to decide whether it is suitable for being used for children dually diagnosed with ASD and hearing loss and making word appropriate replacements in the document to make it more suitable if needed. It is believed that such an application will probably be helpful for parents to decide which materials are more suitable for their children or need to be revised.

Ethics and Consent: This article does not contain any studies involving human participants performed by any of the authors.

Conflicts of Interest and Authorship: No conflicts of interest and authorship to declare.

Funding: No funding has been received for this study.

REFERENCES

Ansar, M., Qamar, U., Bibi, R., & Shaheen, A. (2019). Identification of Difficult English Words for Assisting Hearing Impaired Children in Learning Language. *2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA)*, 60-65. doi: 10.1109/SERA.2019.8886796.

Berent, G. P. (2001). English for deaf students: Assessing and addressing learners' grammar development. In D. Janáková (Ed.), International Seminar on Teaching English to Deaf and Hard-of-Hearing Students at Secondary and Tertiary Levels of Education: Proceedings (pp. 124-134). Prague, Czech Republic: Charles University, The Karolinum Press.

Bermejo, P., Gámez, J. A., & Puerta, J. M. (2011). Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. *Expert Systems with Applications, 38*(3), 2072-2080. doi: 10.1016/j.eswa.2010.07.146

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159. doi: 10.1016/S0031-3203(96)00142-2

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.). Routledge. doi: 10.4324/9780203774441

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273-297. doi: 10.1007/BF00994018

Han, N., Wu, J., Fang, X., Wen, J., Zhan, S., Xie, S., & Li, X. (2020). Transferable Linear Discriminant Analysis. *IEEE Transactions on Neural Networks and Learning Systems, 31*(12), 5630-5638. doi: 10.1109/TNNLS.2020.2966746

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688. doi: 10.1016/j.ijforecast.2006.03.001

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications, 39*(1), 1503-1509. doi: 10.1016/j.eswa.2011.08.040

Li, Y., Liu, B., Yu, Y., Li, H., Sun, J., & Cui, J. (2021). 3E-LDA: Three Enhancements to Linear Discriminant Analysis. *ACM Transactions on Knowledge Discovery from Data, 15*(4), 57. doi: 10.1145/3442347

Ma, C., Du, X., & Cao, L. (2020). Improved KNN Algorithm for Fine-Grained Classification of Encrypted Network Flow. *Electronics, 9*(2), 324. Retrieved from http://dx.doi.org/10.3390/electronics9020324

Martínez-Camblor, P., Pérez-Fernández, S. & Díaz-Coto, S. (2021). The area under the generalized receiver-operating characteristic curve. *The International Journal of Biostatistics*, 20200091. doi: 10.1515/ijb-2020-0091

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282.

McLachlan, G. J. (1992). Discriminant analysis and statistical pattern recognition. Hoboken, NJ, USA: John Wiley & Sons.

McTee, H. M., Mood, D., Fredrickson, T., Thrasher, A., & Bonino, A. Y. (2019). Using Visual Supports to Facilitate Audiological Testing for Children with Autism Spectrum Disorder. *American journal of audiology, 28*(4), 823-833. doi: 10.1044/2019_AJA-19-0047

Myck-Wayne, J., Robinson, S., & Henson, E. (2011). Serving and Supporting Young Children with a Dual Diagnosis of Hearing Loss and Autism: The Stories of Four Families. *American Annals of the Deaf, 156*(4), 379-90. doi: 10.1353/aad.2011.0032

Phillips, H., Wright, B., Allgar, V., McConachie, H., Sweetman, J., Hargate, R., Hodkinson, R., Bland, M., George, H., Hughes, A., Hayward, E., De Las Heras, V., & Le Couteur, A. (2022). Adapting and validating the Autism Diagnostic Observation Schedule Version 2 for use with deaf children and young people. *Journal of autism and developmental disorders, 52*(2), 553-568. doi: 10.1007/s10803-021-04931-y

Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.

Quigley, S. P., & King, C. M. (1980). Syntactic performance of hearing impaired and normal hearing individuals. *Applied Psycholinguistics, 1*(4), 329-356. doi: 10.1017/S0142716400000990

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Szymanski, C., & Brice, P. J. (2008). When Autism and Deafness Coexist in Children: What We Know Now. Odyssey: New Directions in Deaf Education, *9*(1), 10-15.

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168-192. doi: 10.1016/j.aci.2018.08.003

Trudeau, S., Anne, S., Otteson, T., Hopkins, B., Georgopoulos, R., & Wentland, C. (2021). Diagnosis and patterns of hearing loss in children with severe developmental delay. *Am J Otolaryngol, 42*(3), 102923. doi:10.1016/j.amjoto.2021.102923

VanDam, M, & Yoshinaga-Itano, C. (2019). Use of the LENA autism screen with children who are deaf or hard of hearing. *Medicina (Kaunas), 55*(8), 495. doi:10.3390/medicina55080495

Wiley, S., Gustafson, S., & Rozniak, J. (2014). Needs of parents of children who are deaf/hard of hearing with autism spectrum disorder. *Journal of deaf studies and deaf education, 19*(1), 40-49. doi: 10.1093/deafed/ent044

Xia, S., Xiong, Z., Luo, Y., Dong, L. & Zhang, G. (2015). Location difference of multiple distances based k-nearest neighbors algorithm. *Knowledge-Based Systems, 90,* 99-110. doi: 10.1016/j.knosys.2015.09.028