Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 892 – 907

doi:10.55730/1300-0632.4024

A machine learning approach for dyslexia detection using Turkish audio records

Tuğberk TAŞ[1], Abdullah BÜLBÜL[1*], Abas HAŞİMOĞLU[2], Yavuz MERAL[3], Yasin ÇALIŞKAN[3], Gunay BUDAGOVA[4], Mucahid KUTLU[5]

[1]Department of Computer Engineering, Ankara Yıldırım Beyazıt University, Ankara, Turkiye
[2]Department of Child and Adolescent Psychiatry, Bakirköy Prof. Dr. Mazhar Osman Research and Training Hospital for Psychiatric and Neurological Diseases, İstanbul, Turkiye
[3]Department of Child and Adolescent Psychiatry, Basaksehir Cam and Sakura City Hospital, İstanbul, Turkiye
[4]Department of Child and Adolescent Psychiatry, Biruni University School of Medicine, İstanbul, Turkiye
[5]Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkiye

Received: 12.03.2023 • **Accepted/Published Online:** 11.08.2023 • **Final Version:** 29.09.2023

Abstract: Dyslexia is a learning disorder, characterized by impairment in the ability to read, spell, and decode letters. It is vital to detect dyslexia in earlier stages to reduce its effects. However, diagnosing dyslexia is a time-consuming and costly process. In this paper, we propose a machine-learning model that predicts whether a Turkish-speaking child has dyslexia using his/her audio records. Therefore, our model can be easily used by smart phones and work as a warning system such that children who are likely to be dyslexic according to our model can seek an examination by experts. In order to train and evaluate, we first create a unique dataset that includes audio recordings of 12 dyslexic children and 13 nondyslexic children in an 8-month period. We explore various machine learning algorithms such as KNN and SVM and use the following features: Mel-frequency cepstral coefficients, reading rate, reading accuracy, the ratio of missing words, and confidence scores of the speech-to-text process. In our experiments, we show that children with dyslexia can be detected with 95.63% accuracy even though we use single-sentence long audio records. In addition, we show that the prediction performance of our model is similar to that of the humans'. In this paper, we provide a preliminary study showing that detecting children with dyslexia based on their audio records is possible. Once the mobile application version of our model is developed, parents can easily check whether their children are likely to be dyslexic or not, and seek professional help accordingly.

Key words: Dyslexia, machine learning, detection, classification, audio records

1. Introduction

Dyslexia is a specific learning disorder that causes difficulties in word recognition, and poor decoding and spelling abilities [1]. Approximately 80% of individuals with a diagnosis of specific learning disorder have dyslexia [2]. While the rate of incidence of dyslexia in adults is reported as 4%, its rate in children varies between 5% and 15% [3], Thus, it might be the most frequent neurodevelopmental disorder affecting children [4].

As there is no association between dyslexia and IQ [5], even students with high IQ scores can experience the aforementioned difficulties which are likely to decrease their academic success. Consequently, dyslexic students often suffer from frustration and have low self-esteem [6].

*Correspondence: mabulbul@aybu.edu.tr

Detecting dyslexia in earlier stages pave the way for dyslexic people to access special education, which can reduce the effects of dyslexia, and therefore improve their quality of life. However, the diagnosis of dyslexia is a costly and time-consuming process. Child psychiatrists conduct several tests about their reading, spelling, and writing abilities, and consider the presence of any neurological disease or any private condition such as environmental or economic that could complicate learning processes. As child psychiatrists cannot examine all children, we need systems to easily detect children who are likely to be dyslexic and have to be examined by child psychiatrists.

In order to address this problem, several researchers focused on predicting whether a person is dyslexic or not using machine learning techniques. They explored using various data types such as magnetic resonance imaging (MRI) scans [7, 8], electroencephalography (EEG) scans [9–11], movements of eyes while reading a text [12], reading errors, and images of hand-writings [13]. However, many of these require special devices to gather the required data, limiting their large-scale usage. In addition, the methods which can be used on simple devices focus on only a specific language such as Spanish [14] and English [15].

In this work, we propose a novel method to detect dyslexia in children using their audio records in Turkish. In particular, we use accuracy in reading, reading rate, and Mel-frequency cepstral coefficients (MFCC) of the recordings as features and explore various machine learning algorithms including SVM, KNN, and random forest. While deep learning methods, such as deep convolutional neural networks (CNNs), have gained popularity in detection and classification problems, we chose to employ traditional machine learning algorithms for two main reasons. Firstly, the interpretability of the models is crucial in understanding the underlying factors contributing to dyslexia detection. Traditional machine learning algorithms provide interpretable decision boundaries and feature importance measures, which can aid in identifying the relevant markers of dyslexia. Additionally, our dataset size was relatively small, and traditional machine learning algorithms have shown effective performance with limited data availability. In order to train and test our model, we create a dataset covering 25 individuals in an 8-month time frame during the COVID-19 pandemic. Subsequently, we divide each recording into n-grams where each item is a sentence in order to augment the data size.

In our comprehensive experiments, we show that our model can detect dyslexia in children based on a single sentence recording (i.e. unigrams) with a 95.63% accuracy, and its performance increase as we use more sentences per recording. In addition, we show that our model's prediction performance is similar to human prediction based on audio files. Moreover, we observe that MFCC is the most effective feature.

The contributions of this work are as follows.

- We created the very first dataset to detect dyslexia in Turkish-speaking children.
- We propose a model which can detect dyslexia by using audio recordings. Therefore, our model can be implemented on smart phones, enabling its large-scale usage. Consecutively, children who are likely to be dyslexic can be detected at very early ages, and seek treatment by child psychiatrists.
- Following restrictions of our institutional review board (IRB) approval, we share our code and data with researchers for future studies.

The rest of the paper is organized as follows. We present related work in the literature in Section 2. Section 3 describes the data set construction process. In Section 4, we provide the problem definition. We explain our proposed model in Section 5. We present experimental results in Section 6 and discuss the results of our qualitative analysis in Section 7. We conclude in Section 8 and discuss future research directions.

2. Related work

In this section, we discuss related work on applying machine learning models for various medical tasks focusing on studies related to dyslexia and other similar learning disorders.

Regarding studies focusing on autism, Omar et al. [16] propose a model to predict autism in individuals. They train a random forest model with various features including the age and inheritance of the individual and his/her answers to specific questions used in the diagnosis of autism. Similarly, Stevens et al. [17] propose a model to predict autism in individuals. They train an unsupervised model with eight different features where each point corresponding to proficiency in relevant treatment domains. Wall et al. [18] utilize machine learning methods to make the observation-based diagnosis of autism less time-consuming by decreasing the number of questions in the Autism Diagnostic Observation Schedule-Generic (ADOS). They report that they could decrease the question number by 72.4% while keeping the accuracy at 99%. Rahman et al. [19] present a comprehensive review of machine learning methods for feature selection and classification of autism spectrum disorder (ASD). They highlight the importance of early identification and intervention in improving language, communication, and well-being for children with autism. The study explores various machine learning algorithms, including artificial neural networks, support vector machines, a priori algorithms, and decision trees, applied to autism-related datasets to develop predictive models. In our work, we focus on detecting individuals with dyslexia using their audio records.

As voice conveys a huge amount of information about humans, there exist a number of studies utilizing audio records to predict various medical and psychological statuses of humans such as depression [20], Parkinson's disease [21], dysarthria-related diseases [22], and COVID-19 [23]. Voice information is also utilized for detecting learning disorders or assisting people with such disorders. Richard and Mathieu [24] explore the application of machine learning to automate the assessment of learning disabilities, specifically dyslexia and dysgraphia, through the analysis of handwritten text pictures and audio recordings. They demonstrate promising results and state that when there is sufficient data, machine learning can be used to detect learning disorders. Žabkar et al. [25] propose an automated readability index that can be utilized when assessing reading disorders such as dyslexia. Atkar and Priyadarshini [26] examined voice-based reading assistance techniques for dyslexic children aged 5–7 in the Hindi language. They stated that MFCC features can be used to detect accurate word reading and suggested the implementation of a system that can provide assistance to the child when a misread word is detected.

In addition to classification tasks, the improvements in AI also enabled researchers to develop models which can potentially be used for treatment. For instance, Fitzpatrick et al. [27] and Inkster et al. [28] developed text-based automated conversational agents, called Woebot and Wysa, respectively, to help people with symptoms of depression.

In order to detect dyslexia, prior studies utilize a wide range of data types, such as EEG scans [10], MRI scans [29], eye-movements [12, 30], reading errors [31], hand-writing [13], and games [14]. Lakretz et al. [31] focus on reading errors to detect dyslexia. In particular, they first ask individuals to read 196 Hebrew words. Subsequently, they manually analyze the recordings to identify the reading error type. Subsequently, they build a latent Dirichlet allocation (LDA)-based model and two Naive Bayes models using reading errors. They report that while the LDA-based model captures reading error patterns, Naive Bayes model achieves higher accuracy. In our work, we focus on Turkish and directly use audio records instead of manually analyzing them.

Rello and Ballesteros [12] develop an SVM model with various features such as age, font of the text, and others extracted from the eye movements of individuals while reading a text. They report that reading time, the mean time of eye fixations (i.e. resting of eyes when reading a particular part of the given text), and the age of the participant are effective features for detecting dyslexia. Benfatto et al. [30] also utilized eye-tracking to predict dyslexia. In particular, they train an SVM model with low-level features based on fixations and saccades to capture the duration, amplitude, direction, stability, and symmetry of eye movements during reading a text.

Cui et al. [7] train a linear SVM model with features extracted from structural magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI). The features they use include white matter volume, mean, axial, radial diffusivity, and fractional anisotropy. Zahia et al. [29] propose a 3D CNN model trained with functional magnetic resonance imaging (fMRI) data of participants. The fMRI scans were obtained while the participants were going through three different reading tasks. In their experiments with a dataset size of 55 children, they report an accuracy of 72.73%. In another MRI-based study, Usman et al. [32] propose a method to improve the interpretability of dyslexia's neural-biomarkers obtained from MRI datasets. Their approach employs Gaussian smoothing and a modified histogram normalization (MHN) method to enhance the biological interpretation of dyslexia-related neural-biomarkers. Usman et al. [33] further contribute to the field by reviewing the implementation details and challenges of advanced machine learning methods for dyslexia biomarker detection. The study aims to provide insights into the practical aspects of applying machine learning techniques to detect dyslexia-related biomarkers. Di Pietro et al. [34] investigate altered brain connectivity in children with dyslexia compared to those with typical reading skills. Their functional MRI study reveals disorder-specific differences in effective connectivity, particularly in the feedback pathway between the inferior parietal lobule and the visual word form area during print processing.

Rello et al. [14] design a game-based machine learning model to detect dyslexia in Spanish-speaking individuals. In particular, they first ask participants to play a game and then use their game performance measures as features. In their follow-up work [15], they designed another game for English-speaking individuals focusing on cognitive language skills associated with dyslexia and errors generally made by dyslexic individuals. Rauschenberger et al. [35] propose a universal screening tool for dyslexia using a web-based language-independent game combined with machine learning models. The study emphasizes the importance of early screening to enable early intervention for children with dyslexia. By analyzing data gathered from the game, the authors train machine learning models that achieve promising accuracy and F1-scores for detecting dyslexia in German and Spanish.

Frid and Manevitz [9] extract features from event-related potentials (ERP) signals such as the flatness of a frequency spectrum, the total amount of positive amplitudes in selected time frames, and the maximal amplitude value in the signal. They train the KNN algorithm with the ERP-based signals and report around 80% accuracy in their experiments conducted with 32 Hebrew-speaking children. Perera et al. [10] gather electroencephalography (EEG) scans of individuals while they perform a given writing task and train a cubic SVM model. They report 71.88% accuracy in their experiments with 32 adults where 17 of them are dyslexic. Rezvani et al. [11] also used EEG scans collected from 29 dyslexic and 15 nondyslexic participants in grade 3. They report 95% accuracy with an SVM model.

Spoon et al. [13] propose a CNN model to detect dyslexia from the handwriting of students. As their preliminary results, they report 55.7±1.4% accuracy. In their follow-up work [36], they extend their dataset and report 77.6% accuracy.

Table 1. Statistical information about texts read by participants.

Text type	Read by	Number of sentences	Number of words
1	First- and second-grade students	12	123
2	Third-grade students	14	150
3	Fourth- and fifth-grade students	34	351

As another prediction problem about dyslexic individuals, Hamid et al. [37] focus on predicting dyslexic students' engagement towards the learning content using their frontal face images. They collect facial images from 30 participants and explore various algorithms including SVM, Naive Bayes, and KNN.

Based on our literature review, several challenges and limitations can be listed regarding the studies applying machine learning techniques for detecting learning disorders. Firstly, constructing a balanced and sufficiently large dataset is extremely challenging due to the sensitive nature of the task. In addition, limited interpretability of models' predictions makes them harder to apply in real life. Furthermore, as each language has its own linguistic features, language-specific solutions are likely to be required.

Our work differs from the prior work in several aspects. Firstly, we use audio records of individuals, enabling us to implement a model which can run on a smart phone. Secondly, we focus on the Turkish language. To our knowledge, there is no prior work focusing on Turkish-speaking individuals.

3. Data collection

One of the main obstacles to developing AI models in medicine is the lack of annotated datasets. This is mainly because it is extremely challenging to collect data due to privacy issues. Furthermore, the data collection usually requires the involvement of experts. In this work, we constructed the very first dataset for dyslexia detection in Turkish. The data collection process took around 8 months in 2020, i.e. during the COVID-19 pandemic. Now we explain the details of our dataset.

We obtained our IRB approval from the ethics committee of Ankara Yıldırım Beyazıt University before starting the data collection process. All participants were aged between 8 and 11, and the mother tongue of all participants was Turkish. Written consent was obtained from each of the participants via their parents or legal guardian. We also asked for informed consent from the participants to ensure their well-being in the process.

The recordings of participants who are not dyslexic were collected from students in different cities at various locations. On the other hand, we collected audio records of dyslexic participants in three different hospitals. Various smart phones have been used for voice recording.

There are three different Turkish texts for different grades used by experts in examinations for the diagnosis of dyslexia. Participants were asked to read the text appropriate to their grades during the clinical examination. Table 1 provides further information about these texts. In addition, Table 2 shows a part of the original Turkish text read by the third-grade students and its English translation.

After a long period of data collecting process, we could collect recordings of 19 boys and 6 girls. Twelve of the participants were diagnosed with dyslexia while the others were not. The total length of recordings was 74.5 min. As shown in Table 3, the distribution of data according to text types is balanced. In particular, seven participants read the first type of text, seven participants read the second type of text, and the remaining 11 participants read the third type of text. Despite months of data collection process, we were not able to find any girl diagnosed with dyslexia. We share our dataset with the researchers following the restrictions of our IRB

Table 2. A passage belonging to the text read by the third-grade students.

Original text in Turkish	**English translation**
Dün gece çok kar yağdı. Sabah olunca Can, ağabeyi, kız kardeşi, annesi ve babası hep birlikte bahçedeki karlardan kocaman bir kardan adam yaptılar. Daha sonra onun çevresinde dönerek, şarkılar söylediler. Bunun üzerine kardan adam, kendisini önemli hissetti. Az sonra komşu teyze, kahvaltının hazır olduğunu söyleyerek hepsini kahvaltıya çağırdı.	It snowed a lot last night. In the morning, Can, his brother, sister, mother, and father all made a huge snowman out of the snow in the garden. Then they were singing songs by turning around him. Therewith, the snowman felt important. A little later, neighbour aunt invited them all to breakfast by telling them breakfast was ready.

Table 3. Gender, label, and text type distribution of our dataset.

Text type	**Participant info**			
	Boy		***Girl***	
	Dyslexic	***Not dyslexic***	***Dyslexic***	***Not dyslexic***
1	***2***	***2***	***0***	***3***
2	***2***	***4***	***0***	***1***
3	***8***	***1***	***0***	***2***

approval. We note that the dataset is totally anonymous such that any information about the identity of the participants is hidden.

4. Problem definition

Diagnosing dyslexia is a long and expensive process that includes a series of tests and observations, as mentioned before. Therefore, in this study, we work on whether we can correctly predict whether a person is dyslexic or not by just using their audio recordings. Thus, our work can be considered a binary classification problem where audio recordings of an individual and the corresponding text are given as input.

We propose a machine-learning model for this problem, as detailed in the following section. In order to train and test machine learning models, we need large amounts of data. However, due to the severe challenges we faced during the data collection phase, we could collect data only from 25 participants. Therefore, we first divided each audio recording of participants into multiple recordings where each one belongs to a single sentence of the corresponding text, yielding 556 sentences in total. Subsequently, we consider each of these sentence-length recordings as a separate data item and use them to train and test our model. Therefore, in our work, we investigate whether we can predict whether a person is dyslexic or not by using the audio recording of a single sentence.

5. Proposed method

As our dataset is not large enough to effectively train data-hunger deep learning models, we propose a feature-based machine learning model to detect dyslexic people. In particular, we utilize five different groups of features and explore various machine learning algorithms including SVM, KNN, and random forests (RF). Now we explain our features in detail.

Reading accuracy. People with dyslexia are more likely to make mistakes in reading than people without dyslexia. Therefore, in this feature, we first convert the given audio recording to text using the Speech-to-Text (STT) API of Google. Next, we remove all punctuation and convert all letters to lowercase in both the converted text and the corresponding ground-truth text. Subsequently, we calculate Levenshtein Distance [38] between these two texts to quantify how accurately the individual could read the given text. To normalize the results, the calculated distance is divided by the expected text length. Furthermore, the min-max normalization method is applied to scale the Levenshtein distance between zero and one.

Reading rate. We expect that dyslexic people will read the texts slower than other people. Therefore, in this feature, we count the number of syllables read per second, based on the Speech-to-Text output. We also apply the min-max normalization to scale the reading rate range between zero and one.

Ratio of read words vs. ground truth. Dyslexic participants have a higher chance to skip a word or even a line in the text without reading. In some cases, they can also read a word multiple times to read it correctly. Therefore, the number of words read by participants might be an important indicator for predicting dyslexia. In this feature, we calculate the ratio of words read by the individual to the number of words in the corresponding text. We use the Speech-to-Text output for this calculation. We also apply min-max normalization.

STT confidence scores. We rely on the output of STT API for the features mentioned above. However, STT does not always work correctly, affecting the features we calculate. Therefore, in order to reduce the negative impact of mistakes during STT process, we use the confidence scores provided by Google's STT API for each conversion[1]. The confidence scores are between zero and one. Values closer to one imply that expressions are clear in the audio recording, and it is expected that the API can convert speech to text correctly. We again apply min-max normalization.

Mel-frequency cepstral coefficients (MFCC). MFCC are coefficients to represent the short-term power spectrum of a sound. MFCC helps model to deal with noisy audio signals. In order to extract MFCC features from an audio signal, we first break the sound signal (i.e. input files) into overlapping frames. Next, we apply fast Fourier transform (FFT) to each window. Humans are more sensitive to changes in audio signal energy at lower levels. Thus, Mel-filter is applied to frequencies and its log values are calculated to mimic human ear perception of sound. Finally, we apply discrete Cosine transform (DCT) and apply mean normalization to obtain MFCC features which have a size of 20.

6. Experiments

6.1. Implementation

The voices of the participants are recorded with different devices, causing different file formats and sample rates. Thus, we converted all sound files to *wav* format and set the sample rate to 16,000 Hz, following the recommendations of Google STT API[2]. We used Librosa, a Python library to load *wav* audio files and extract MFCC features, and scikit-learn[3] library for machine-learning algorithms.

[1] https://cloud.google.com/speech-to-text/docs/basics
[2] https://cloud.google.com/speech-to-text/docs/best-practices
[3] https://scikit-learn.org/stable/user_guide.html

6.2. Experimental setup

Our dataset consists of 566 sentences from 25 individuals. In order to achieve a reliable evaluation, we create four different datasets using our data:

- **Original dataset ($D_{Original}$):** We use all the recordings we collected without any further preprocessing or filtering.
- **Dataset with reduced noise ($D_{ReducedNoise}$):** The audio recordings of the dyslexic participants were taken in a hospital environment, whereas the audio recordings of the nondyslexic participants were mostly taken in a school environment. Therefore, there might be background noise that our models learn to identify dyslexic participants. This might cause an unrealistic evaluation of our models. Thus, we clean the background noise such that models are forced to make their predictions based on the reading of participants instead of the background noise. Noise reduction is done using noisereduce library[4] in Python, which implements a commonly used method [39], where fast Fourier transforms and masking are used with spectral gating to denoise the audio files.
- **Dataset with just boys (D_{Boys}):** As seen in Table 3 our dataset does not include any records belonging to dyslexic girls. Thus, voice differences between boys and girls provide an important clue about the label of data items in our dataset. In order to prevent this situation, we filtered out recordings belonging to girls and used only those of boys for this dataset.
- **Dataset with just boys and reduced noise ($D_{Boys\_ReducedNoise}$):** Similar to $D_{ReducedNoise}$ dataset, we reduce background noise in D_{Boys}, yielding a dataset in which data bias coming from both genders of participants and recording environment has been eliminated.

In order to achieve reliable evaluation, we take the following steps in each dataset mentioned above. Firstly, we use 5-fold cross-validation and report average performance across five trials. Secondly, we randomize the data to prevent any bias due to the order of the data. Thirdly, we ensure that no data item belonging to the same person is in both the training and test set. Lastly, we remove the cases with very low voices because Google STT API was not able to produce any results for these cases.

6.3. Experimental result

In our first experiment, we build models using KNN, SVM, and random forest with all proposed features and assess their performance on four datasets described in Section 6.2. Table 4 shows accuracy and F_1 scores of each model. Our observations are as follows. Firstly, random forest model outperforms KNN in all cases and SVM model in $D_{Original}$ and D_{Boys} datasets. Secondly, SVM model outperforms others in terms of accuracy in datasets in which we reduce noise. Thirdly, cleaning the background noise or removing girl participants' data from the dataset decreases the classification success of KNN and random forest algorithms in terms of both F_1 score and accuracy. This suggests that these algorithms actually utilize gender and background noise information when estimating dyslexia. However, we observe the opposite situation for the SVM model such that its performance in $D_{Original}$ and D_{Boys} is lower than that in other datasets.

In our next experiment, we conduct a feature ablation experiment to measure the effect of each feature on the performance. In this experiment, we train a model using each feature separately with the most successful

[4]https://pypi.org/project/noisereduce/

Table 4. Accuracy and F_1 scores for KNN, SVM, and random forest models with all proposed features on the four datasets. The best result for each case in written in bold.

Model	$D_{Original}$		$D_{ReducedNoise}$		D_{Boys}		$D_{Boys\_ReducedNoise}$	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
KNN	92.91	93.99	82.09	81.38	85.37	75.91	73.93	71.67
SVM	88.48	89.68	**89.57**	**91.44**	85.24	72.45	**85.18**	74.18
RF	**95.63**	**96.03**	89.36	89.27	**93.47**	**79.45**	76.87	**77.07**

Table 5. F_1 score when we use a single feature group to train models.

Selected feature	$D_{Original}$	$D_{ReducedNoise}$	D_{Boys}	$D_{Boys\_ReducedNoise}$
MFCC	96.38	91.58	79.33	75.99
Reading accuracy	60.81	49.02	74.37	74.37
Reading rate	48.00	47.05	69.76	69.76
Ratio of read words vs. ground truth	59.90	49.02	74.14	74.14
STT confidence scores	51.73	47.91	72.80	72.80

algorithm (in terms of F_1 score) for each dataset. In particular, we use SVM for $D_{ReducedNoise}$ and RF for the other three datasets. Table 5 shows the results of feature ablation experiment. In all datasets, we observe that the reading rate is the least effective feature while MFCC is the most effective feature. In addition, reading accuracy is the second most effective feature.

In our experiments so far, we used voice recordings for each sentence. Now we assess the impact of recording length on our model's prediction. In particular, we first create *n-grams* of each recording of participants where each item is a sentence. Using a similar experimental setup, we train a model with all features and calculate its F_1 score. Again we use SVM for $D_{ReducedNoise}$ and RF for the other three datasets due to their high performance in these datasets (see Table 4). We conduct this experiment for all datasets and increase n from 1 to 4. Note that the experiments mentioned above is a specific version of this experiment where n is set to 1. The results are shown in Figure 1. Although the performance increase in $D_{ReduceNoise}$ dataset is very slight, the performance of our approach increases in all datasets as the number of sentences in each record increases. Our results suggest that our model is able to capture more characteristics of reading patterns of dyslexic participants as we use additional sentences.

6.4. Human evaluation

To the best of our knowledge, there is no prior study in the literature predicting whether a person is dyslexic or not from audio recordings for Turkish. Therefore, we are not able to compare our proposed method with any prior work. To address this issue and have a reference value to assess the performance of our model's performance, we conducted an experiment measuring how human beings perform in this prediction task. Specifically, we randomly selected 47 sentences from our dataset. We asked three groups of people to predict whether a person is dyslexic or not by listening to the person's recorded voice. These three groups include 1) nonexpert people without any prior training, 2) nonexpert people who are trained by listening to several records of known cases of dyslexia (including both positive and negative cases), and 3) two child psychiatrists who have

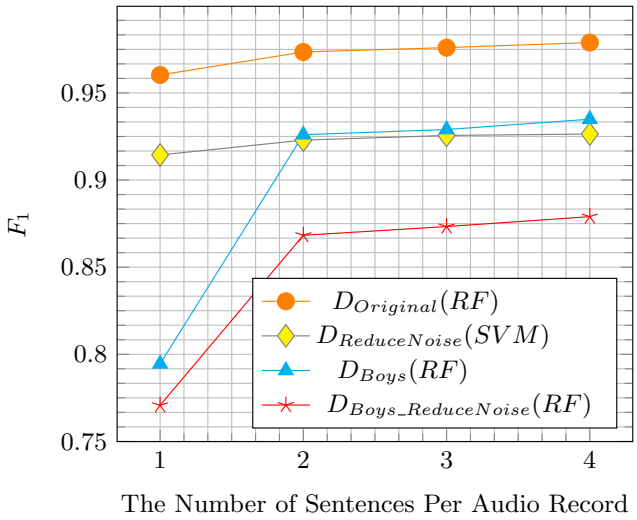

Figure 1. F_1 score of our proposed model when the number of sentences per audio record increased from one to four.

prior experience in diagnosing dyslexia. The total number of participants in the first two groups is 15. Table 6 shows accuracy of all three groups.

As expected, experts outperform nonexperts, and nonexperts with prior training have better prediction accuracy than those without training. Our experiment shows that even experts fail frequently due to just using a single sentence for the prediction. While we cannot compare these results directly with our models' accuracy scores, the performance of humans suggests that our models' have high prediction accuracy and can effectively use the voice signals in diagnosis.

Table 6. Accuracy of human judgments.

	Accuracy
Nonexpert with no prior training	*60.4*
Nonexpert with prior training	*68.9*
Expert	*75.53*

7. Qualitative analysis

In order to have more insights into how dyslexia can be diagnosed based on voice records and our model's performance, we investigate which letters, syllables, and words in Turkish are more difficult to read for dyslexic individuals compared to nondyslexic individuals. Now we report details of our analysis and discuss our findings.

We first divide our datasets into two subsets: recordings of nondyslexic individuals and dyslexic individuals. Subsequently, we convert each recording to text using Google STT. Next, we apply case-folding on the resultant text and compare it to the ground truth text to detect misread words, syllables, and characters. We use SequenceMatcher[5] to detect misread words and python-syllable[6] to detect syllables in words.

[5]https://docs.python.org/3/library/difflib.html
[6]https://github.com/ftkurt/python-syllable

We first focus on misread letters and calculate how many times nondyslexic and dyslexic individuals misread each letter. Subsequently, we calculate the ratio of misreading frequency between nondyslexic and dyslexic individuals for each letter by the following formula:

$$R^i = (N^i_{dyslexic} - N^i_{non-dyslexic})/N^i \tag{1}$$

where N^i represents the number of letter i in the ground truth text, $N^i_{Non-dyslexic}$ is the number of times nondyslexic individuals misread the letter i and $N^i_{Dyslexic}$ is the number of times dyslexic individuals misread the letter i. The results are shown in Figure 2. Note that a negative ratio means that the corresponding letter is misread at a higher rate by nondyslexic participants. We observe that the letter "j" is the most frequently misread letter by dyslexic individuals. Interestingly, there exist some letters which are misread more frequently by nondyslexic individuals. We note that this might be because of mistakes of speech-to-text conversion. We provide further discussion on this issue below. Nevertheless, our results show that the mistakes can be at different levels for each letter. Therefore, it might be important to design the text to be read for the diagnosis process.

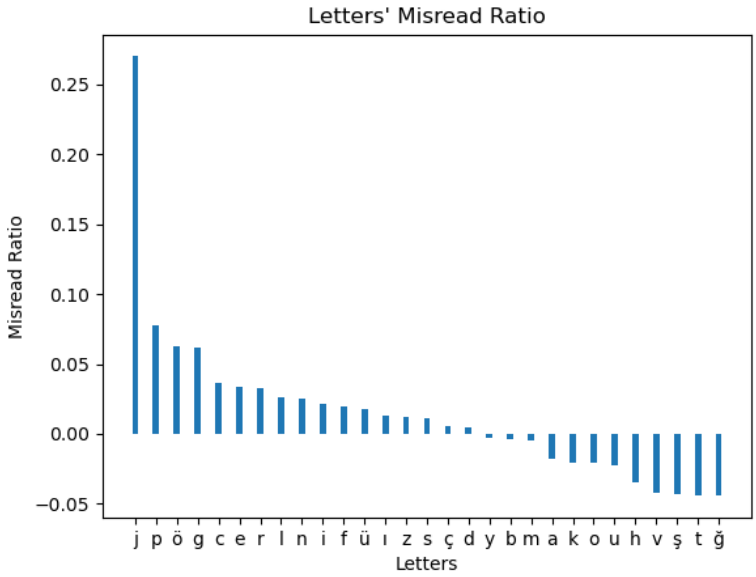

Figure 2. Letters' misread ratio.

As Turkish is an agglutinative language, words might have several suffices affecting the semantics and syntactic roles of words. Furthermore, words with many suffices can be harder to read due to their length. Therefore, we also detected the most misread words and syllables for dyslexic and nondyslexic individuals. Figures 3 and 4 show the word clouds for misread words and syllables for each group, respectively, where words/syllables with higher misread ratio are shown in larger fonts. In order to find out which words and syllables are misread especially by dyslexic individuals, we generated additional word clouds (Figures 3c and 4c) which are generated by taking the difference of the misread rates by dyslexic individuals from the ones by nondyslexic individuals. This approach helps eliminate the errors caused by STT and reveals the words and syllables in our dataset that are difficult to be read correctly by dyslexic individuals.

(a) Dyslexic **(b)** Nondyslexic **(c)** Normalized

Figure 3. Word clouds show words misread by participants. In (a), you can see words misread by Dyslexic participants. (b) shows words misread by nondyslexic participants. The word cloud in (c) is created by subtracting the number of misreadings of the word by nondyslexic participants from the number of misreadings of the word by dyslexic participants.

(a) Dyslexic **(b)** Nondyslexic **(c)** Normalized

Figure 4. Word clouds show syllables misread by participants. In (a), you can see syllables misread by dyslexic participants. (b) shows syllables misread by nondyslexic participants. The word cloud in (c) is created by subtracting the number of misreadings of the syllable by nondyslexic participants from the number of misreadings of the syllable by dyslexic participants.

The words misread by nondyslexic participants might be due to the limitations and mistakes of STT library. Therefore, we randomly selected 5 audio records from each group, i.e. dyslexics and nondyslexics, and listened to them to check whether the words and syllables considered misread are due to the individuals or STT. We focused on the five words and five syllables with the highest misread ratio for each group. The five words with the highest misread ratio for dyslexic and nondyslexic participants are {'can', 'işiten', 'adamsa', 'gülümsemeyle' and 'birikintisinin'} and {'gülümsemeyle', 'rüzgarları', 'bahçesindekilere', 'çama', 'ışıklarıyla'}, respectively.

We observed that the most misread five words of dyslexic participants are actually incorrectly transcribed by the STT in three cases. Regarding nondyslexic participants, the most misread five words appear 12 times in total and 10 of these words are actually read correctly by the participants.

The five syllables with the highest misread ratio for dyslexic and nondyslexic participants are {'can', 'öp','fiy', 'siy', 'yin'} and {'siy', 'sıy','yen', 'kış', 'tıp'}, respectively. The most misread five syllables of dyslexic participants appeared nine times in the recordings that we manually checked and four of them are really misread by the participants. On the other hand, the most misread five syllables of nondyslexic participants appeared six times in our analysis and only one of these words was really misread.

Overall, in 42.86% of the cases of dyslexic participants where the converted text do not match with the ground truth, the participants really misread the given text, whereas the nondyslexic participants really misread only 16.66% of the cases that we inspected.

In our further analysis of the mistakes of STT, we noticed some challenges due to the Turkish language. For instance, the word "bahçesindekilere" (meaning: to those which are in his/her garden) has been correctly read by the participants. However, the STT converted it to "bahçesinde kilere", i.e. inserting a space within the word. While these two words are also valid Turkish words ("bahçesinde" means "in his/her garden" and "kilere" means "to cellar"), the STT is not able to form the longer word and produces two semantically unrelated words. Furthermore, we noticed that the linguistic phenomenon called haplology also was one of the reasons that caused STT to fail. In particular, the conjunction 'ile' ("with") can be written as a separate word. However, if the previous word ends with a vowel, it can also be written as a suffix where the letter 'i' is turned into 'y'. For instance, "arabayla" and "araba ile" are both valid phrases and have exactly the same meaning ("with (a) car"). In our analysis, we found out that in the text participants read, the conjunction 'ile' is written as a suffix to the previous word, but the STT mostly transcribes it separately. For instance, in Figure 4, we notice many words in this form (e.g., gülümsemeyle, annesiyle, keyfiyle, edayla, ışıklarıyla, and soğumasıyla). We also observe that 'fiy', 'siy', and 'sıy' are among the syllables which are frequently detected as misread. Therefore, our analysis suggests that the text read by the participants should be designed carefully to reduce the negative impact of these special linguistic cases. Furthermore, as more advanced STT tools are developed, it is likely that the performance of our model will increase.

8. Conclusion

In this paper, we present the first dataset for detecting dyslexia in Turkish-speaking individuals using their audio records. In addition, we propose a machine learning model which uses reading accuracy, MFCC, and reading rate as features. In our comprehensive experiments, we achieve an accuracy of 95.63% and an F_1 score of 96.03% in detecting dyslexia.

The feature ablation experiment reveals that reading rate is the least effective feature, while MFCC (Mel-Frequency Cepstral Coefficients) is the most effective. This information helps us understand the significance of different features in dyslexia classification. Increasing the number of sentences in each recording improves model performance in most datasets, indicating the importance of capturing more reading patterns in dyslexic participants. Comparing the model's accuracy with human performance, the models exhibit high prediction accuracy, outperforming even experts who rely on single-sentence prediction. Features such as reading accuracy and reading rate which are commonly used by experts when diagnosing dyslexia are found to yield accuracies similar to that of the experts in the human evaluation experiment, while with the additional features, the proposed approach increases its detection performance which can guide future studies in terms of diagnostic methods. Overall, the machine learning-based results provide valuable insights for dyslexia diagnosis using voice recordings. The findings emphasize the importance of specific features, recording length, and attention to linguistic factors in the assessment process.

As our model uses only audio records of children, it can be easily implemented as a smart phone application and can be used by many people. While our model cannot be used for diagnosis of dyslexia, it can be used as a warning system such that the parents of children who are likely to be dyslexic according to our model can seek examination of child psychiatrists. Thus, dyslexic children can be diagnosed and access special education in the early stages of dyslexia, which might potentially make a huge impact on individuals' life.

In addition to estimating dyslexia, we have performed a qualitative analysis and demonstrated mostly misread letters, words, and syllables in Section 7. The outcomes of this analysis have the potential to provide insights into dyslexia estimation based on voice records. It also contributes to the understanding of dyslexia-

related reading challenges in Turkish and provides valuable information for further research and improvements in the field of dyslexia detection.

There are some limitations of our study. The program we developed may serve as a preliminary tool for identifying potential reading difficulties in children. However, it is not intended to be a substitute for professional diagnostic assessments. The possibility of achieving more accurate results lies in the development of a more comprehensive dataset. For instance, incorporating slow-reading students without dyslexia into the system may lead to more precise outcomes. It should be noted, though, that the article does not delve into making specialized inferences about other difficulties, such as reading comprehension, as it falls outside the scope of the study.

Regarding future work, our work can be extended in several directions. Firstly, while the datasets used in training directly impact the performance of models, we experienced that collecting audio records of dyslexic children is an extremely challenging and long process. Therefore, we plan to extend the dataset. Having larger datasets will also enable utilizing deep learning models. Secondly, each language has different phonological features and might cause different challenges in reading and writing. Therefore, we leave building datasets and developing models for other languages as future work.

References

[1] Lyon GR, Shaywitz SE, Shaywitz BA. A definition of dyslexia. Annals of dyslexia 2003; 53 (1): 1–14.

[2] Shaywitz SE, Gruen JR, Shaywitz BA. Management of dyslexia, its rationale, and underlying neurobiology. Pediatric Clinics of North America 2007; 54 (3): 609–623.

[3] Diagnostic and statistical manual of mental disorders : DSM-5™. American Psychiatric Publishing, a division of American Psychiatric Association, Washington, DC, 5th edition. 2013 - 2013. ISBN 9780890425541.

[4] Shaywitz SE. Dyslexia. New England Journal of Medicine 1998; 338 (5): 307–312.

[5] Tanaka H, Black JM, Hulme C, Stanley LM, Kesler SR et al. The brain basis of the phonological deficit in dyslexia is independent of IQ. Psychological science 2011; 22 (11): 1442–1451.

[6] Hamid SSA, Admodisastro N, Kamaruddin A. A study of computer-based learning model for students with dyslexia. In 2015 9th Malaysian Software Engineering Conference (MySEC). IEEE, 2015; 284–289.

[7] Cui Z, Xia Z, Su M, Shu H, Gong G. Disrupted white matter connectivity underlying developmental dyslexia: a machine learning approach. Human brain mapping 2016; 37 (4): 1443–1458.

[8] Płoński P, Gradkowski W, Altarelli I, Monzalvo K, van Ermingen-Marbach M et al. Multi-parameter machine learning approach to the neuroanatomical basis of developmental dyslexia. Human Brain Mapping 2017; 38 (2): 900–908.

[9] Frid A, Manevitz LM. Features and machine learning for correlating and classifying between brain areas and dyslexia. arXiv preprint arXiv:1812.10622 2018; .

[10] Perera H, Shiratuddin MF, Wong KW, Fullarton K. EEG signal analysis of writing and typing between adults with dyslexia and normal controls. International Journal of Interactive Multimedia and Artificial Intelligence 2018; 5 (1): 62.

[11] Rezvani Z, Zare M, Žarić G, Bonte M, Tijms J et al. Machine learning classification of dyslexic children based on EEG local network features. BioRxiv 2019; page 569996.

[12] Rello L, Ballesteros M. Detecting readers with dyslexia using machine learning with eye tracking measures. In Proceedings of the 12th International Web for All Conference. 2015; pages 1–8.

[13] Spoon K, Crandall D, Siek K. Towards detecting dyslexia in children's handwriting using neural networks. In Proceedings of the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, CA, USA. 2019; 1–5.

[14] Rello L, Ballesteros M, Ali AX, Serra M, Sánchez DA et al. Dytective: diagnosing risk of dyslexia with a game. In PervasiveHealth. 2016; 89–96.

[15] Rello L, Romero E, Rauschenberger M, Ali A, Williams K et al. Screening dyslexia for english using hci measures and machine learning. In Proceedings of the 2018 international conference on digital health. 2018; 80–84.

[16] Omar KS, Mondal P, Khan NS, Rizvi MRK, Islam MN. A machine learning approach to predict autism spectrum disorder. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2019; 1–6.

[17] Stevens E, Dixon DR, Novack MN, Granpeesheh D, Smith T et al. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. International journal of medical informatics 2019; 129: 29–36.

[18] Wall DP, Kosmicki J, Deluca T, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Translational psychiatry 2012; 2 (4): e100–e100.

[19] Rahman MM, Usman OL, Muniyandi RC, Sahran S, Mohamed S et al. A review of machine learning methods of feature selection and classification for autism spectrum disorder. Brain Sciences 2020; 10 (12). ISSN 2076-3425.

[20] Al Hanai T, Ghassemi MM, Glass JR. Detecting depression with audio/text sequence modeling of interviews. In Interspeech. 2018; pages 1716–1720.

[21] Cantürk İ, Karabiber F. A machine learning system for the diagnosis of Parkinson's disease from speech signals and its application to multiple speech signal types. Arabian Journal for Science and Engineering 2016; 41 (12): 5049–5059.

[22] Vizza P, Tradigo G, Mirarchi D, Bossio RB, Lombardo N et al. Methodologies of speech analysis for neurodegenerative diseases evaluation. International journal of medical informatics 2019; 122: 45–54.

[23] Laguarta J, Hueto F, Subirana B. Covid-19 artificial intelligence diagnosis using only cough recordings. IEEE Open Journal of Engineering in Medicine and Biology 2020; 1: 275–281.

[24] Richard G, Mathieu S. Dyslexia and Dysgraphia prediction: A new machine learning approach. arXiv preprint arXiv:2005.06401 (2020).

[25] Žabkar J, Urankar T, Javornik K, Babuder MK. Identifying Reading Fluency in Pupils with and without Dyslexia Using a Machine Learning Model on Texts Assessed with a Readability Application. Center for Educational Policy Studies Journal. 2023.

[26] Atkar G, Priyadarshini J. Advanced machine learning techniques to assist dyslexic children for easy readability. International Journal of Scientific & Technology Research,(IJSTR), ISSN 2020; pages 2277–8616.

[27] Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. JMIR mental health 2017; 4 (2): e7785.

[28] Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth and uHealth 2018; 6 (11): e12106.

[29] Zahia S, Garcia-Zapirain B, Saralegui I, Fernandez-Ruanova B. Dyslexia detection using 3d convolutional neural networks and functional magnetic resonance imaging. Computer Methods and Programs in Biomedicine 2020; 197: 105726.

[30] Nilsson Benfatto M, Öqvist Seimyr G, Ygge J, Pansell T, Rydberg A et al. Screening for dyslexia using eye tracking during reading. PloS one 2016; 11 (12): e0165508.

[31] Lakretz Y, Chechik G, Friedmann N, Rosen-Zvi M. Probabilistic graphical models of dyslexia. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015; pages 1919–1928. @articlesainburg2020finding, title=Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires, author=Sainburg, Tim and Thielk, Marvin and Gentner, Timothy Q, journal=PLoS computational biology, volume=16, number=10, pages=e1008228, year=2020, publisher=Public Library of Science

[32] Usman OL, Muniyandi RC, Omar K, Mohamad M. Gaussian smoothing and modified histogram normalization methods to improve neural-biomarker interpretations for dyslexia classification mechanism. Plos one 2021; 16 (2): e0245579.

[33] Usman OL, Muniyandi RC, Omar K, Mohamad M. Advance machine learning methods for dyslexia biomarker detection: A review of implementation details and challenges. IEEE Access 2021; 9: 36879–36897.

[34] Di Pietro SV, Willinger D, Frei N, Lutz C, Coraj S et al. Disentangling influences of dyslexia, development, and reading experience on effective brain connectivity in children. NeuroImage 2023; 268: 119869. ISSN 1053-8119.

[35] Rauschenberger M, Baeza-Yates R, Rello L. A universal screening tool for dyslexia by a web-game and machine learning. Frontiers in Computer Science 2022; 3. ISSN 2624-9898.

[36] Spoon K, Siek K, Crandall D, Fillmore M. Can we (and should we) use ai to detect dyslexia in children's handwriting? In Proc. Artif. Intell. Social Good (NeurIPS). 2019; 1–6.

[37] Hamid SSA, Admodisastro N, Manshor N, Kamaruddin A, Abd Ghani AA. Dyslexia adaptive learning model: student engagement prediction using machine learning approach. In International Conference on Soft Computing and Data Mining. Springer, 2018; 372–384.

[38] Levenshtein VI et al. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, volume 10. Soviet Union, 1966; pages 707–710.

[39] Sainburg T, Thielk M, Gentner TQ. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLoS computational biology. 2020 Oct 15;16 (10):e1008228.