Automatic detection of dyslexia based on eye movements during reading in Russian

Anonymous ACL submission

Abstract

Dyslexia, a common learning disability, requires an early diagnosis. However, current screening tests are very time- and resource-consuming. We present an LSTM model that aims to automatically classify dyslexia based on eye movements recorded during natural reading combined with basic demographic information and linguistic features of the fixated words. The proposed model outperforms the state-of-the-art model and reaches the AUC of 0.93. We additionally discuss the outcomes of several ablation studies assessing which features are critical for model performance.

1 Introduction

One of the most common learning disabilities is dyslexia, a difficulty that specifically affects reading and spelling in individuals with otherwise intact cognitive abilities. The origin of the difficulty is believed to lie in phonological decoding (International Dyslexia Association, 2024). The prevalence of dyslexia is estimated to be between 9% and 12% (Katusic et al., 2001; Shaywitz et al., 1998). Early diagnosis is the key factor for getting the needed support and staying on track in the educational system (Glazzard, 2010; Torgesen, 2000; Vellutino et al., 2004).

Various testing batteries exist, but most must be administered by a trained specialist, who is not always present at school. Moreover, such batteries are still often evaluated using paper-and-pencil methods, which are time-consuming and error-prone. Without a cheap, fast, and reliable mass testing method, the only way to get proper support for a struggling reader is through the educator, who may notice reading difficulties and recommend additional testing. This route crucially depends on the educator, and will fail more often for educators who are overworked. In particular, that means that reading difficulties will more often be left unmitigated in the already disadvantaged districts and schools.

Several machine-learning solutions have already been proposed for the mass screening for dyslexia based on eye movements recorded during naturalistic reading (Asvestopoulou et al., 2019; Nilsson Benfatto et al., 2016; Haller et al., 2022; Jothi Prabha and Bhargavi, 2022; Raatikainen et al., 2021; Rello and Ballesteros, 2015; Shalileh et al., 2023). Yet almost all of these models were trained on very modest, at least by machine learning standards, samples of 61 (Asvestopoulou et al., 2019) to 185 participants (Nilsson Benfatto et al., 2016).

This paper presents a comparison of two models that aim to automatically classify dyslexia on a dataset comprising eye-movements while reading from 293 young readers of different ages.

2 Problem Setting

We study the task of inferring whether a child has dyslexia from eye movements and a stimulus text that was presented during the recording of the eye movements. Since this problem is a binary classification task, the model's performance can be characterized by a false positive rate and a true positive rate. By altering the decision threshold, one can observe a receiver operator characteristic curve (ROC curve). The area under the ROC curve (AUC) provides an aggregated measure of performance for all possible classification thresholds.

3 Methods

3.1 Reference method

As a baseline, we use a state-of-the-art (SOTA) SVM-RFE with a linear kernel described and implemented by Haller et al. (2022). This approach was first proposed by Nilsson Benfatto et al. (2016), who reported 96% accuracy on a balanced dataset. As input, the SOTA model uses the means and standard deviations of 12 eye-movement features, such

as first fixation duration, first-pass reading time, etc. (for the full list, refer to Haller et al. 2022). In the reader-prediction setting (see Section 4.3), 12 features are aggregated across all sentences read by a given participant. In the sentence-prediction setting, the same eye-movement features are aggregated within each separate sentence read by a given participant.

Note that Haller et al. had a homogenous data set of age-matched readers, and they did not include either age or grade into the model. Given that grade is an important predictor of reading skill, and the present dataset includes readers from grade 1 to 6, we report the performance of the SOTA model both without grade, for full comparability with Haller et al.'s results, and with grade, for a fairer comparison.

3.2 Proposed model

The proposed model input is a participant's fixation sequence on a sentence. Each input vector consists of demographic information, gaze-specific and linguistic features. In total, the fixation vector consists of 26 features: the participant's age, grade, gender, fixation duration, fixation horizontal and vertical coordinates on the screen, fixation landing position on the word, fixated word length in letters, fixated word predictability and frequency, number of morphemes comprising the word, next fixation distance, next saccade amplitude, angle, velocity, and direction.

The proposed model architecture is a bidirectional Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997). The mean of the hidden states is fed into two sequential linear layers, projecting it down to a single sigmoid output to represent the label prediction. Optimized hyperparameters and search space are reported in Appendix A.

4 Experiments

4.1 Eye-movement data

The cross-sectional dataset comprises eye movements while reading in 293 school children, native speakers of Russian, from the 1st to the 6th grade (Shalileh et al., 2023). In Russia, grades 1 through 4 correspond to primary school, and grades 5 and 6 – to secondary school. Based on reading speed and accuracy, children were classified as typically developing ($N = 221$) or having developmental dyslexia ($N = 72$). Classification was based on the Standardized Assessment of Reading Skills test (SARS, Kornev and Ishimova 2010) and recent normative cutoff levels reported by Dorofeeva et al. (2019). The Standardized Assessment of Reading Skills requires a test-taker to read a short text aloud as quickly and as accurately as possible. The number of words read accurately in the first minute is taken as a measure of reading fluency. If a child scores at least 1.5 standard deviations below their corresponding age mean, a dyslexia label is assigned.

For all children, nonverbal intelligence scores were obtained using Raven's Colored Progressive Matrices (Raven, 2003). All children had nonverbal intelligence scores within the normal range.

Typically-developing children. All children in this group had age-appropriate reading fluency and comprehension. The parents or primary caretakers reported no history of reading disorders. The composition of the group can be seen in Table 1.

Grade	1 (N=50)	2 (N=40)	3 (N=37)	4 (N=39)	5 (N=31)	6 (N=24)
Gender						
Female: N (%)	22 (44%)	24 (60%)	19 (51%)	18 (46%)	12 (39%)	10 (42%)
Age						
Mean ± SD	$7.32_{0.51}$	$8.35_{0.48}$	$9.30_{0.46}$	$10.18_{0.56}$	$11.29_{0.78}$	$12.00_{0.59}$
Nonverbal intelligence						
Mean ± SD	$29.88_{3.99}$	$31.00_{3.23}$	$31.24_{3.50}$	$31.90_{3.59}$	$32.81_{2.12}$	$33.17_{2.39}$
Reading speed (wpm)						
Mean ± SD	$63.80_{27.06}$	$79.0_{17.54}$	$95.57_{13.93}$	$119.28_{20.67}$	$122.48_{29.38}$	$124.62_{23.50}$

Table 1: Composition of the control group, split by grade.

Children with developmental dyslexia. In this group, the reading speed was lower than the population's average by at least 1.5 SD. The detailed composition of the group can be seen in Table 2.

Grade	1 (N = 8)	2 (N = 10)	3 (N = 20)	4 (N = 28)	5 (N = 6)
Gender					
Female: N (%)	2 (25%)	2 (20%)	12 (60%)	9 (32%)	2 (33%)
Age					
Mean ± SD	$7.25_{0.46}$	$8.40_{0.84}$	$9.30_{0.57}$	$10.25_{0.59}$	$11.17_{0.41}$
Nonverbal intelligence					
Mean ± SD	$29.75_{4.74}$	$29.00_{3.74}$	$31.40_{5.75}$	$32.14_{3.33}$	$28.50_{4.85}$
Reading speed (wpm)					
Mean ± SD	$17.38_{8.52}$	$30.70_{10.68}$	$52.20_{20.48}$	$57.50_{22.29}$	$56.50_{16.60}$

Table 2: Composition of the group of children with dyslexia, split by grade.

4.2 Reading materials

All children were asked to read the same set of 30 sentences comprising the Child Russian Sentence Corpus (Lopukhina et al. 2022). Reading took them from 10 to 30 minutes. In some cases, the reading session was terminated before a child read all sentences due to various reasons. Since the rate of early termination was somewhat higher in the dyslexia group (presumably due to reading difficulties), we decided to keep the data from the early terminated sessions. The number of sentences each

child read ranged from 10 to 30, with the median of 27; 86% of all eye-movement recordings had fewer than 30 sentences (83% recordings in the control group, 96% recordings in the dyslexia group).

Sentence difficulty was at the level of 3rd to 4th grade, according to an automatic text difficulty measurement developed for Russian (Laposhina and Lebedeva, 2021), and estimated to be 7.42 on the Flesch-Kincaid scale adapted to Russian (Readability Test). The sentences were between six and nine words long ($M = 7.6$, $SD = 0.85$), with 50 characters per sentence ($SD = 5.16$) on average. In total, the children read 227 words, which contained 182 unique word forms (as words could be repeated across sentences). Individual words were on average 5.6 letters long (range 1–13), and had an average lemma frequency of 50.29 items per million (median: 0.73, range: $0.0001 - 667$). The frequency was calculated from the subcorpus of texts for children of the years 1920–2015 of the Russian National Corpus.

Corpus materials were morphologically annotated: 54 words consisted of a single morpheme, 81 words consisted of two morphemes, 45 words – of three morphemes, 34 words – of four morphemes, nine words – of five morphemes, and four words consisted of six morphemes. Finally, for every word in every sentence, word predictability was estimated using an online cumulative cloze task with 46 children (24 girls, $M = 11.3$, range 9–12) who did not participate in the eye-tracking study. Predictability was measured as the number of correct guesses divided by the total number of guesses. Zero cloze probabilities were replaced with $\frac{1}{2}\times$ the number of guesses for the word.

4.3 Model evaluation

The models are evaluated in two settings: prediction of the reader's status based on a single sentence data (sentence prediction setting) or based on all available reading data (reader prediction setting). All models are evaluated and tuned using 10-fold nested cross-validation and random grid search (see Appendix A). Data from the same person is always constrained to one fold, so that the models always make predictions for unseen participants. The ratio of persons with/without dyslexia is balanced across all folds.[1]

[1]All code is available online: https://anonymous.4open.science/r/RDC-A08A/

4.4 Results

For all methods, we report AUC (chosen as a metric invariant to class imbalance, see Richardson et al. 2024) for reader- and sentence-level settings (see Table 3). A visual summary of ROC AUC performance can also be found in Figure 1. For all evaluated models, classification performance in the reader-prediction setting was numerically higher than in the sentence-prediction setting. However, according to an unpaired one-tailed t-test, the difference between settings was not significant in any model or configuration (LSTM: $t(15.55) = 1.22$, $p = 0.12$; $\text{SOTA}_{\text{+Grade}}$: $t(16.21) = 0.81$, $p = 0.21$; $\text{SOTA}_{\text{-Grade}}$: $t(17.83) = 0.24$, $p = 0.41$). The SOTA model that included information about grade performed numerically better than the same model without grade information, but the difference was not significant (reader prediction setting: $t(17.96) = 1.03$, $p = 0.16$; sentence prediction setting: $t(16.64) = 0.71$, $p = 0.24$). Importantly, the proposed LSTM outperformed the $\text{SOTA}_{\text{+Grade}}$ model in both reader-prediction ($t(12.146) = 2.12$, $p = 0.028$) and sentence-prediction settings ($t(17.92) = 2.20$, $p = 0.021$).

4.4.1 LSTM ablation experiments

In the reader-prediction setting, we run three additional ablation studies ($\text{LSTM}_{\text{-Saccade}}$, $\text{LSTM}_{\text{-Ling}}$, and $\text{LSTM}_{\text{-Demographic}}$), assessing model performance without saccade-related measures (next fixation distance, next saccade amplitude, next saccade angle, next saccade velocity, and next saccade direction), without linguistic information (word length, frequency, predictability, and the number of morphemes comprising the word), and without demographic information (age, grade, and gender). In all ablation studies, AUC score was lower numerically, but the decrease was not statistically significant ($\text{LSTM}_{\text{-Saccade}}$: $t(14.70) = 0.95$, $p = 0.17$; $\text{LSTM}_{\text{-Ling}}$: $t(17.80) = 0.06$, $p = 0.47$; $\text{LSTM}_{\text{-Demographic}}$: $t(16.14) = 1.66$, $p = 0.058$).

5 Discussion

The finding that information about reader's grade did not significantly improve either SOTA or LSTM model's performance is rather surprising because in the present dataset, dyslexia was diagnosed based on the age-specific normative cut-offs in reading speed (see Section 4.1). Consequently, information about reader's grade should be crucial for the classification performance. Grade-invariant performance

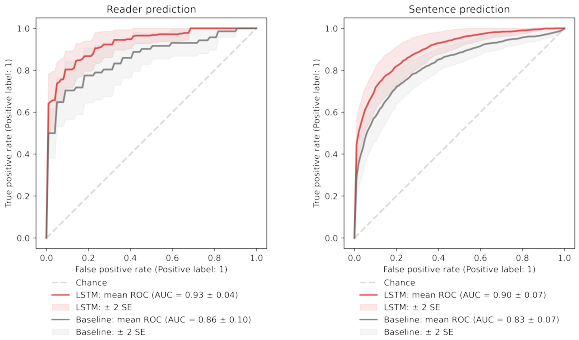

Figure 1: Summary of model performance. SOTA baseline model used grade information.

		AUC
Reader	SOTA	0.86±0.10
	SOTA$_{\text{-Grade}}$	0.81±0.11
	LSTM	0.93±0.05
	LSTM$_{\text{-Ling}}$	0.92±0.05
	LSTM$_{\text{-Saccade}}$	0.91±0.07
	LSTM$_{\text{-Demographic}}$	0.90±0.06
Text	SOTA	0.83±0.07
	SOTA$_{\text{-Grade}}$	0.80±0.10
	LSTM	0.90±0.07

Table 3: Summary of model performance metrics in the reader- and sentence-prediction settings.

might potentially reflect that the model has captured some invariant property of the eye movements of readers with dyslexia that is shared between all grades. For the SOTA model trained exclusively on aggregated features, this explanation is unlikely. For the LSTM model trained on a sequence of separate fixation events, this explanation is more likely, but it is precisely the LSTM that shows greater numerical decrease in performance when the information about grade is removed. Overall, we still believe that a successful model should be able to uncover the relationship between reading speed, grade, and dyslexia label.

Another surprising outcome is the lack of difference between the sentence- and reader-prediction settings in all the tested models. Given that reader-prediction setting relies on $10\times$ to $30\times$ more data, we would expect performance to be higher. The increase in performance may not be significant due to the relatively small size of the dataset and insufficient statistical power. Alternatively, the lack of difference may reflect a true limit in model performance, where additional information beyond a single sentence is of little to no added value.

The finding that removing linguistic features did not significantly affect LSTM model's performance is less surprising. Arguably the most crucial feature for dyslexia classification, a representation of a word's degree of orthographic transparency (Borleffs et al., 2017), was not present in the feature set. Including a measure of word orthographic transparency might prove to be a promising next step in improving model performance.

The lack of difference between the performance of LSTM with and without saccade-related information might indicate that either some saccade-related information is represented by the model implicitly (saccade distance can trivially be represented as a distance between x and y coordinates of two consecutive fixations) or that saccade-related information is irrelevant for classification purposes.

Most importantly, the proposed LSTM outperformed the SOTA model, and, based on the outcomes of ablation experiments, we can conclude that the increase in performance is due to the more detailed information about the eye movements, but not due to added information about the linguistic stimulus.

6 Ethical considerations

Using demographic variables, such as age and gender, could lead to reproducing existing biases. For example, males are diagnosed with dyslexia more frequently than females, but at least part of the difference may be attributed to referral bias (**??**). One way to ensure that the model is bias-free is to withhold the potentially biasing feature. The ablation experiment that removed the demographic information performed on par with the full model. Therefore, we conclude that the model at least does not enhance the bias that might be present in the data set.

7 Conclusions

The model of automatic dyslexia detection proposed in this paper has outperformed the SOTA model. Importantly, unlike most of the models proposed so far (Nilsson Benfatto et al., 2016; Haller et al., 2022; Asvestopoulou et al., 2019; Jothi Prabha and Bhargavi, 2022), the present LSTM was trained on an unbalanced dataset of eye movements of children who were also not age-matched, and might therefore be more robust and potentially more appropriate for the real-world applications.

Limitations

This decision to include information from participants who did not read all 30 sentences could potentially lead to data leakage: The model may learn that incomplete sessions are more likely to come from a child with dyslexia. We think that this is unlikely for two reasons: First, the proportions of incomplete sessions are not drastically different between the two groups. Second, this potential data leakage should only affect the reader-prediction setting (where the model expects to see 30 sentences), not the sentence-prediction setting (where the model expects to see one sentence). In the present case, there was no significant difference in performance between the reader prediction and the sentence prediction settings (see Section 4.4), so the reader-prediction setting is unlikely to have an unfair advantage.

References

Thomais Asvestopoulou, Victoria Manousaki, Antonis Psistakis, Ioannis Smyrnakis, Vassilios Andreadakis, Ioannis M Aslanides, and Maria Papadopouli. 2019. Dyslexml: Screening tool for dyslexia using machine learning. *arXiv preprint arXiv:1903.06274*.

Elisabeth Borleffs, Ben AM Maassen, Heikki Lyytinen, and Frans Zwarts. 2017. Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and writing*, 30:1617–1638.

Svetlana V Dorofeeva, Victoria Reshetnikova, Margarita Serebryakova, Daria Goranskaya, Tatiana V Akhutina, and Olga Dragoy. 2019. Assessing the validity of the standardized assessment of reading skills in russian and verifying the relevance of available normative data. *The Russian Journal of Cognitive Science*, 6(1):4–24.

Jonathan Glazzard. 2010. The impact of dyslexia on pupils' self-esteem. *Support for learning*, 25(2):63–69.

Patrick Haller, Andreas Säuberli, Sarah Elisabeth Kiener, Jinger Pan, Ming Yan, and Lena Jäger. 2022. Eye-tracking based classification of Mandarin Chinese readers with and without dyslexia using neural sequence models. *arXiv preprint arXiv:2210.09819*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

International Dyslexia Association. 2024. Dyslexia basics. https://dyslexiaida.org/. Accessed: 2024-07-03.

A Jothi Prabha and R Bhargavi. 2022. Prediction of dyslexia from eye movements using machine learning. *IETE Journal of Research*, 68(2):814–823.

Slavica K Katusic, Robert C Colligan, William J Barbaresi, Daniel J Schaid, and Steven J Jacobsen. 2001. Incidence of reading disability in a population-based birth cohort, 1976–1982, rochester, minn. In *Mayo Clinic Proceedings*, volume 76, pages 1081–1092. Elsevier.

AN Kornev and OA Ishimova. 2010. Metodika diagnostiki disleksii u detey [methods of diagnosis of dyslexia in children]. *St. Petersburg: Publishing house of the Polytechnic University*.

Antonina N Laposhina and Maria Yu Lebedeva. 2021. Textometr: An online tool for automated complexity level assessment of texts for russian language learners. *Russian Language Studies*, 19(3):331–345.

Anastasiya Lopukhina, Nina Zdorova, Vladislava Staroverova, Nina Ladinskaya, Anastasiia Kaprielova, Sofya Goldina, Olga Vedenina, Ksenia Bartseva, and Olga Dragoy. 2022. Benchmark measures of eye movements during reading in russian children.

Mattias Nilsson Benfatto, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson. 2016. Screening for dyslexia using eye tracking during reading. *PloS one*, 11(12):e0165508.

Peter Raatikainen, Jarkko Hautala, Otto Loberg, Tommi Kärkkäinen, Paavo Leppänen, and Paavo Nieminen. 2021. Detection of developmental dyslexia with machine learning using eye movement data. *Array*, 12:100087.

Jean Raven. 2003. Raven progressive matrices. In *Handbook of nonverbal assessment*, pages 223–237. Springer.

Readability Test. Readability test for russian texts. Accessed: 2024-07-04.

Luz Rello and Miguel Ballesteros. 2015. Detecting readers with dyslexia using machine learning with eye tracking measures. In *Proceedings of the 12th international web for all conference*, pages 1–8.

Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. 2024. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6):100994.

Russian National Corpus. Russian national corpus. Accessed: 2024-07-04.

Soroosh Shalileh, Dmitry Ignatov, Anastasiya Lopukhina, and Olga Dragoy. 2023. Identifying dyslexia in school pupils from eye movement and demographic data using artificial intelligence. *Plos one*, 18(11):e0292047.

Sally E Shaywitz, Bennett A Shaywitz, Kenneth R Pugh, Robert K Fulbright, R Todd Constable, W Einar Mencl, Donald P Shankweiler, Alvin M Liberman, Pawel Skudlarski, Jack M Fletcher, et al. 1998. Functional disruption in the organization of the brain for reading in dyslexia. *Proceedings of the National Academy of Sciences*, 95(5):2636–2641.

Joseph K Torgesen. 2000. Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning disabilities research & practice*, 15(1):55–64.

Frank R Vellutino, Jack M Fletcher, Margaret J Snowling, and Donna M Scanlon. 2004. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, 45(1):2–40.

A Model parameters

Model search space is summarized in Table 4.

Batch size	8, 16, 32, 64, 128
Learning rate	$15 \times \mathcal{U} \sim (1e^{-5}, 1e^{-1})$
LSTM hidden layer size	30, 40, 50, 60, 70

Table 4: Hyperparameter search space.

The optimal parameters can be found in Table 5.

Batch size	Learning rate	Hidden layer size
Reader prediction setting		
64	0.001	40
16	0.001	40
64	0.01	30
64	0.001	40
64	0.001	40
64	0.001	40
16	0.01	30
32	0.01	30
16	0.01	50
64	0.001	40
Sentence prediction setting		
8	$7.07e^{-05}$	30
8	0.0003	50
128	$4.21e^{-05}$	70
64	0.0025	50
8	$7.07e^{-05}$	30
64	0.0025	50
128	0.0025	30
8	$7.07e^{-05}$	30
32	$5.34e^{-05}$	70
8	$4.21e^{-05}$	50

Table 5: Resulting optimal parameters.