ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Screening Dyslexia for English: Impact of Heterogeneity in Demographic Variables

Enrique Romero * and Luz Rello

Abstract. Dyslexia is a complex learning disorder that can be challenging to diagnose. In this way, it is crucial to gather knowledge about the impact of key attributes. This work focuses on demographic variables used in a pioneering computer-based linguistic game designed for the screening of dyslexia using Machine Learning. The analysis reveals the heterogeneity in these variables, offering valuable insights for future Machine Learning approaches. It emphasizes key contributions, such as strategies to mitigate biases and effectively address heterogeneity, suggesting the formation of subgroups based on interaction data collected.

1 Introduction

The application of Machine Learning to detect dyslexia has experienced remarkable growth over the past decade, particularly with methods leveraging diverse types of data [1]. Early methods used eye-tracking data and data derived from brain imaging including MRI and EEG. More recent approaches favor cost-effective, accessible data, such as metrics from computer-based linguistic games.

One of the pioneering studies to use computer-derived measures via linguistic games is described in [2], which utilized Machine Learning to identify dyslexia risk. Following the implementation and application of this model in Spanish-speaking populations [3], it was observed that conducting an initial analysis of key attributes could have optimized the process and facilitated the development of more reliable and less biased Machine Learning models for dyslexia detection.

This study presents an attribute analysis applied to English, focusing on key attributes relevant to dyslexia detection including age, gender, native language, and whether the individual has failed language-related assessments. The results offer valuable insights and specific contributions to the implementation of predictive models for dyslexia detection. For example, the analysis can inform decisions on the number of models required or assess the risk of potential biases, thereby improving the reliability and effectiveness of future models.

2 Data Description

The data set used in this work contains 267 examples, corresponding to demographic data and measures of the participants in a study of dyslexia prediction

*The PERMEPSY project is supported under the frame of ERA PerMed by: Instituto de Salud Carlos III (ISCIII), Spain; German Federal Ministry of Education and Research (BMBF), Germany; Agence Nationale de la Recherche (ANR), France; National Centre for Research and Development (NCBR), Poland; Agencia Nacional de Investigación y Desarrollo (ANID), Chile. This paper is part of project PID2022-143299OB-I00, financed by MCIN/AEI/10.13030/501100011033/FEDER, UE.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

with *Dytective for English*, an online game designed to that end [2]. Participants in the study ranged from 7 to 60 years old. There are four demographic features: *Gender*, *Age*, *Second Language* (with values 'yes' or 'no' depending on whether the participant has a second language other than English or not) and *Language Subject* (with value 'yes' if the participant declares that failed language-related assessments and 'no' otherwise). The rest of the features were the obtained measures of 37 attention and linguistic exercises addressing different indicators related to dyslexia. For each exercise, six features were collected: Clicks, Hits, Misses, Score, Accuracy and Missrate, leading to a total number of 226 features (including the four demographic attributes). Each participant was marked as D if the participant has dyslexia, N if not, and M (maybe) if the participant suspects that he or she has dyslexia but is not diagnosed. A detailed description of the data can be found in [2].

3 Data Preprocessing and Feature Extraction

A number of preprocessing steps were performed. Examples marked as M (maybe) were assigned to the class D (dyslexia). Participants over 15 years of age were joined together. Categorical variables with only two possible values (*Gender, Second Language* and *Language Subject*) were codified as $\{-1, +1\}$. Missing values were set to 0.

The number of features was very large compared to the number of examples. Additionally, the variables associated to the 37 exercises contain redundant information. In this way, a feature extraction process was performed, trying to summarize the information contained in each of the 37 exercises in two new generated features.

1. Efficiency, computed as $(\text{Hits} - \text{Misses})/(\text{Hits} + \text{Misses})$. By construction, the values of this attribute are in the interval $[-1, +1]$. The sign of this attribute indicates whether the number of hits is larger than the number of misses or not.

2. Speed, computed as $\min(\text{Hits} + \text{Misses}, 30)$. Since every exercise had a limited time to be completed, the total number of responses is an indicator of speed. The maximum value is limited to 30, which is already a very high value.

Therefore, the data set analyzed in this work contains 267 examples and 78 features. Although the number of features is still large compared to the number of examples, the information is much more compact than in the original data.

4 Data Analysis

A descriptive analysis of demographic variables is conducted with respect to the two extracted features: Efficiency and Speed. This analysis aims to detect heterogeneity among different categories and provide deeper insights into the

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

data. Figure 1 shows the mean Efficiency and Speed (across the 37 exercises) for *Gender, Second Language, Language Subject* and *Age* (from top to bottom and left to right). Yellow and blue bars correspond to dyslexic and non-dyslexic samples, respectively. Note that two different categories (such as Male/Female for *Gender*, etc.) have similar behavior if they have similar global shapes (*i.e.* the values can be different between dyslexics or non-dyslexics but will be similar between the dyslexics of all categories and the non-dyslexics for all categories.

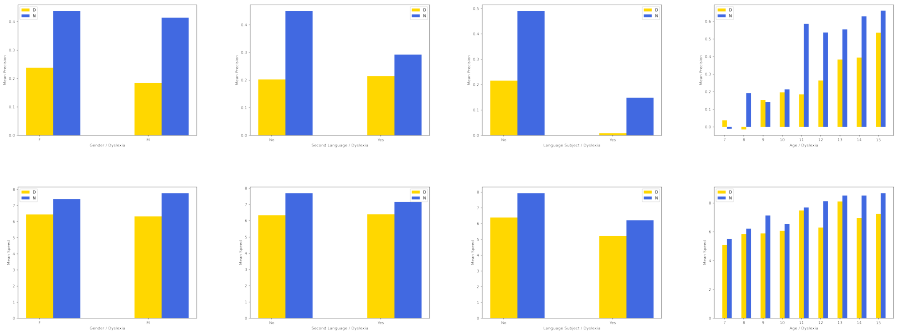

Fig. 1: Mean Efficiency (top row) and Speed (bottom row) for *Gender* (first column), *Second Language* (second column), *Language Subject* (third column) and *Age* (fourth column). See the text for details.

A number of conclusions can be extracted from this analysis. In general, it can be seen that dyslexic participants have less efficiency and speed than non-dyslexic ones, as expected. In addition, no relevant differences are observed when between genders (see Figure 1, first row). However, the rest of demographic variables show important and interesting differences that are an indicator of the heterogeneity in the data. Remarkably, these differences are more noticeable in efficiency rather than in speed. Regarding *Second Language*, efficiency is higher for non-dyslexics which do not have a second language than for non-dyslexics bilingual, while dyslexics have similar efficiency regardless of whether they have a second language or not. This can be explained because the difficulties of dyslexia transfer to the second language if phonological awareness in the first language has not been well established.

With respect to *Language Subject*, it is clear that non-dyslexics without troubles in language classes at school have the highest efficiency. In contrast, dyslexics with school problems have low efficiency values. Interestingly, dyslexics without troubles in language classes perform better, both in efficiency and speed, than non-dyslexics with difficulties. This is a clear indicator of the significant challenge in detecting dyslexia, especially when individuals have received treatment. It also demonstrates how the symptoms of dyslexia can be overcome once identified and addressed.

Age analysis also shows very interesting information. First, it can be observed that speed increases with age, something that is somewhat expected. However, note that it happens for both non-dyslexics and dyslexics and the differences

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

do not seem to be too large, being also an indicator of the difficulty of the problem. Second, efficiency has clear differences depending on the ranges of ages considered. At age 7, efficiency is very low, regardless of whether one is dyslexic or not. At age 8, efficiency is much larger for non-dyslexics than for dyslexics. At ages 9 and 10 efficiencies are very similar, but at the age of 11 there is an abrupt change in the behavior which leads to a very large difference between dyslexics and non-dyslexics. Finally, it can be observed that these differences tend to decrease as long as age increases. Again, this is a clear indicator of heterogeneity in the data which may lead to wrong conclusions if it is not taken into account. These findings are consistent with prior scientific studies indicating that language acquisition occurs in developmental stages.

In summary, this analysis reveals the existence of a high degree of heterogeneity in the data with respect to demographic variables, notably *Second Language, Language Subject* and *Age*, suggesting that different models should be constructed in order to accommodate these differences in a reasonable manner. The analysis also shows that, in general, efficiency is more important than speed.

5 Experiments

The most interesting conclusions of the analysis performed in the previous sections, from the point of view of dyslexia, are those related to age since *Age* is a determining factor in defining the stages of language acquisition [4].

In this way, a series of experiments were designed to evaluate the impact of the heterogeneity observed in this variable. To that end, and motivated by the difference in their behavior observed in Figure 1, data was split into three groups of age: AgeGroup1 (7-8 years old), AgeGroup2 (9-10 years old) and AgeGroup3 (over 11 years old). The main idea of the experiment is to learn with the data of one of the groups and test with the other ones. Large differences among the respective results would indicate that the samples come from different populations, which should be taken into account when constructing general screening dyslexia systems. For the sake of comparison, the same experiment was performed for Gender (GenderM and GenderF subsets), which did not show such heterogeneity, and for the whole data set (Complete).

The statistics related to the number of examples and percentages of dyslexic participants in each group are: Complete (267, 26.59%), AgeGroup1 (57, 24.56%), AgeGroup2 (45, 46.67%), AgeGroup3 (165, 15.78%), GenderMale (141, 20.57%) and GenderFeMale (126, 25.40%).

A series of six Machine Learning experiments were conducted using the described data subsets. In each experiment, one data set was used for training and another for testing, except for the Complete dataset, as explained below. The procedure can be easily illustrated with an example. For the AgeGroup1 data subset, a model selection process was performed to select a set of good parameters. This model selection process was conducted with a Cross-Validation (CV) scheme, giving the results for the AgeGroup1 as the mean values in the validation sets (a double CV scheme was not feasible due to limited data). The

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

final selected model was tested on the other age subsets: AgeGroup2 and AgeGroup3. This procedure was repeated for the rest of age and gender groups. For the Complete subset, since there was no test set available, the results were those of the validation sets in the CV model selection process.

Support Vector Machine classifiers [5] (SVMCs) were used as Machine Learning models. SVMCs with linear and Gaussian kernels were constructed for every training data group. Model selection was performed with a grid search over the γ and C parameters of the SVMC. 30 repetitions of a 5-fold CV were performed for every experiment in order to collect enough statistics of the results. The features included in each subset were: *Second Language, Language Subject* and Efficiency. Therefore, all subsets have the same 39 attributes. Since the classes have a high degree of unbalance, we considered that the most suitable measure for assessing the experiments was the balanced accuracy. Due to space limitations, only the results of SVMCs with linear kernels are shown, but similar performances are obtained with Gaussian kernels. The mean balanced accuracy for the Complete data set was 83.81%. The results for the rest of subsets can be seen in Tables 1 and 2 for the Age and the Gender subsets, respectively. Each row corresponds to an experiment with the corresponding training set and each column to the validation or test results (see above).

Training	AgeGroup1	AgeGroup2	AgeGroup3
AgeGroup1	84.64%	66.00%	74.04%
AgeGroup2	85.37%	70.08%	83.91%
AgeGroup3	63.24%	59.95%	87.85%

Table 1: Mean balanced accuracy for Age subsets (see the text for details).

Training	GenderM	GenderF
GenderM	85.01%	81.00%
GenderF	78.64%	84.25%

Table 2: Mean balanced accuracy for Gender subsets (see the text for details).

Several conclusions can be drawn from the experiments. A large difference in the results of the age subsets can be observed. Models trained with AgeGroup1 have very different performance when tested with AgeGroup2 and AgeGroup3, not only with respect to AgeGroup1 but between AgeGroup2 and AgeGroup3. Models trained with AgeGroup3, in contrast, have similar (and poor) performance when tested with AgeGroup1 and AgeGroup2. Somewhat surprisingly, models trained with AgeGroup2 have a very good behavior in AgeGroup1 and AgeGroup3. This clearly indicates that these three groups are very different among them, being AgeGroup2 the most difficult to predict. The scarcity of the data could also be a possible explanation for these differences, but then one would also have the same picture for the gender subsets, which is not the case. In this way, gender subsets seem to be much more similar between them than age subsets. Finally, note that global results are comparable to those in [2], indicating that the efficiency feature contains relevant information for the task.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

6 Conclusions

In this study, we analyzed and characterized groups within two populations - individuals with and without dyslexia. The findings offer recommendations for optimizing Machine Learning models.

Absence of Gender Bias. The study found no significant gender bias in dyslexia-related tasks like linguistic games, despite higher diagnosis rates in males [6]. This suggests gender balancing in datasets is unnecessary, especially in contexts with limited data availability, streamlining study design.

Age-Based Differentiation. Differences between dyslexic and non-dyslexic individuals were most evident in specific age groups, particularly between ages 9 and 11, and became more pronounced by age 11. These findings emphasize the importance of age as a factor in dyslexia research [4]. From a machine learning perspective, this age-based differentiation highlights the necessity of designing predictive models tailored to specific age groups. A single, generalized model trained on a heterogeneous dataset may fail to capture these nuances, especially with limited data volumes. Instead, stratifying data by age could improve model accuracy and reliability, as shown in the experiments.

To sum up, the study highlights the critical role of thorough preliminary data analysis in addressing population heterogeneity and mitigating biases. These insights lead to recommendations for tailoring subsequent Machine Learning methods, thereby enhancing their reliability and accuracy. Such advancements offer significant benefits to both the scientific and educational communities.

References

[1] S. Kaisar. Developmental dyslexia detection using machine learning techniques: A survey. *ICT Express*, 6(3):181–184, 2020.

[2] L. Rello, E. Romero, M. Rauschenberger, A. Ali, K. Williams, J.P. Bigham, and N. Cushen-White. Screening Dyslexia for English Using HCI Measures and Machine Learning. In *International Digital Health Conference*, pages 80–84, 2018.

[3] L. Rello, Baeza-Yates R., Ali A., Bigham J. P., and Serra M. Predicting risk of dyslexia with an online gamified test. *PLoS ONE*, 15(12):e0241687, 2020.

[4] S. E. Shaywitz and B. A. Shaywitz. Dyslexia (specific reading disability). *Biological psychiatry*, 57(11):1301–1309, 2005.

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, UK, 2000.

[6] A. B. Arnett, B. F. Pennington, R. L. Peterson, E. G. Willcutt, J. C. DeFries, and R. K. Olson. Explaining the sex difference in dyslexia. *Journal of Child Psychology and Psychiatry*, 58(6):719–727, 2017.