RESEARCH ARTICLE

**Identifying dyslexia in school pupils from eye movement and demographic data using artificial intelligence**

**Fatma ashraf, Salma Harb1\*, Dmitry Ignatov2, Anastasiya Lopukhina3, Olga Dragoy1,4**

**1** Center for Language and Brain, HSE University, Moscow, Russia, **2** School of Data Analysis and Artificial Intelligence, Faculty of Computer Science, Moscow, Russia, **3** Rastle lab, Royal Holloway, University of London, London, United Kingdom, **4** Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

**Abstract**

This paper represents our research results to achieve the following goals: (i) implement a new multi-source dataset to overcome the shortcomings of previous datasets, (ii) propose a robust AI-based solution for dyslexia detection in primary school students, (iii) explore our psycholinguistic knowledge by examining features importance in the best detection of dyslexia using our AI model. To achieve the first goal, we collected and annotated a new set of data during eye movements and reading. In addition, we collected demographic information, including nonverbal intelligence, to form three data sources... Our dataset is the largest eye movement dataset worldwide. Unlike before for the second purpose, we formulated the dyslexia prediction task as regression and classification problems and examined the performance of 12 classification and eight regression methods. We used a Bayesian optimization method to refine the hyperparameters of the models: and reported the mean and standard deviation of our estimation metrics through stratified tenfold cross-validation. Our research showed that multilayer perceptron, random forest, gradient boosting, and k-nearest neighbor form the group with the most acceptable results. Furthermore, we showed that while using each data source separately did not produce accurate results, their combination led to a reliable solution. We also determined the importance of the features of our best classifier: our results showed that IQ, gender, and age are the three most important features; We have also shown that fixation along the y-axis is more important than other fixation details. Dyslexia detection, eye fixation, eye movement, demographics, classification, regression, artificial intelligence.

**Introduction: Background, previous work, and motivation**

Developmental dyslexia is a learning disorder characterized by a specific reading disorder despite normal intelligence and oral language ability [1]. Children with dyslexia suffer from slow and difficult reading and word recognition problems; therefore, text comprehension, the goal of reading, is impossible to achieve, which negatively affects educational success, mental health, and social integration [2, 3]. Therefore, early detection of dyslexia is very important for the implementation of appropriate interventions [4-6]. Traditionally, dyslexia is detected during a formal assessment that covers several linguistic and cognitive tasks using phonological and visuospatial processing., reading words, non-words and texts, spelling etc. [7]. Such assessment batteries require a trained expert (usually a psychologist or other educational specialist), are time-consuming, and require children's overt responses during some toxic behaviors (e.g., reading nonwords). Thus the a review of the application of artificial intelligence is dyslexia detection. MRI, fMRI, facial video or image, reading test errors, test scores, EEG and eye tracking are seven types of data used to train AI algorithms. Regarding the number of individual datasets in our data, the three most common are eye-tracking-based (seven datasets, including the present study), EEG (six datasets), and MRI (five datasets). Data types used. Artificial intelligence methods for dyslexia detection the author of publication [9] examined 13 artificial intelligence-based solutions for dyslexia detection at the end of 2019. A recent study [8] deeply reviewed 22 resolutions until the beginning of 2021, including 13 trials. According to [8], support vector machines [10], artificial neural networks, and random forests [11] are, in descending order, the three most used AI classification algorithms. Calculation of precision, accuracy and recall using a 10-fold cross-validation procedure is the most used evaluation framework. Most articles published after the previous two studies also follow similar guidelines. Specifically, El Hmimdi et al. (2021) [12] analyzed raw eye tracking datasets [13, 14] and proposed new eye image parameters as input features to their set of classical classification algorithms and achieved approximately 82% accuracy. Raatikainen et al. [15] presented a new eye movement dataset for Finnish indigenous peoples. They used random forests to find the most informative features and fed them into support vectors to detect dyslexia. AlGhamdi 2022 [16] used a publicly available data set [17] derived from online gaming test results and proposed a new group recommendation for dyslexia detection with almost perfect classification accuracy, while Kaisar and Chowdhury (2022) [18] used the same data set. achieved an initial lower accuracy, then systematically investigated the effects of different oversampling methods, and proposed a hybrid method using oversampling and ensemble learning, which then achieved higher accuracy, although not as high as Alghamdi’s results. The authors of [19] ] collected handwritten signs. . data from Chinese children and created a multilevel multivariate model. Vajs et al. [20] used the VGG16 neural network [21] on a slightly different version of the proposed dataset [22] and achieved 87% accuracy. Later, in [23], they proposed a new feature space and obtained a logistic regression ROC with an AUC value of0.96. These authors confirmed previous results from two different datasets [24]. Previous eye tracking studies of Russian-speaking children with and without dyslexia are few and focused on comparisons of fixation durations, progressive saccades and regressions in these studies. two groups of participants ([25–27]). Their results were consistent with those of other studies conducted in alphabetic languages. Namely, all three studies agreed that dyslexic children produced longer fixations and were more sensitive to word length and frequency compared to typically developing readers. Also, Parshina et al. (2022) used the Scan Path method to investigate which global reading processes 1-5. a class of children with and without dyslexia who learned to read whole sentences. The authors identified five reading processes and concluded that children with dyslexia relied on the same processes as their typically developing peers, but with a reading delay of 3 years. Importantly, previous studies on reading in Russian have never attempted to classify readers with or without dyslexia based on their eye movements.

In the upcoming sections we tackled the motivation, datasets, flowchart, computational setting, apparatus, demographic data, methods under consideration, feature importance, test results, conclusion.

# **Motivation and contribution**

faster than traditional dyslexia diagnostic methods, most of them have several disadvantages. This study aimed to (i) address some of the shortcomings of previously developed solutions, (ii) propose a reliable AI-based solution for dyslexia detection, and (iii) explore psycholinguistic information using our best AI model. To clarify our goals and contribution, we first review the shortcomings of previous solutions in detail. We classify these flaws into (i) data related and (ii) artificial intelligence related. Among plausible data types, it is natural to use eye movement data to analyze reading disorders such as dyslexia, and we focus on this data type. We summarized the characteristics of the six previously reported eye movement datasets in Table 1. Regarding data-related issues, the following findings in Table 1 require special attention: (i-a) size of datasets, (i-b) synthetically balanced dataset - excluding [ 15 ] , (i-c) characteristics of target values , (i-d) age distribution and (i-e) limited to a specific language. Specifically (i-a), it is well known that the greater the amount of data, the greater the ability of the AI model to recognize patterns [31, 32]. Our dataset is the largest of its kind, so it should increase the power of AI models. Regarding (i-b), although synthetic balancing of data representations is a popular method for class imbalance problems; To our knowledge, there is no strict mathematical definition for deciding which samples should be selected for further sampling. Current techniques can result in the model giving more weight to some data points in the synthetically manipulated data, and there is no guarantee that the new data representation matches the true distribution of the unknown underlying data. The results of a recent and comprehensive review [33] on this topic are partially compatible with our idea and confirmed our claim. That's why we increased the size of our dataset and kept a somewhat unbalanced data presentation.

In (i-c), instead of binary class data of typically developing and dyslexic, we introduced three categories: 1) typically developing readers, 2) readers at low risk for dyslexia, and 3) readers at high risk for dyslexia. ; In addition, we introduced a continuous target variable of reading speed, which is a direct measure of reading aloud. This arrangement allowed us to formulate the problem as both classification and regression tasks. Our goal was to create a margin between the two traditional categories by introducing a low-risk category. Our dataset (i-d) covers a wider age group, so we expected to detect dyslexia in its earlier stages at school. students Therefore, our newly introduced dataset can be considered as our first contribution. Our second main goal and contribution addresses the shortcomings of AI-related previous releases. To our knowledge, this is the most comprehensive empirical study investigating the performance of 12 classification and eight regression methods for AI-aided dyslexia detection. All considered AI methods are fine-tuned by the Bayesian optimization search method, and the corresponding tuned values are accurately presented. Our third goal and contribution is to apply Shapley's additive explanatory method. determine the importance of AI methods in this field of research to explore our psycholinguistic knowledge using our best AI method.

# **Data sets**

Part of the current data set, e.g., data from 144 participants, was reported in [27]. In this article, the authors analyze the eye movements of typical readers and dyslexic children using Scan Path [34] and clustering methods. This study has different objectives and adds data on 163 new participants. All data collection for this study was approved by the HSE Interuniversity Research and Empirical Research Ethics Committee and carried out in accordance with the Declaration of Helsinki (World Medical Association, 2013). Participants were recruited between March 2020 and March 2022. Their parents signed informed consent prior to the study. The authors have access to the eye tracking and behavioral data of the participants, their age, grade, gender, and social security. They do not have access to data that could identify individual participants. The entire dataset used in this study, as well as the Python code to implement all the methods and metrics considered, is publicly available in the following GitHub repository: https://github .com/Sorooshi/DD.

# **Flowchart**

A diagram of a process

Description automatically generated

#### **Computational setting**

Their computations consisted of two components (i) fine-tuning the hyperparameters of the methods under consideration and (ii) a comprehensive evaluation of the fine-tuned methods.

A diagram of a model

Description automatically generated

# **Apparatus and stimuli**

The eye-tracking dataset was collected under well-controlled experimental conditions. Participants' eye movements were recorded using an Eye link 1000 Plus or Eye link Portable Duo eye tracker (SR Research, Canada) at a sampling rate of 1000 Hz. Participants were seated 55 cm from the camera and their heads were secured with a chin rest. Only right eye movements were monitored [35]. Natural reading ability was measured: participants silently read 30 different sentences from the child's Russian sentence corpus [36, 37]. The selected sentences were suitable for primary school children and had a diverse grammatical structure specific to the reader. The sentences were presented to each participant in random order. Ten sentences were followed by a two-choice comprehension question to check engagement in the task. The task took approximately 20 minutes to complete. All participants' data were included in the analysis, regardless of their accuracy on the comprehension questions. Using Eye Link Data Viewer 4.2.1 (Oakville, Ontario, Canada: SR Research Ltd.), we generated a fixation report (also called eye fixation in this article), a region of interest report (also referred to). to as IA or IA data in this article) from those collected for raw eye movements, see [38, 39] for details. Fixation, IA, and demographic data, including a measure of nonverbal intelligence (IQ), consisted of the three sources. of the dataset presented in this article. We combined demographic data with fixed and impact assessment reports to test the added value of demographic data to eye tracking data. We also tested each participant in a standardized way to obtain an independent, direct, and continuous reading. Reading Skills Assessment Tool (SARS) [40]. The children had to read a 227-word text ("How I got crabs") in printed form as quickly and accurately as possible. The number of words read accurately in the first minute was taken as a measure of each child's reading ability.

# **Demographic data**

Our data includes 307 Russian-speaking elementary school students from the first to the sixth grade. All children had different but age-appropriate non-verbal intelligence as assessed by Ravens color progression matrices [41]. Parents of the participants did not report abnormal vision and had no neurological or psychiatric disorders. They also confirmed that their children were monolingual. Based on the SARS test [40] and recent normative cutoffs obtained [42], individual reading scores were divided into three groups: 1) typically developing children (TD);2 ) children who are at risk of developmental disability (DR); 3) children with developmental disabilities (DD). The TD group, which we sometimes call typical readers in this article, consists of 213 students, 100 girls and 113 boys. There are 22 students in the DR group, seven girls and 15 boys. There are 72 students in the DD group, 27 girls and 45 boys. We summarized the characteristics of our data set in Table 2. The DR group includes those students whose reading scores on the SARS were 1–1.5 standard deviations (SD) below the population mean. There are 22 students in the group, seven girls and 15 boys. The last group, DD, consists of students whose reading speed was less than 1.5 SD of the population mean performance. There are 72 students in the group, 27 girls and 45 boys. Boundaries between groups were based on SARS test guidelines.

# **Methods under consideration**

One of the main objectives of this study was to conduct a comprehensive set of experiments to empirically investigate the performance of different AI methods to find a robust AI-based solution for dyslexia detection. To this end, we investigate the performance of four families of models: (A) artificial neural networks: multilayer perceptron and convolutional neural network; (B) non-parametric: random forest, AdaBoost, Gradient Boosting, k-nearest neighbor, and support vector machines; (C) linear: linear regression and logistic regression; (D) Bayesian: Gaussian, polynomial and complement naive Bayes. Polynomial and complement-naive Bayesian models did not give satisfactory results; therefore, we excluded them from the paper. Although we obtained similar results for both the classification and regression tasks, for the sake of brevity we only focused on explaining the classification tasks. The remainder of this section describes the principles behind the models.

# **Feature importance**

We investigated the importance of features in MLP predictions, one of our best models, using the SHAP library clown notation with the demographic fixed dataset. Due to the computational complexity of the SHAP approach, it was not possible to use the entire trainset as background data, so as suggested by the author of SHAP, we first trained the MLP of the entire trainset and then passed. trained a model with 500 randomly selected data points from the train data distribution as background data for the kernel predictor, and we used the entire test distribution to determine the shape values. The results are illustrated in figure 4. The left side of this figure shows a summary bar graph of the mean absolute SHAP (MAS) values for each feature per class, and the right side shows the hive summary. intrigues Classes TD, DR, and DD. According to the presented results, demographic characteristics accounted for a total of 93.2% of MAS values and fixed data for the remaining 6.8%. Specifically, considering the sub-images presented by MAS summary bars, IQ, age, and gender\_2, i.e., being male, had cumulative MAS values of 0.24, 0.23 and 0.23, respectively. And among the six school grades, the first and fourth grades (both MAS = 0.16) were more important than the others. Considering the summary hive plots (b, d, and f), we found that paternity was the most important trait in predicting TD and DD classes. Age was the most important trait in predicting DR and the second most important feature in detection. TD. class ÄO was the second most important trait in predicting DR and DD and the third most important trait in predicting TD. Its two extremes had an inverse effect on the model forecasts, especially TD and DD forecast. It is worth emphasizing that we believe that demographic characteristics are confounding rather than confounding. A deeper study of this topic is our future work program. In terms of fixation characteristics, fixation on the y-axis had a greater effect with an approximate MAS of 0.052 than fixation duration with an approximate MAS of 0.036 and fixation. along the y-axis. x-axis with a MAS value of 0.025. In our data set, we found that students with dyslexia looked at the screen on average lower than normal readers while reading. They had more eye movement spikes on the y-axis than normal readers. These two reasons justify why the model placed more emphasis on this function.

# **Independent test results**

Because of the real-world importance of the dyslexia screening problem, and before starting a clinical trial of our proposed solution, we evaluated the performance of the tuned MLP classifier on an independent (and newly collected) test set using demographic and fixed data. This new test set consisted of nine typical readers (five girls) and seven students with dyslexia (four girls). For fair evaluation, we randomly selected one of the ten train test groups. We then trained the MLP classifier with hyperparameters previously tuned on the training set. After training, we used two frameworks to evaluate our model. In the first frame, we used the entire independent test set to evaluate the performance of the trained MLP. The Mara frame does not mimic real-world conditions where the percentage of typically developing students is higher than that of students with dyslexia. To address this problem, we determined the number of typical readers and randomly selected three students with dyslexia. We repeated this process ten times and calculated the mean and standard deviation of the metrics.

A group of colorful bars

Description automatically generated with medium confidence

# **Conclusion and future work**

The main goals of this research were (i) to improve the shortcomings of previously implemented datasets, (ii) to propose a robust AI-based solution to detect dyslexia in its early stages, and (iii) to explore our psycholinguistic knowledge. performance of our best AI model. (i) for clarity, the vast majority of previous data consisted of small numbers of participants, and the distribution of control and dyslexic participants was synthetically balanced. More importantly, the age distribution of most datasets is not suitable for developmental delay correction. Therefore, we presented for the first time in Russia a new eye movement dataset consisting of 307 participant data annotated by experts, making it the largest dataset in its category with the most accurate eye tracking data. Not only does this mimic the actual unbalanced data distribution of normal and dyslexic groups, but it also includes a wider and more appropriate age group (first through sixth grade elementary school students). Our dataset consists of three data sources: 1) eye fixations, 2) domain of interest, and 3) demographic information, including IQ. We also introduced a new class, separating traditional dyslexia into patients with low and high risk of developing dyslexia. To achieve (ii), we examined the performance of 12 classification methods (from four model families) in individual subgroups. our dataset and their combination. In all these cases, we fine-tuned the models using the BO method; We then trained and evaluated each model using a tenfold cross-validation procedure and reported the mean and standard deviation of the evaluation metrics. Our experiments showed that although no model obtained completely satisfactory results for detecting dyslexia from each of our individual data sources, CNN with F1 score = 0.637 and ROC AUC = 0.758 obtained the best and relatively satisfactory results for predicting dyslexia in fixed data. And GB got almost the same output Front Ground Results from GB in IA data. The combination of fixation and demographic data sources produced acceptable results for the four models. Specifically, our proposed AI model is MLP with average F1 score = 0.912 and ROC AUC = 0.983, while RF, GB and KNN are also reliable alternatives. When we combined the demographics of interest, we observed the same patterns and results. Regarding the advantages and disadvantages of the approaches used, although our dataset is the\largest of its kind, due to limited training samples, we had to limit our experiments to lower neural networks, which are less prone to over tuning than deeper networks. However, the neural networks used in our experiments performed slightly better than their competitors based on ensemble learning, and interpreting their weights without tools like Shapley values is quite difficult - if not impossible - to interpret and visualize decision trees. with limited size is much easier. To achieve our third goal, we used the SHAP method to determine the importance of one of our best classification functions on a fixed demographic dataset. In short, we found that IQ, age, and masculinity are the three main (and probably confounding) demographic characteristics. We also found that fixation along the y-axis is more important than the x-axis, with the total.

References

1.Frazier M. Dyslexia: Perspectives, challenges, and treatment options. Nova Biomedical; 2016.

2.Undheim AM. A thirteen-year follow-up study of young Norwegian adults with dyslexia in childhood: reading development and educational levels. Dyslexia. 2009; 15(4):291–303. https://doi.org/10.1002/ dys.384 PMID: 19301419

3.Riddick B. Living with dyslexia: The social and emotional consequences of specific learning difficulties/ disabilities. Routledge; 2012.

4.Glazzard J. The impact of dyslexia on pupils’ self-esteem. Support for learning. 2010; 25(2):63–69. https://doi.org/10.1111/j.1467-9604.2010.01442.x

5.Snowling MJ, Hulme C. Interventions for children’s language and literacy difficulties. International Journal of Language & Communication Disorders. 2012; 47(1):27–34. https://doi.org/10.1111/j.1460-6984. 2011.00081.x PMID: 22268899

6.Vellutino FR, Fletcher JM, Snowling MJ, Scanlon DM. Specific reading disability (dyslexia): What have we learned in the past four decades? Journal of child psychology and psychiatry. 2004; 45(1):2–40. https://doi.org/10.1046/j.0021-9630.2003.00305.x PMID: 14959801

7.Roitsch J, Watson SM. An overview of dyslexia: definition, characteristics, assessment, identification, and intervention. Science Journal of Education. 2019; 7(4). https://doi.org/10.11648/j.sjedu.20190704. 11

8.Usman OL, Muniyandi RC, Omar K, Mohamad M. Advance machine learning methods for Dyslexia bio- marker detection: a review of implementation details and challenges. IEEE Access. 2021; 9:36879– 36897. https://doi.org/10.1109/ACCESS.2021.3062709

9.Kaisar S. Developmental dyslexia detection using machine learning techniques: A survey. ICT Express. 2020; 6(3):181–184. https://doi.org/10.1016/j.icte.2020.05.006

10.Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3):273–297. https://doi.org/ 10.1007/BF00994018

11.Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/ A:1010933404324

12.El Hmimdi AE, Ward LM, Palpanas T, Kapoula Z. Predicting dyslexia and reading speed in adolescents from eye movements in reading and non-reading tasks: A machine learning approach. Brain Sciences, 11(10), p.1337. https://doi.org/10.3390/brainsci11101337 PMID: 34679400

13.Kapoula Z, Bucci MP, Jurion F, Ayoun J, Afkhami F, and Bre´mond-Gignac D., Evidence for frequent divergence impairment in French dyslexic children: deficit of convergence relaxation or of divergence per se? Graefe’s Archive for Clinical and Experimental Ophthalmology. 2006; 245:931–936. https://doi. org/10.1007/s00417-006-0490-4 PMID: 17186259

14.Bucci MP., Bre´mond-Gignac D. and Kapoula Z. Poor binocular coordination of saccades in dyslexic children. Graefe’s archive for clinical and experimental ophthalmology. 2008; 246:417–428. https://doi. org/10.1007/s00417-007-0723-1 PMID: 18046570

15.Raatikainen Peter and Hautala Jarkko and Loberg Otto and Ka¨ rkka¨ inen Tommi and Leppa¨ nen Paavo and Nieminen Paavo Detection of developmental dyslexia with machine learning using eye movement data. Array. 2021; 12:100087. https://doi.org/10.1016/j.array.2021.100087