7000CEM INDIVIDUAL RESEARCH PROJECT PREPARATION

DETECTING DYSLEXIA USING MACHINE LEARNING TECHNIQUES PROJECT PROPOSAL

11796778

SULEIMAN ADAMU YAKUBU

1. Problem Statement and Research Question

1.1 Problem Statement

Dyslexia is a learning disability that can cause difficulty with learning tasks such as reading, writing, and spelling. The National Health Service (NHS) estimates that one in ten people in the United Kingdom have dyslexia to some degree. It is important to note that dyslexia is not in any way a learning disability, this means it does not affect cognitive ability. Many people with dyslexia have average or even above average intelligence (National Health Service, 2022).

Dyslexic people could struggle to comprehend words and can have difficulty reading, sounding out words (phonology), understanding what they have read, or even pronouncing the words. They can struggle with memory tasks as well. It is a lifelong condition; the cure is not known. The best medical practitioners and educators can do at the moment is to tailor the method of learning to the individual so that symptoms of the disorder are decreased as much as possible. Furthermore, dyslexia can run along family lines. Families with a history of dyslexia are more likely to produce dyslexic children.

People living with dyslexia, especially children, may deal with discrimination, bullying, and self esteem issues. This is because dyslexia is usually accompanied by poor performance at school which can lead to feelings of isolation and inadequacy (Usman et al., 2021)). A lot of this trauma could be circumvented if the issue is diagnosed earlier and alternate methods of learning are applied.

Researchers and educators have come up with several ways to detect dyslexia. They are as follows:

1. Web-based and mobile games.
2. Eye movement tracking
3. Detailed scanning of subjects' brains: e.g Electroencephalogram (EEG) and Magnetic Resonance Imaging (MRI) scans.
4. Standardised tests (Usman et al., 2021).

The most common detection methods for dyslexia focus on the symptoms of the disorder and are usually used in schools. The reading and writing ability of a pupil is tested using standardised tests which are then evaluated by a human expert. Phonological awareness (how someone sounds out words) and the working memory of the pupil are also evaluated during these tests. These are time consuming, expensive, and can misdiagnose some dyslexics (Rello et al., Apr 23, 2018). This is because each person affected experiences dyslexia differently and the tests rely on the interpretation of a specialist.

1.2 Research Question

The question to be addressed by this research work is: "How can dyslexia be detected early, cheaply and efficiently? ".

2. Primary Research Plan

The first step in answering the research question is to conduct in-depth research using the resources at my disposal; the school library, online journals, and publicly available datasets to discover what methods are used presently for detecting dyslexia, especially in young children.

For the second step, the methods selected in the first step will then be scrutinised and the best one picked for further analysis. One interesting question to ask will be: what will determine which method is the best? The best method, for our purposes will be the method that scores the highest in a combination of these metrics:

1. Affordability: since the goal of the research is to come up with ways to detect dyslexia, the method should be available to the average school and should not require expensive machinery, or expert knowledge to be effective
2. Effectiveness: is the method correctly classifying dyslexic children? What is the accuracy of the method?
3. Speed: say a method rates highly in the above metrics. It will still be counter-intuitive to select a method that takes a long time to return results.

Given that there are multiple detection methods in general use presently, I plan to explore ways to make the selected method better, with respect to the metrics I have selected previously. Some ways to do this can include hyperparameter tuning, ensemble methods (using several machine learning algorithms together), or the development and/or selection of an entirely new algorithm.

2.1 Sources of Data

This work, as with Machine Learning projects in general, requires data to be able to function properly. The data will be used to train the selected model for detecting dyslexia. The dataset used by (Spoon et al., 2019) can be used and extended. A simple web page will be created for parents and guardians to upload previously handwritten samples for their wards. This part will be unstructured. We can also collect structured handwriting data from the students. This will involve them writing down words read to them. This part of the dataset will contain the same words for all the students.

2.2 Data Preprocessing and Model Training

A binary classification approach using Xing & Qiao's (2016) method for matching handwriting to writers is proposed. In this method, for each sample, the writer will either be dyslexic, or not. Since we plan to get more data to bolster the dataset in the research by (Spoon, et al., 2019), and also explore a host of ensemble methods, the model is expected to produce much better accuracy than the model presented by (Spoon, et al., 2019).

The models proposed include Random Forest Classifier, SVM, etc.

3. Ethical and Legal Considerations

As this project possibly involves collecting data about individuals, it raises some ethical and legal questions. These are discussed below:

1. Informed consent: The data should also be collected with the full informed consent of the individuals involved. Parental or guardian consent is needed in the case of minors.
2. Anonymity: The data collected will be purged of any personally identifiable characteristics. This means the data will be anonymous.
3. Ownership of data: Who owns the data that is collected? Can a subject, for instance, decide years after that they no longer want their data to be used in the detection process?
4. Validity: even though the parents and guardians of the students will be instructed not to help their wards with the writing tasks, there is always the risk that the legal guardians disregard those instructions.

A huge question is the question of confidentiality, since any data collected will be made publicly available (the available data is already publicly available), this means the research potentially violates confidentiality clauses. The way I have found to deal with this is the purging of personally identifiable features in the data as they are unnecessary to the detection of dyslexia.

4. Literature Review

4.1 Introduction

The detection of dyslexia has been a bone of contention for many years. Several different methods have been proposed. This report is mainly focused on the ones that took a Machine Learning (ML) approach.

4.2 Related Works

This section covers a review of previous works on the detection of dyslexia using machine learning methods. I attempt to provide an in-depth review of the most recent, most promising works on the topic.

(Spoon et al.,), 2019) is the work that is closest to the project that I am trying to undertake. In this piece of work, the authors collected unstructured handwritten samples from various pupils. Here, the parents or legal guardians of the child were asked to upload previously written text in the child's handwriting. They then gave each child a writing task containing words and a paragraph read by their parents or a researcher. These are used to build a dataset of handwriting. The dataset also contained demographic data such as the age of the child, grade, gender, dominant hand (i.e the hand they wrote with), ethnicity, parents' education history, income, whether they read to the child, marital status, etc. The technique they used was based on a previous handwriting recognition technique called DeepWriter by (Xing & Qiao,) 2016). This technique was developed to match analysed handwriting to the owners. The authors then tweaked this technique to assume there are only two writers; one with dyslexia, and one without. To predict the dyslexic students, the authors split each image of writing into lines and then generated random patches using (Arvanitopoulos & Susstrunk, Sep 2014) seam carving technique for historical manuscripts (Sep 2014). This technique was used to extract lines and then individual words from the handwriting data. The words were then randomised to make them language agnostic. The authors applied a Convolutional Neural Network and the model achieved an accuracy of 55.7% ±1.4%. This study suffered from a paucity of data and could benefit from an influx of data samples to improve its accuracy. Also, using a different algorithm or an ensemble method could improve the work.

(Jothi Prabha & Bhargavi, 2019) used a Machine Learning approach as well to detect dyslexia as well. However, these authors focused on analysing eye movement for the purposes of detecting dyslexia. The authors used a recording device called the Ober-2, which is a "wearable corneal reflection infra-red based eye tracking device". This device measured the horizontal (L_x and R_x) and vertical (L_y and R_y) positions of the eye while the subjects were given passages appropriate to their age level to read. The subjects were then asked questions about the passage they just finished reading to gauge their reading comprehension. The dataset built then contains raw eye tracking data for 185 subjects, 97 in the high risk (HR) group and the rest in the low risk group (LR). The architecture of the system is shown in the figure below

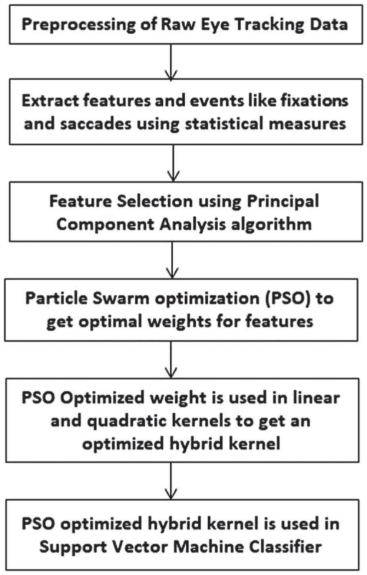
Preprocessing of Raw Eye Tracking Data
Extract features and events like fixations and saccades using statistical measures
Feature Selection using Principal Component Analysis algorithm
Particle Swarm optimization (PSO) to get optimal weights for features
PSO Optimized weight is used in linear and quadratic kernels to get an optimized hybrid kernel
PSO optimized hybrid kernel is used in Support Vector Machine Classifier

Figure 1. Architecture of the system (Jothi Prabha & Bhargavi, 2019)

As the architecture shows, features are extracted from the data, then Principal Component Analysis (PCA) is used on the dataset. This selects the most important features from the dataset for use with the ML algorithm. Particle Swarm Optimization, which is an optimization algorithm is then run on the dataset. This helps us get optimal weights for each of the features selected by the PCA algorithm. The optimised kernel is used in a Support Vector Machine (SVM) classifier and the performance of the optimised and linear models are compared with the results below:

Performance metrics	Linear SVM	Hybrid Kernel SVM-PSO
Accuracy	0.90	0.95
Sensitivity	0.94	1.00
Specificity	0.70	0.89
Positive predicted value	0.76	0.86

Negative predicted value	0.92	1.0

Table 1: Comparison of the performance of the Linear SVM and Hybrid SVM-PSO kernels (Jothi Prabha & Bhargavi, 2019)

The models both achieved admirable results. However, the hybrid SVM-PSO model outperformed the linear one by a significant margin. While this work achieved good accuracy, the subjects had to go through standardised testing to identify high risk and low risk individuals. The over reliance on expert opinion, along with the need for specialised equipment such as the Ober-2 is part of why this research work is being undertaken.

(Rello et al., Apr 23, 2018) used a series of features generated from a group of 267 participants playing a computer game to detect dyslexia. The game is a web game created using HTML, CSS, and Javascript with PHP and MySQL for the backend. Of the 267 subjects (aged 7 to 60), 54 were diagnosed with dyslexia, 206 did not have dyslexia, and 9 of the participants were judged at risk of having dyslexia. The group without dyslexia were used as the control group and all the participants had English as a first language although 84 of them could speak a second language. Each group was assigned a label; D (Dyslexic) for the dyslexic group, N (Not dyslexic) for the group without dyslexia, and M (Maybe) for the at risk group. The group were made to play a video game remotely and the following data points were recorded; number of clicks per item, number of correct answers (hits), number of incorrect answers (misses), accuracy (calculated as the number of hits divided by the number of clicks), and (miss rate) were assessed. When playing the game, players tried to maximise their scores by solving a linguistic problem as many times as possible within a 25-second window. The authors used a SVM algorithm on the dataset and obtained a prediction accuracy of 84.62%. However, The linguistic problem in the game might be too complex for 3-5 year olds to follow. This is the group we are targeting in our research as we are trying to get dyslexia diagnosed earlier to ensure that treatment and improvement can start as soon as possible.

(Christodoulides et al., 2022) used another approach. The authors acquired EEG signals from the brains of their study participants using a combination of a Brain Computer Interface device and an Interactive Linguistic Software tool. In the study, the EEG signals of the brain were measured for 12 students with dyslexia and 14 without. Using the software tool, the participants were measured in three experimental conditions; auditory discrimination, visual recognition, and visual recognition with background music. The features extracted from these experiments were then used to train a variety of models including a Random Forests Classifier, KNN classifier, Naive Bayes, Decision Trees, Support Vector Machine (SVM), and a type of Artificial Neural Network called a MultiLayer Perceptron (MLP) . The models achieved good accuracy with the Random Forest Classifier achieving the best across the board with at least 95% accuracy. The EEG scanned different sections of the brain using a commercial wearable EEG device called the Emotiv EPOC +. Using the device approximately 11.5 hours of signal records were taken from the participants (5 hours 51 minutes for participants with dyslexia, and 5 hours 47 minutes for the control group). This study seems to be the one with the most possibility for accuracy. However, the study was carried out on only university students. The dyslexic subjects had all been diagnosed early and have had interventions. Furthermore, the subjects were all right-handed. This might indicate that the study group was not random and we do not know if the success of this study will translate in the "real world". In addition, the study requires the use of an EEG device and a specialist which is what my study is trying to avoid.

4.3 Conclusion

The works covered in this literature review mostly required specialist intervention and sometimes specialised equipment. They focused on testing dyslexia using an online computer game, EEG diagrams, eye movement while reading, and finally the analysis of the subjects' handwriting.

Of the methods used, the handwriting analysis scored the least by a significant margin, indicating that there is the possibility of more work to be done on this method. The interest in this method is as a result of handwriting being easy to collect and the relative ease with which it can be analysed.

Bibliography

Arvanitopoulos, N., & Susstrunk, S. (2014). Seam carving for text line extraction on color and grayscale historical manuscripts. Paper presented at the 726-731. https://doi.org/10.1109/ICFHR.2014.127 https://ieeexplore.ieee.org/document/6981106

Christodoulides, P., Miltiadous, A., Tzimourta, K. D., Peschos, D., Ntritsos, G., Zakopoulou, V., Giannakeas, N., Astrakas, L. G., Tsipouras, M. G., Tsamis, K. I., Glavas, E., & Tzallas, A. T. (2022). *Classification of EEG signals from young adults with dyslexia combining a brain computer interface device and an interactive linguistic software tool*. Elsevier BV. https://doi.org/10.1016/j.bspc.2022.103646

El Hmimdi, A. E., Ward, L. M., Palpanas, T., & Kapoula, Z. (2021). Predicting dyslexia and reading speed in adolescents from eye movements in reading and non-reading tasks: A machine learning approach. *Brain Sciences, 11*(10), 1337. https://doi.org/10.3390/brainsci11101337

Jothi Prabha, A., & Bhargavi, R. (2019). Prediction of dyslexia from eye movements using machine learning. *Journal of the Institution of Electronics and Telecommunication Engineers,* , 1-10. https://doi.org/10.1080/03772063.2019.1622461

National Health Service. (2022). *Dyslexia.* https://www.nhs.uk/conditions/dyslexia/. Accessed: 27th March 2022.

Rello, L., Romero, E., Rauschenberger, M., Ali, A., Williams, K., Bigham, J., & White, N. (Apr 23, 2018). Screening dyslexia for english using HCI measures and machine learning. Paper presented at the 80-84. https://doi.org/10.1145/3194658.3194675 http://dl.acm.org/citation.cfm?id&#61;3194675

Spoon, K., Crandall, D., & Siek, K. (2019). Towards detecting dyslexia in children's handwriting using neural networks. Paper presented at the *International Conference on Machine Learning AI for Social Good Workshop,&nbsp;* pp. 1–5.

Szalma, J., & Weiss, B. (Jun 02, 2020). Data-driven classification of dyslexia using eye-movement correlates of natural reading. Paper presented at the 1-4. https://doi.org/10.1145/3379156.3391379 http://dl.acm.org/citation.cfm?id&#61;3391379

Usman, O. L., Muniyandi, R. C., Omar, K., & Mohamad, M. (2021). Advance machine learning methods for dyslexia biomarker detection: A review of implementation details and challenges. *IEEE Access, 9*, 1. https://doi.org/10.1109/ACCESS.2021.3062709

Xing, L., & Qiao, Y. (2016). DeepWriter: A multi-stream deep CNN for text-independent writer identification. Paper presented at the *&nbsp;15th International Conference on Frontiers in Handwriting Recognition,* https://doi.org/10.1109/ICFHR.2016.105