Array 12 (2021) 100087

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

Detection of developmental dyslexia with machine learning using eye movement data[☆]

Peter Raatikainen[a], Jarkko Hautala[b], Otto Loberg[c], Tommi Kärkkäinen[a], Paavo Leppänen[d], Paavo Nieminen[a,*]

[a] *Faculty of Information Technology, University of Jyväskylä, PO Box 35, FI 40014, Finland*
[b] *Niilo Mäki Institute, University of Jyväskylä, PO Box 35, FI 40014, Finland*
[c] *Department of Psychology, Bournemouth University, Talbot Campus, Fern Barrow, Poole, BH12 5BB, United Kingdom*
[d] *Department of Psychology, Faculty of Education and Psychology, University of Jyväskylä, PO Box 35, FI 40014, Finland*

ARTICLE INFO

MSC:
68T05
68T10

Keywords:
Support Vector Machine
Random Forest
Dyslexia

ABSTRACT

Dyslexia is a common neurocognitive learning disorder that can seriously hinder individuals' aspirations if not detected and treated early. Instead of costly diagnostic assessment made by experts, in the near future dyslexia might be identified with ease by automated analysis of eye movements during reading provided by embedded eye tracking technology. However, the diagnostic machine learning methods need to be optimized first. Previous studies with machine learning have been quite successful in identifying dyslexic readers, however, using contrasting groups with large performance differences between diagnosed and good readers. A practical challenge is to identify also individuals with borderline skills. Here, machine learning methods were used to identify individuals with low performance of reading fluency (below 10 percentile from a normal distribution) using their eye movement recordings of reading. Random Forest was used to select most important eye movement features to be used as input to a Support Vector Machine classifier. This hybrid method was capable of reliably identifying dysfluent readers and it also provided insight into the data used. Our best model achieved accuracy of 89.7% with recall of 84.8%. Our results thus establish groundwork for automatic detection of dyslexia in a natural reading situation.

1. Introduction

Dyslexia is a neurocognitive learning disorder characterized by reading and spelling impairments despite normal intelligence [1]. It is one of the most common learning disorders [2] with estimated prevalence of 5% to 12% [3,4]. Dyslexia often has negative consequences on the academic and occupational success [5], the self-esteem [6], and the social-emotional development [7] of an individual. Studies [6,8,9] have shown that the earlier dyslexia is detected and support given in teaching, the more its negative effects can be mitigated [10].

One emerging educational technology is eye tracking, which is a method to measure a user's gaze position on, e.g., a computer screen based on analysis of pupil movement and infrared-induced corneal reflection on eye video. The two main types of eye movements in reading are 200 to 300 ms stay-put moments of gaze, termed fixations [11–13] intervened with rapid ballistic (15 to 80 ms) movements called saccades [12,14].

Readers with dyslexia exhibit quite different eye movement behavior compared to typical readers [14], namely by exhibiting substantially more and longer fixations, shorter saccade duration and length, and more backward directed saccades, i.e. regressions, than typical readers [13–15]. The underlying reason for the eye movement abnormalities is proposed to be due to the difficulties the person has in decoding and recognizing printed words [14,16].

The long-standing promise of eye tracking is to combine it with computational methods to provide fine-grained information of the individual's cognitive processes [14]. An important step to realize this promise is the creation of methods to reliably identify reading difficulties from eye movements. Machine learning methods have been successfully implemented in the detection of dyslexia from eye movements with promising results by [17–19] and [20]. For reviews, see [21–23].

1.1. Related work

Even if the prevalence rate of dyslexia in children is 5%–12% [3,4], many recent studies that apply nonlinear classifiers for the detection of dyslexia are trained with nearly equal proportions. Such direct circumvention of the class-imbalance problem negatively affects the

[☆] Communicated by Franck Patrick Vidal.
[*] Corresponding author.
E-mail address: paavo.j.nieminen@jyu.fi (P. Nieminen).

https://doi.org/10.1016/j.array.2021.100087
Received 4 June 2021; Accepted 24 August 2021
Available online 30 August 2021

performance of learning algorithms, hindering the generalization of the trained models to the population level [24].

For instance, in [17], a controlled trial was performed with 97 participants (aged 10–54), out of which 48 had a diagnosed dyslexia. The assessment setting consisted of 12 readings of 60 words in Spanish by each participant. Using a Support Vector Machine (SVM) model, the authors obtained a 10-fold cross-validation accuracy of 80.2% to identify the dyslexics, using trial-and-error based selection of reading time, mean of fixation time, and age of the participant as features.

In another study [18], altogether 185 Swedish children (aged 9–10) with 97 poor readers (5th percentile in word decoding) and 88 typical readers with an average or above average word decoding skill, were studied. Not surprisingly, due to large differences in the reading skill between the groups and almost equal group sizes, the SVM classifier scored 95.6% 10-fold cross-validation accuracy. A diverse set of eye movement features relating to progressive and regressive saccade lengths and their corresponding fixation durations was used in the analysis. The same dataset was reanalyzed with Particle Swarm Optimization (PSO) based Hybrid Kernel SVM-PSO method in [19], ending up again with 95% classification accuracy. In this study, the eye movement features were transformed into principal components.

In [20], 69 Greek children (aged 8–12), out of which 32 (46 %) were clinically diagnosed as having dyslexia, were studied. The reading skill difference between the groups was substantial as children of the control group were also clinically confirmed not to have reading problems. Eye movements were collected during reading two texts summing up to 324 words. With SVM and the LASSO technique for feature selection, 97% accuracy was obtained with features of saccade length, number of short forward movements, and number of multiply fixated words.

In higher education, [25] reported an ambitious study trying to separate highly skilled university students from low skilled students in their literacy proficiency from eye movements of reading ($n = 61$). Interestingly, these two groups of students did not differ in their overall eye movement parameters such as mean fixation duration or saccade length, but the identification was based on more subtle eye movement patterns related to reading comprehension processes. These features were sentence or paragraph specific forward fixation time, first-pass rereading time, second-pass fixation time, and regression path reading time, leading to the classification accuracy of 80.3% with the SVM method.

Recently, two studies with large test groups of young dyslexics were published. In [26] the assessment of 2679 children (aged 7–9) concluded that fixation duration had the highest correlation with the reading speed and accuracy. An approach not utilizing eye movements was reported in [27], where reading exercises in an online gamified test with 32 linguistic exercises in Spanish for 3644 respondents (aged 7–17) were processed. Dyslexia diagnosis was given to 392 (10.8%) of the participants. Random Forest (RF) classifier with a rich set of 196 features in total, where 4 represented demographic features and 192 performance features from the interaction during playing, scored in 10-fold cross-validation 79.7/79.1% precision and 80.4/78.4% recall in Dyslexia/No dyslexia separation. The analysis of the RF model showed that the two most important features were gender and general performance in Spanish classes.

A fresh review on the eye tracking techniques and applications in [28, Table 3] summarizes SVM as the most commonly used technique, the rising popularity of the convolutional neural networks and deep learning techniques, but identifying only three preliminary studies with the Random Forest technique.

1.2. Our contribution

The promising results of the previous studies with eye tracking and machine learning have largely relied on using small, balanced groups of clinically diagnosed dyslexics with a strictly different control group of non-dyslexics. Our present study extends the machine learning approach to identify reading difficulties based on an arbitrary, albeit rather generally used, cut-off criterion on the reading fluency continuum. While posing a maximal difficulty for the detection, such an identification task is also of utmost practical importance. In most clinical and educational settings, diagnosis is based on a similar generally used arbitrary cut-off score (e.g., below 10 percentile performance in standardized reading tests) in a normally distributed skill [29]. This issue concerns especially reading fluency which is the hallmark of dyslexia in transparent orthographies, where dyslexia is characterized as slow yet rather accurate word reading [30]. Instead, in opaque orthographies dyslexia is characterized by a large number of word reading errors, which is due to the complex correspondence between spoken and written language [31]. Importantly, the cut-off scores based on normal distribution for a reading difficulty are often a major factor in administrative decision-making, i.e., for deciding whether a student is deemed eligible for special education services or not [32].

In Section 2, we describe the dataset and the machine learning methodology we used. In Section 3, we present our computational results which are further discussed in Section 4. Section 5 briefly concludes this paper. The method and results were originally created as part of the Master's Thesis of the first author [33]. For this paper, we have distilled and clarified the major findings of the thesis work.

2. Materials and methods

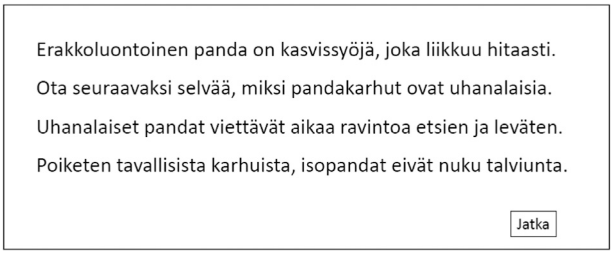
2.1. eSeek Internet reading skill data

The data set used in this research was gathered by the project eSeek from the Department of Psychology at the University of Jyväskylä. The project studied Internet reading skills among Finnish students with and without learning disorders. The data had been obtained over the course of three years from 165 youngsters with an average age of 12.5 years. It includes results of the Internet reading skill tests, eye-movement data, and a partial analysis of these. The students had been chosen from a class of about 400 students. Of the chosen students, 30 (18 %) met the criteria for a reading disorder based on choosing the 10th worst percentile of the reading fluency performance score. This criteria was used to label the students as either dyslexic or typical readers. The eye movements of the participants were recorded using an EyeLink 1000 eye-tracker with a sampling frequency of 1000 Hz. A Dell Precision T5500 workstation with an Asus VG-236 monitor (1920 × 1080, 120 Hz, 52 × 29 cm) at the viewing distance of 60 cm was used for displaying the stimuli. The calibration of the device was performed before the experiment and repeated between trials, if visible head movements were made, a drift was detected on the researcher's screen used for following the eye movements, or the calibration error exceeded 0.30 visual degrees. [34]

During the experiment, participants completed a practice task and then 10 simulated information search tasks. The tasks consisted of reading a contextualized question and then selecting a search result (out of four options) that would help them answer the question. An example of the given question is “Find out why pandas are endangered?” [34]

The eye movement data was obtained from the question page shown to the participants. Fig. 1 shows one question page in which four sentences and a “Continue” button were displayed. The second and the third sentence had an important role; one contained the *task* (question) for this information search, the other was a *distractor* with irrelevant information. The placement of these varied between the tasks, i.e., the task could also be on the third row and the distractor on the second. The *first* and *last* sentences provided some context and narrative for the task assignment. Like the distractor, these sentences contained information that is relevant to the context but not to the task.

In the case of Fig. 1, the distractor is the third sentence. The first sentence reads “The reclusive panda is a herbivore that moves slowly”. The second sentence, which is the task, reads “Next, find out why pandas are endangered”. The distractor sentence translates to “The



Erakkoluontoinen panda on kasvissyöjä, joka liikkuu hitaasti.

Ota seuraavaksi selvää, miksi pandakarhut ovat uhanalaisia.

Uhanalaiset pandat viettävät aikaa ravintoa etsien ja leväten.

Poiketen tavallisista karhuista, isopandat eivät nuku talviunta.

Fig. 1. Example of the question page shown during the task (in Finnish).

endangered pandas spend their time looking for food and resting". Finally, the last sentence reads "Contrary to normal bears, giant pandas do not hibernate".

To avoid confusion with the "task sentence", each of the 10 information search tasks and the practice task that an individual participant completed are henceforth referred to as a *trial*. We also use the shorthand notation T1, T2,…, T11 to refer to each of these 11 trials.

Before use in this study, we removed participants with incomplete reading fluency test results or technically failed trial recordings. After such cleaning, 161 students were left in the data file. Of these, 30 (i.e., 18.6%) were recognized as having a reading disorder based on their reading fluency performance score being in the 10th lowest percentile.

2.2. Machine learning methods

After selecting an arbitrary threshold in reading fluency (10th percentile in this study) and labeling the trials accordingly, our goal becomes that of creating a binary classifier using supervised machine learning. While a massive number of alternatives exist, we found good reasons to use the ones discussed next.

Support Vector Machine (SVM) [35] was a very clear first choice for us because it has been used in all prior works on dyslexia detection based on eye-tracking that we found and overviewed in Section 1.1. Thus, using SVM as a baseline method puts us in a comparable zone with other systems under development. SVM maps its input vectors into a high dimensional feature space through a chosen non-linear mapping and then finds an optimal hyperplane to separate the classes with a maximal margin which reduces generalization error [36]. In this study, we used the SVM implementation in the Python module Scikit-learn [37] which applies the LIBSVM library [38] internally.

Random Forest (RF) [39] is a classifier comprising an ensemble of randomized decision trees, which make a joint decision on the class. Like SVM, RF has been used with good results in various tasks [40]. Our reasons to use RF were plenty. First of all, we wanted to compare SVM with at least one other popular method. The timeframe of our project would not allow a fully comprehensive comparison of every method available, so we opted to pick one with greatest promises. In addition to its demonstrated practical usefulness, we were intrigued by the reported robustness, generalization capability, and intrinsic feature selection opportunities embedded in decision tree-based methods [41, 42]. In this study, we picked also the RF implementation from the Scikit-learn Python module [37] as we did with SVM.

2.3. Hyperparameter optimization

Appropriate configuration of hyperparameters is necessary in order to produce a model with the best performance for the problem. In order to strike a compromise between adequate performance and time spent in exploring the hyperparameter space, we opted to use grid search for two influential parameters in both methods and leave other hyperparameters to the default settings in Scikit-learn. In such a low-dimensional setting, grid search is presumed to be superior to manual search both in efficiency and reliability [43].

For this research, we chose to use the default kernel type of Scikit-learn which is the radial basis function. Exhaustive search of the most suitable kernel type was deliberately left out of our current scope. For the two hyperparameters, `C` (regularization cost) and `gamma` (radial basis width), a grid search was performed.

For RF, we chose two parameters that we regarded most influential. The number of trees in the forest is defined by the `n_estimators` parameter. Having a larger number of trees is usually better, but that also increases the computation time for the model [41]. When splitting a node in the decision tree, the feature used for the split is selected from a random subset of features. The number of features chosen into this subset is determined by the `max_features` parameter [41]. Other parameters of the RF were left to the default settings of Scikit-learn.

2.4. Cross-validation

While 10-fold cross-validation is usual [44], other numbers of folds can be used according to use scenario [45]. We chose to use 5 folds in order to conserve computational time and to increase the number of underrepresented dyslectic cases in each fold (6 instead of 3). The latter effect enabled more intelligible probing of fold-wise results which was useful at the early method development stage reported here. We used stratified subsampling in creating the folds.

2.5. Training and evaluation algorithm

Algorithm 1 was used in simultaneously training and evaluating the models and searching for optimal hyperparameters. We implemented it using Python and Scikit-learn and additionally the Pandas module for input data handling. The source code is available at https://r.jyu.fi/DHV.

The training and evaluation are done in cycles; each cycle consists of training the model and obtaining the results. The `cycles` constant determines how many times the whole cross-validation cycle is done. `p1` and `p2` are two hyperparameters chosen to be optimized (`C` and `gamma` for SVM; `n_estimators` and `max_features` for RF).

Algorithm 1 Training and evaluation with cross-validation.

```
for i = 1, ... , cycles do
    for p1, p2 in hyperParameters do
        Create five cross-validation folds
        for each cross-validation fold do
            Create classifier
            Fit model with data
            Store resulting predictions
            Calculate and store confusion matrix
        end for
    end for
end for
Sort resulting models according to the recall score
```

The combinations of hyperparameters are compared against each other by a performance metric. In our case, we used the recall score of dyslexics predicted. Recall is the fraction of correctly predicted samples out of all the samples of the positive class. This was chosen as the performance metric in this research as it was deemed more important to correctly detect the dyslexics than typical readers. In addition to the recall score, we also observed the overall accuracy of the model. Using only the accuracy score is not enough, because the classes are unbalanced in our data. It would be possible to obtain an accuracy of 81.4% by just declaring all of the test subjects as typical readers. This would give a false picture of the model's performance.

In the case of RF, the algorithm also calculates the feature importances for each model created in the cross-validation folds. We shall return to the topic of feature selection using RF in Section 2.7.

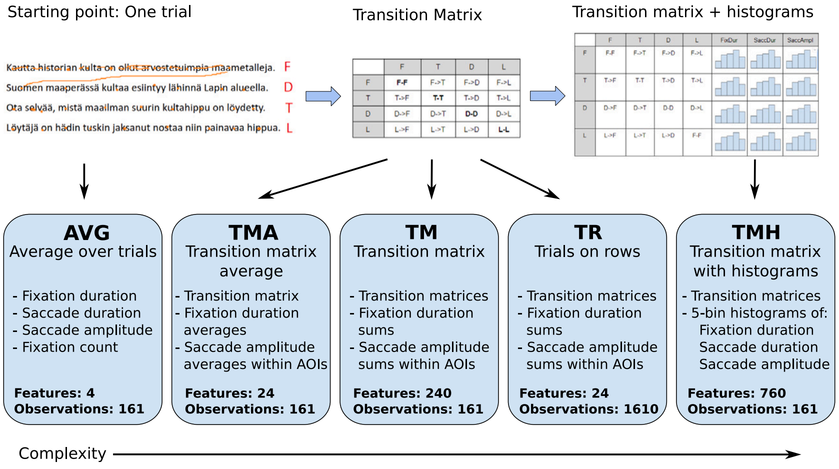

Starting point: One trial
Kautta historian kulta on ollut arvostetuimpia maametalleja. F
Suomen maaperässä kultaa esiintyy lähinnä Lapin alueella. D
Ota selvää, mistä maailman suurin kultahippu on löydetty. T
Löytäjä on hädin tuskin jaksanut nostaa niin painavaa hippua. L
Transition Matrix
F T D L
F-F F->T F->D F->L
T->F T-T T->D T->L
D->F D->T D-D D->L
L->F L->T L->D L-L
Transition matrix + histograms
FixDur SaccDur SaccAmpl
AVG
Average over trials
- Fixation duration
- Saccade duration
- Saccade amplitude
- Fixation count
Features: 4
Observations: 161
TMA
Transition matrix average
- Transition matrix
- Fixation duration averages
- Saccade amplitude averages within AOIs
Features: 24
Observations: 161
TM
Transition matrix
- Transition matrices
- Fixation duration sums
- Saccade amplitude sums within AOIs
Features: 240
Observations: 161
TR
Trials on rows
- Transition matrices
- Fixation duration sums
- Saccade amplitude sums within AOIs
Features: 24
Observations: 1610
TMH
Transition matrix with histograms
- Transition matrices
- 5-bin histograms of: Fixation duration Saccade duration Saccade amplitude
Features: 760
Observations: 161
Complexity

Fig. 2. Features in increasing level of complexity.

	F	T	D	L
F	11	1	2	3
T	2	16	5	6
D	1	7	8	0
L	0	4	8	4

Fig. 3. Example of a transition matrix used in this study.

2.6. Feature extraction

A major part of this study was the exploration of possible features, i.e., classifier inputs, that enable discrimination in our binary classification task. Fig. 2 outlines our procedure of creating feature sets with increasing complexity, based on earlier studies (Section 1.1) and also some newly devised ideas.

The left side of Fig. 2 introduces the most obvious and simplistic feature set, labeled AVG in result analysis later on. The AVG set contains simple averages and sums of fixations and saccades observed in the eye-tracking recordings. There are 4 features for each 161 participants. The saccade amplitude and saccade duration are partially tied to each other, as the larger the amplitude, the longer the saccade lasts.

In the center of Fig. 2, we have first computed a transition matrix [11] representing the transitions of the participant's gaze from one area of interest (AOI), i.e., sentence, to another. Contrary to tradition, we include on the diagonal the times that the gaze has shifted within one AOI. This was considered relevant because dyslexics have been known to have more fixations while reading. The abbreviations F, T, D, and L denote the first, task, distractor, and last sentence, respectively. Fig. 3 shows an example of a transition matrix. We can see that the eyes of the participant have moved within the first sentence 11 times, moved from the first sentence to the task 1 time, first sentence to distractor two times, and so on.

We conjectured that the transitions between sentences could be useful in separating the readers with difficulties from typical readers. The reasoning is that dyslexics would have more difficulty finding the task sentence out of the four than typical readers. This would cause them to have a more erratic gaze movement among the sentences, and by comparing transition matrices it should be possible to notice this difference.

Our feature set labeled TMA contains averages of the transition matrix values (16 features) and fixation duration averages and saccadic amplitude averages within each area of interest (AOI), i.e., each sentence, (+8 features) averaged over all the trials. The intention behind averaging was to lower the dimensionality of the feature set and possibly reduce the noise.

The feature set labeled TM contains these $16 + 8 = 24$ features separately for each trial, thus the dataset size is 240 features for each 161 participants.

For the feature set TR, we also rephrased and relabeled the classification task to identify not participants but individual trials, so we have 24 features for each 1610 individual trials. For evaluation, a participant's result is determined by a vote: Each of the 10 separate trials of a participant are classified, and the final prediction is dyslexic when more than five (i.e., half of the trials) say so.

The rightmost feature set of Fig. 2, labeled TMH, includes not averages but 5-bin histograms of 3 values: fixation duration, saccade duration, and saccade amplitude. The histograms were created separately for each 4 sentences in each 10 trials. The 160 transition matrix values of TM were also included in TMH. This causes the feature set to have $5 \times 3 \times 4 \times 10 + 160 = 760$ features for each 161 participants. The bin intervals were calculated beforehand by evenly dividing the entire range of data values into five equally sized partitions. Each feature is min–max scaled to range $[0, 1]$ before use.

2.7. Feature selection

In addition to the feature sets described in Section 2.6, we created some reduced ones via feature selection [46]. The basic idea is to use the RF classifier to select the most important features, which are then given to SVM for classifying the data in a manner similar to [47]: We calculate the feature importances with RF for every fold, every cycle, and every hyperparameter combination. Of these feature importances, the 10 most important ones are saved at each fold rotation. Once the hyperparameter combination with the best recall value for the dyslexia class has been found, the n most frequent features are picked into their own feature set, where n is the number of features chosen. We tried out $n = 10, 20, 30, 35,$ and 40 to heuristically search for the optimal feature set. In result analysis, we use the labels RFF10, RFF20, RFF30, RFF35, and RFF40, respectively.

3. Results

We applied the method of Section 2.5 with 100 cycles and a 5-fold cross-validation for various classifiers and feature sets. Table 1 shows an overview of the best results. The "Method" column indicates the machine learning method used to produce the model. The "Bal" tag indicates that the class weights were balanced for the Scikit-learn library SVM by adjusting them inversely in proportion to class frequencies. The "Feat" column holds the names of the feature sets as given in Sections 2.6 and 2.7. The "Accuracy" column holds the average fraction of correct predictions for all of the 100 models created in the

Table 1
Best models created with their accuracy and recall scores.

Method	Feat.	Accuracy	Recall
SVM	RFF35	89.8% ± 4.7%	75.9% ± 17.1%
	TR	86.4% ± 1.8%	55.7% ± 6.4%
SVM Bal	RFF35	89.7% ± 4.0%	84.8% ± 14.0%
RF	RFF35	86.9% ± 4.6%	54.0% ± 20.4%

Table 2
Results for SVM using features selected by Random Forest.

Feat.	Accuracy	Recall	C	γ
RFF10	85.7% ± 5.7%	57.5% ± 20.6%	8000	0.05
RFF20	86.5% ± 5.0%	61.4% ± 20.6%	30	1.0
RFF30	89.9% ± 4.6%	73.8% ± 17.5%	30	1.1
RFF35	89.8% ± 4.7%	75.9% ± 17.1%	30	1.09
RFF40	89.5% ± 4.7%	74.5% ± 17.1%	30	0.9
RFF35bal	89.7% ± 4.0%	84.8% ± 14.0%	1	1

Table 3
Results for SVM using feature sets apart from RFFn.

Feat.	Accuracy	Recall	C	γ
AVG	85.0% ± 2.1%	42.8% ± 18.1%	100000	0.05
TMA	80.9% ± 2.8%	46.5% ± 19.6%	500	0.09
TM	78.2% ± 3.9%	38.5% ± 19.3%	1000	0.001
TMH	85.0% ± 3.1%	41.6% ± 19.5%	200	0.009
TR	86.4% ± 1.8%	55.7% ± 6.4%	50000	0.1

Table 4
Results for RF using the generated feature sets.

Feat.	Accuracy	Recall	maxF	#Est
AVG	80.7% ± 5.5%	50.2% ± 19.2%	4	10
TMA	83.6% ± 5.0%	41.3% ± 19.2%	18	30
TM	81.7% ± 5.3%	36.9% ± 20.1%	240	20
TMH	84.5% ± 4.6%	39.9% ± 19.2%	550	20
TR	86.7% ± 1.1%	36.3% ± 5.1%	24	20
RFF35	85.4% ± 1.1%	42.6% ± 19.7%	5	20

algorithm cycles. The error given is the standard deviation of these accuracy scores. Similarly, the “Recall” column contains the average recall scores with standard deviation.

The best results for SVM were achieved by the RFFn feature sets. These results are displayed in Table 2. The first column holds the name of the feature set. The two last columns contain the SVM hyperparameter values that yielded the best result. By balancing the class weights, the recall score of the RFF35 model was improved significantly with a minuscule decrease in accuracy. Table 3 displays the results obtained with SVM by using the rest of the feature sets. The best accuracy and recall scores were obtained with the TR feature set.

The best results obtained by the Random Forest classifier are displayed in Table 4. The two last columns contain the hyperparameters optimized with grid-search and used by each model. As can be seen, the results are not as good as with SVM. Yet, using RF as a feature selection mechanism enabled us to find the best predictions using SVM.

4. Discussion

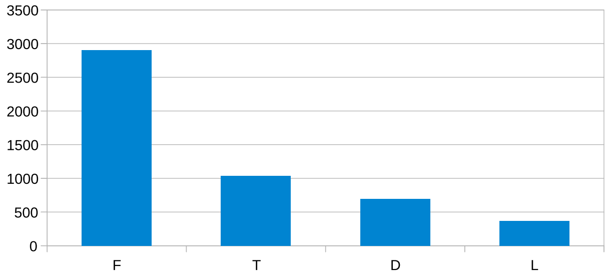


Fig. 4. Importance of each sentence.

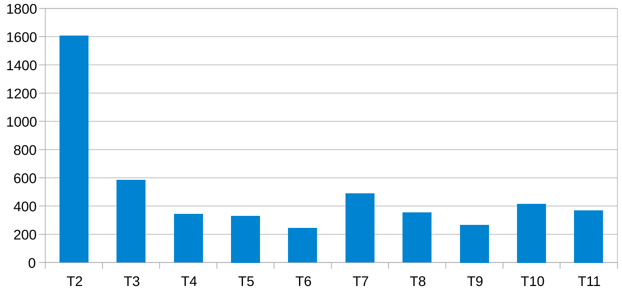


Fig. 5. Importance of each trial.

Analyzing the most relevant features obtained in the RFFn sets enables us to make observations that may be of interest in designing future tests for dyslexia screening. Fig. 4 displays the number of features related to each sentence chosen for the top 10 most important features each fold rotation. The total number of features chosen is $10 \times folds \times cycles = 10 \times 5 \times 100 = 5000$. Of these, 2900 (58%) were related to the first sentence on the task question page. This relation means that the feature was generated from gaze activity in the first sentence. The other sentences had a much lesser effect. This finding may be related to the usage of contextual information in reading. Knowing the context helps readers read faster as they are able to predict upcoming words [48]. In first sentence, the context is not yet established and therefore readers have to rely mostly on their word recognition, decoding and syntactic language skills, which are known to be primarily affected in dyslexia. In other words, the finding suggests that dyslectic readers may rely relatively more on context in their reading as a compensatory reading strategy [49].

The different trials T2–T11 seemed to be another factor in feature importance. Fig. 5 presents the number of features related to each trial chosen for the top 10 each fold rotation. The features generated from T2 data (i.e., first actual trial) occur most often (32%), indicating a high significance in classifying the two classes correctly. For the rest of the trials, the feature count stays somewhat in the same range, with low points at T6 and T9. The high importance of T2 is speculated to be the result of the participants not having yet established a context and cognitive schema for the information searching task. Again, dyslexics suffer more from not knowing the context of the text than fluent readers. This could explain the importance of T2 trial features in separating the two classes; at this point the participants had not yet seen enough trials to establish the context and form of the question page text. Later on in the experiment, the context is established and thus it is harder to distinguish the readers with difficulties from typical readers.

Fig. 6 shows the number of times the most important features were picked. We can see that features concerning the first sentence (indicated by an “F”) and ones from T2 have occurred most often, indicating their importance, as stated above. In addition, by looking at the histogram bin numbers in the feature names, we can also see that in the case of saccadic features, the most frequent bin is the first. Respectively, for the features that are created from fixation data, the most important bin is the last. These observations indicate that the shortest saccades and the longest fixations help the classification the most. This is a conclusion that agrees with the results obtained by [13,15].

We can also notice that features extracted from saccadic data are more important than ones from fixation data, although also a reverse pattern has been reported [26]. The use of transition matrices did not contribute greatly to the classification; only one feature in the 35 most important features is from a traditional transition matrix. The other three features in this list (T2F-F, T10T-T and T3D-D) are the numbers of fixations made within the indicated sentence.

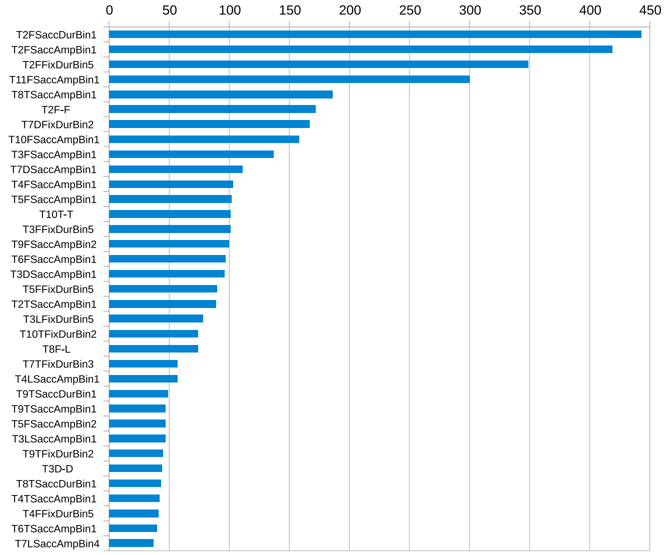


Fig. 6. Importance of fine-grained features.

5. Conclusions

In this study, we developed a classifier to identify dyslexic readers from eye movement data. Importantly, we defined dyslexia here by an arbitrary threshold in reading fluency score which is a realistic and practically important choice, leading to an inherently difficult classification task. Our feature extraction augments the traditionally used transition matrices by using gaze patterns within AOIs and histograms rather than plain average values of fixation and saccade measures. An SVM classifier using most relevant eye movement features selected using RF met an accuracy of 89.7% and a recall score of 84.8%. The result is promising, and deeper analysis of the feature importances provides insight that can be used in guiding future research towards fast and reliable dyslexia screening tools.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work.

Acknowledgments

This research was supported by the Academy of Finland, grants #274022, #311877, and #317030.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.array.2021.100087.

References

[1] Frazier M, editor. Dyslexia : Perspectives, Challenges and Treatment Options. Nova Science Publishers, Inc.; 2016.

[2] Handler SM, Fierson WM, et al. Joint technical report—Learning disabilities, dyslexia, and vision. Pediatrics 2011;peds–2010.

[3] Shaywitz SE. Dyslexia. N Engl J Med 1998;338(5):307–12.

[4] Katusic SK, Colligan RC, Barbaresi WJ, Schaid DJ, Jacobsen SJ. Incidence of reading disability in a population-based birth cohort, 1976–1982, Rochester, Minn. In: Mayo Clinic Proceedings, vol. 76, Elsevier; 2001, p. 1081–92.

[5] Morris D, Turnbull P. A survey-based exploration of the impact of dyslexia on career progression of UK registered nurses. J Nurs Manag 2007;15(1):97–106.

[6] Glazzard J. The impact of dyslexia on pupils' self-esteem. Support Learn. 2010;25(2):63–9.

[7] 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.

[8] Snowling MJ, Hulme C. Interventions for children's language and literacy difficulties. Int J Lang Commun Disord 2012;47(1):27–34.

[9] Torgesen JK. Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. Learn. Disabil. Res. Pract. 2000;15(1):55–64.

[10] Vellutino FR, Fletcher JM, Snowling MJ, Scanlon DM. Specific reading disability (dyslexia): What have we learned in the past four decades? J. Child Psychol Psychiatry 2004;45(1):2–40.

[11] Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J. Eye Tracking: A Comprehensive Guide to Methods and Measures. OUP Oxford; 2011.

[12] Eden G, Stein J, Wood H, Wood F. Differences in eye movements and reading problems in dyslexic and normal children. Vis Res 1994;34(10):1345–58.

[13] Deans P, O'Laughlin L, Brubaker B, Gay N, Krug D. Use of eye movement tracking in the differential diagnosis of attention deficit hyperactivity disorder (ADHD) and reading disability. Psychology 2010;1(04):238.

[14] Rayner K. Eye movements in reading and information processing: 20 years of research. Psychol Bull 1998;124(3):372.

[15] De Luca M, Borrelli M, Judica A, Spinelli D, Zoccolotti P. Reading words and pseudowords: An eye movement study of developmental dyslexia. Brain Lang. 2002;80(3):617–26.

[16] Hyönä J, Olson RK. Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. J Exp Psychol: Learn Mem Cogn 1995;21(6):1430.

[17] Rello L, Ballesteros M. Detecting readers with dyslexia using machine learning with eye tracking measures. In: Proceedings of the 12th Web for All Conference. W4A '15, ACM; 2015, 16:1–16:8.

[18] Nilsson Benfatto M, Öqvist Seimyr G, Ygge J, Pansell T, Rydberg A, Jacobson C. Screening for dyslexia using eye tracking during reading. PLoS One 2016;11(12):e0165508.

[19] Jothi Prabha A, Bhargavi R. Prediction of dyslexia from eye movements using machine learning. IETE J. Res. 2019;1–10.

[20] Asvestopoulou T, Manousaki V, Psistakis A, Smyrnakis I, Andreadakis V, Aslanides IM, Papadopouli M. DysLexML: Screening tool for dyslexia using machine learning. 2019, arXiv preprint arXiv:1903.06274.

[21] Perera H, Shiratuddin MF, Wong KW. Review of the role of modern computational technologies in the detection of dyslexia. In: Information Science and Applications (ICISA) 2016. Springer; 2016, p. 1465–75.

[22] Gündüz A, Najjar T. Analysis eye movements during reading by machine learning algorithms: A review paper. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). 2018, p. 1069–75.

[23] Prabha AJ, Bhargavi R. Prediction of dyslexia using machine learning—A research travelogue. In: Proceedings of the Third International Conference on Microelectronics, Computing and Communication Systems. Springer; 2019, p. 23–34.

[24] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21(9):1263–84.

[25] Lou Y, Liu Y, Kaakinen JK, Li X. Using support vector machines to identify literacy skills: Evidence from eye movements. Behav Res Methods 2017;49(3):887–95.

[26] Strandberg A. Eye movements during reading and reading assessment in swedish school children: a new window on reading difficulties. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications. 2019, p. 1–3.

[27] Rello L, Baeza-Yates R, Ali A, Bigham JP, Serra M. Predicting risk of dyslexia with an online gamified test. Plos One 2020;15(12):e0241687.

[28] Klaib AF, Alsrehin NO, Melhem WY, Bashtawi HO, Magableh AA. Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and internet of things technologies. Expert Syst Appl 2021;166:114037, (in press).

[29] de Jong PF, van Bergen E. Issues in diagnosing dyslexia. In: Segers E, van den Broek P, editors. Developmental Perspectives in Written Language and Literacy. 2017, p. 349–61.

[30] Wimmer H, Schurz M. Dyslexia in regular orthographies: manifestation and causation. Dyslexia 2010;16(4):283–99.

[31] Seymour PH, Aro M, Erskine JM, collaboration with COST Action A8 Network. Foundation literacy acquisition in European orthographies. Br J Psychol 2003;94(2):143–74.

[32] Weber MC. The IDEA eligibility mess. Buff. L. Rev. 2009;57:83.

[33] Raatikainen P. Automatic detection of developmental dyslexia from eye movement data. University of Jyväskylä; 2019.

[34] Hautala J, Kiili C, Kammerer Y, Loberg O, Hokkanen S, Leppänen PH. Sixth graders' evaluation strategies when reading internet search results: an eye-tracking study. Behav. Inform. Technol. 2018;1–13.

[35] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

[36] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw 1999;10(5):988–99.

[37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12(Oct):2825–30.

[38] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2(3):27.

[39] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[40] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests. Pattern Recognit 2011;44(2):330–49.

[41] Louppe G. Understanding random forests: From theory to practice. 2014, arXiv preprint arXiv:1407.7502.

[42] Cutler A, Cutler DR, Stevens JR. Random forests. In: Ensemble Machine Learning. Springer; 2012, p. 157–75.

[43] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(Feb):281–305.

[44] Kohavi R, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. 14, Montreal, Canada; 1995, p. 1137–45.

[45] Marcot BG, Hanea AM. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Comput Statist 2020.

[46] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.

[47] Yang J, Yao D, Zhan X, Zhan X. Predicting disease risks using feature selection based on random forest and support vector machine. In: International Symposium on Bioinformatics Research and Applications. Springer; 2014, p. 1–11.

[48] Hawelka S, Schuster S, Gagl B, Hutzler F. On forward inferences of fast and slow readers. An eye movement study. Sci Rep 2015;5.

[49] Nation K, Snowling MJ. Individual differences in contextual facilitation: Evidence from dyslexia and poor reading comprehension. Child Develop. 1998;69(4):996–1011.