Kristu Jayanti Journal of Computational Sciences Volume 1 2021, Pages 29-36

A Study on different Classification Models for predicting Dyslexia

Vani Chakraborty

Department of Computer Science[PG], Kristu Jayanti College
vanichakraborty@kristujayanti.com

Abstract

Eye tracking technology is used to record the eye positions and the movements of the eye using the optical tracking of corneal reflections. Eye tracking data collected this way can be used in a wide variety of applications like gaming, marketing, cognitive ability and psychology. One of the applications of eye tracking data is to predict whether an individual has a learning disability like Dyslexia. Dyslexia is the most common neurological learning disability which manifests in the form of difficulty in reading and spelling. Although eye tracking data is recorded and available, there is a scarcity of studies done in analysing the data and understanding the hidden relationship and classifying it appropriately. This research intends to study different classification models like Logistic Regression, Gaussian NB, SVC, Decision Tree and so on. that are applied in the prediction of risk of Dyslexia. The paper also presents the result of the accuracy of different classification models in predicting the risk of Dyslexia.

***Keywords* :** *Dyslexia, Eye tracking, detection , Machine Learning, Classification models*

1. Introduction

According to International Dyslexia Association [1], Dyslexia is a neurological condition caused by a different wiring of the brain. There is no cure for Dyslexia but once diagnosed at a right stage, coping mechanisms can be devised [1]. According to the British Dyslexia Association, the number of individuals with Dyslexia in the UK is around 10% which is roughly 7.3 million people. But this number is also not considered as a true representation. It is said that around 16% of the worldwide population has Dyslexia[2]. The number of Dyslexics in India is roughly 15% among school going children. So the number would be 35 million and more [4]. According to Dr. Richard K. Wagner, Florida State University and Florida Center for Reading research, individuals with dyslexia are commonly misdiagnosed or even missed entirely. This problem is attributed to the fact that there is unreliability in diagnosis where only one single indicator is used for measuring [3]. No specific computational model exist in the literature for predicting dyslexia.

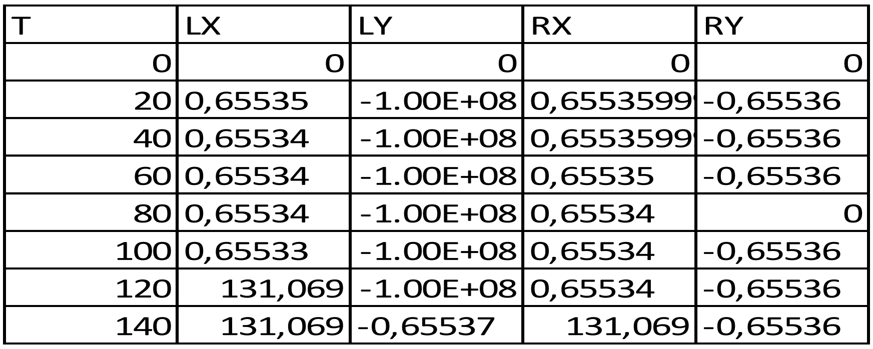
There are many methods to predict dyslexia using the conventional methods. The first and foremost method is to diagnose dyslexia usng oral and written assessments, taking the help of a trained psychologist. MRI scans can also be used as a good measurement for predictying dyslexia. All of these conventional tests are time consuming, expensive and involves lot of persistence from all the stake holders involved. One of the earliest studies suggested the relation between the tracking of the eye movements and its application in various domain. An important application of eye tracking data is to predict whether an individual is dyslexic or not. It is seen that there is a difference in the eye movement recording of individuals with and without dyslexia. When the eye tracking data is anlaysed using machine learning algorithms, we can predict with a reasonable amount of precision whether an individual is having dyslexia or not.

2. Literature Review

Nilsson Benfatto Et al did a study for collecting eye tracking data from 97 high risk subjects with reading difficulties and 88 low risk subjects. They used predictive modeling and statistical resampling techniques to develop classification models from eye tracking records with good accuracy. Did their work on predicting reading mistakes in children with reading difficulties based on eye-tracking data from real-world reading. The data used for this experiment stems from noisy readings outside the controlled lab conditions. They identified that gaze data improves the performance more than any other feature group and their models achieved good performance. Thomais Asvestopoulou analyzed eye movements during text reading to understand whether reading disorders can be predicted. They developed DysLexML, a screening tool for developmental dyslexia that applies various ML algorithms to analyze fixation points recorded via eye-tracking. They examined a large set of features based on statistical properties of fixations and saccadic movements and identified the one with prominent predictive power. Katie Spoon, David Crandall Et al have devised a way to detect dyslexia using handwriting. They collected data from K-6 children’s handwriting. Both controlled data set and experimental data set was collected. They have devised an automated early screening technique to be used in conjunction with other approaches to accelerate the detection process.

3. Dataset

We have used a dataset where the eye movement data is recorded of 185 subjects out of which 97 were high risk dyslexics and the rest 88 are low-risk subjects. The dataset was collected as part of Kronoberg reading development project, a longitudinal research project on reading development and reading disability in Swedish school children running between 1989 and 2010. The eye tracking was done using google-based infrared cornea reflection system, Ober-2. All the subjects were made to read one and the same text presented on a single page of white paper with high contrast. The raw eye tracking movements for each of the subjects is available in the form of a csv file with the following data. The first one is T, standing for the frequency or the time interval. For each of the time interval, lx,ly,rx and ry representing the left eye’s x and y positions and right eye’s x and y positions were recorded. In order to extract features from this raw eye data, we used a velocity threshold identification algorithm.

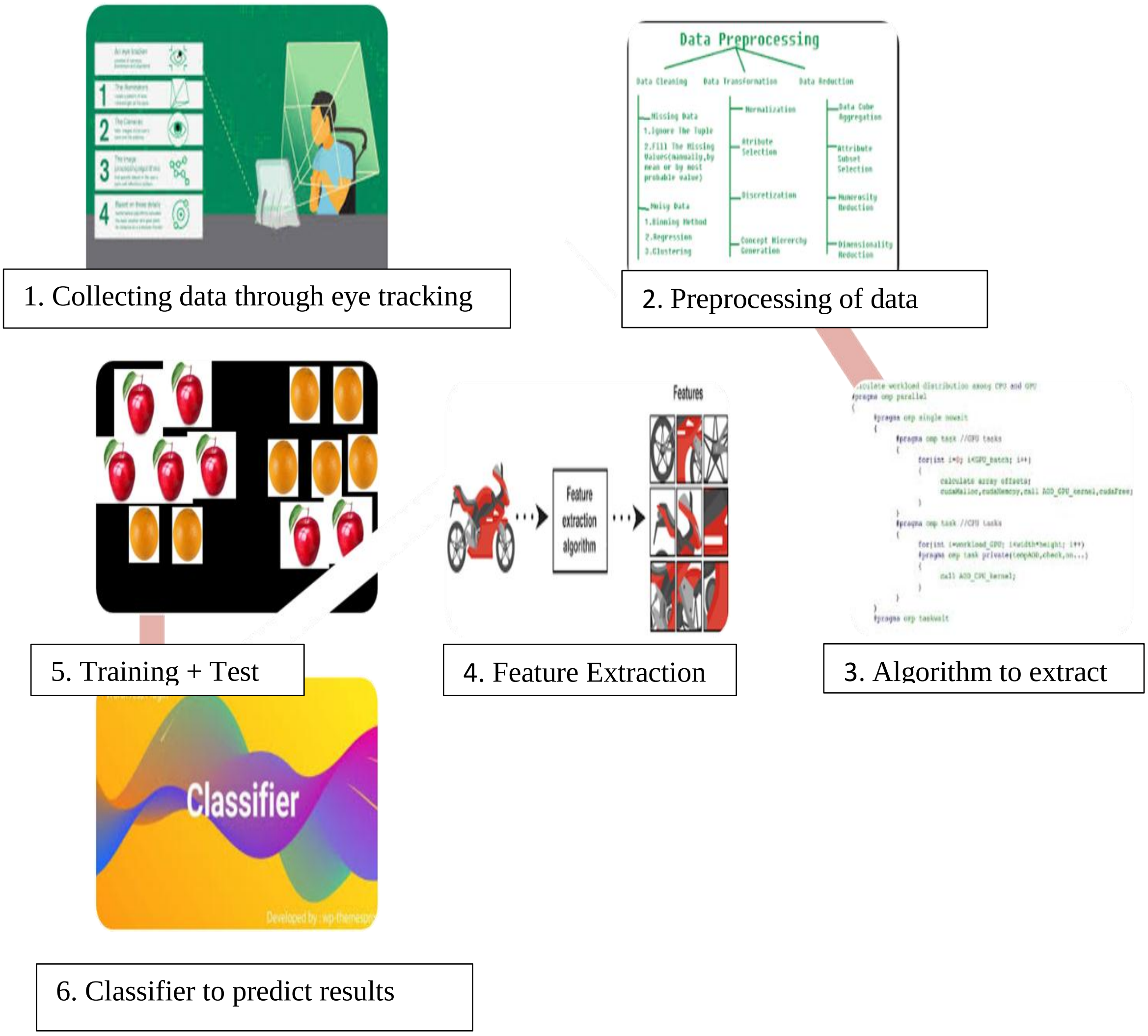


T	LX	LY	RX	RY
0	0	0	0	0
20	0,65535	-1.00E+08	0,6553599	-0,65536
40	0,65534	-1.00E+08	0,6553599	-0,65536
60	0,65534	-1.00E+08	0,65535	-0,65536
80	0,65534	-1.00E+08	0,65534	0
100	0,65533	-1.00E+08	0,65534	-0,65536
120	131,069	-1.00E+08	0,65534	-0,65536
140	131,069	-0,65537	131,069	-0,65536

Figure 2: Snapshot of the raw data obtained from tracking of eye movements

A python program was written to open each of the files holding the raw eye movements of each of the individual. Using the I-VT algorithm two important metrics are extracted. One is fixation and the other is saccade. Fixation is the time taken for processing a particular point in the image by our eyes. The time interval between two fixations is called a saccade. In order to identify fixations and saccades, the first step is to calculate the distance travelled between two consecutive time intervals of the recorded eye movement. If the distance is below a particular threshold (decided based on standard data), then it is considered as a fixation and if it is not it is considered as saccade. With this differentiation, from the data set available, a new csv file was created with each row being the entry for one subject. So the csv file has 185 rows and 101 features. One of the information is the label which is associated with each row, which is either a 0 or 1, 1 being the subject is high-risk for dyslexia and 0 is low-risk for dyslexia. That label is removed from the data frame to be used in the classification model. One more feature is gender which is again not going to influence the output and so that is also removed from the data frame. So now there are 99 features that are available for applying the classification model to predict whether a subject is high-risk or low-risk.

4. Proposed Methodology



```
alculate workload distribution among CPU and GPU
#pragma omp parallel
{
    #pragma omp single nowait
    {
        #pragma omp task //GPU tasks
        {
            for(int i=0; i<GPU_batch; i++)
            {
                calculate array offsets;
                cudaMalloc,cudaMemcpy,call ADD_GPU_kernel,cudaFree;
            }
        }
        #pragma omp task //CPU tasks
        {
            for(int i=workload_GPU; i<width*height; i++)
            #pragma omp task private(tempADD,check,nn...)
            {
                call ADD_CPU_kernel;
            }
        }
    }
    #pragma omp taskwait
```

Figure 1 : Proposed Methodology

The following are the detailed explanation of the steps in the proposed methodology.

1. Eye tracking data collected from individuals while reading a standard textual data.
2. The raw eye tracking data need to converted into metrics using standard algorithms.
3. From the eye metrics, all the possible features that can be extracted should be done.
4. From the available features, feature extraction should be performed to choose only those which are highly influential to the result.
5. The feature set should be divided as training + test data.
6. Classifiers to be applied on the test data to predict results after they are trained with the training data.

5. Classification Models

After extracting the features from the eye tracking data, a data set was obtained with 101 features including a label indicating the risk of Dyslexia or not.. Normalization was done on the data and our own implementation of the Logistic Regression algorithm was applied on the data . The accuracy of the classification obtained using that model was 91.89%.

```
iteration: 100
cost: 0.22563144672706406
```

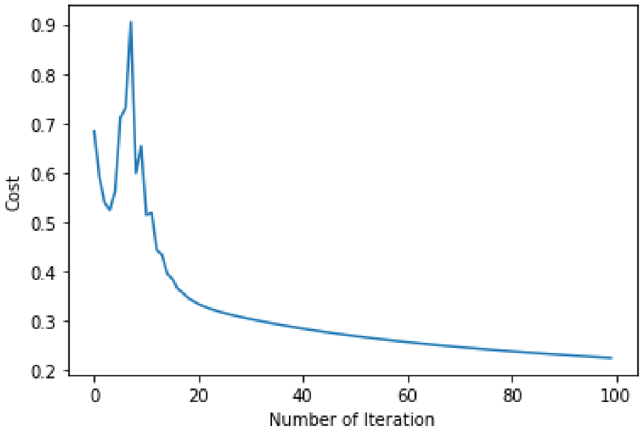


Figure 3 : Manual Test Accuracy: 91.89%

Six different ML algorithms were identified and used from the scikit package in Python. The algorithms used are Logistic Regression, Random Forest, KNeighborsClassifier, DecisionTreeClassifier, GaussianNB and SVC. RandomForest is an important technique used in supervised classification algorithms. KNN is one of the simplest machine learning algorithms. It is a non-parametric algorithm because there is no assumption made on the underlying data. KNN stores all the available data points and when a new data point arrives for classification, it does it based on the similarity.

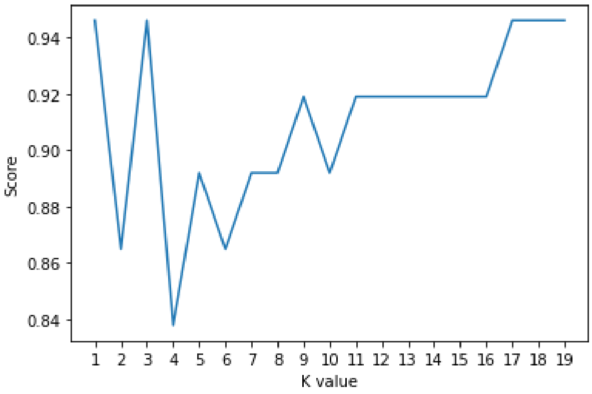


Figure 4 : Maximum KNN Score is 94.59%

Support Vector Classifier is again one of the most popular supervised machine learning algorithms. This algorithm is able to generate the best line or decision category that can differentiate or divide n-dimensional space into classes so that any new data point that arrives is placed in the correct category in the future. This best decision boundary is called a hyperplane.

DecisionTree classifier has a number of decision trees and each of these trees is based on various subsets of the given dataset and in order to improve the accuracy of the data set, average is calculated. The final output is predicted based on the prediction from each tree.

LogisticRegression is based on predicting a dependent variable which is categorical, based on a set of independent variables. LogisticRegression is used for solving classification problems. The following graph depicts the comparison of the performance of the above mentioned Machine Learning algorithms.

```
Input Data Shape =  (185, 101)
Training Data Label
0.0    60
1.0    69
dtype: int64
Test Data Label
0.0    28
1.0    28
dtype: int64
```

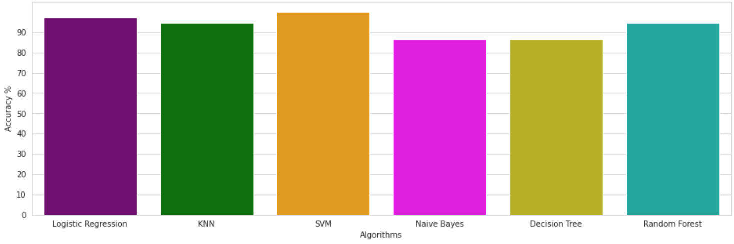


Figure 5 : Comparison of Different ML algorithms

From the above graph, one can see that the performance of Naïve Bayes and Decision Tree are not as efficient as other algorithms and they do not have to be pursued further. The confusion matrix obtained thus of applying the different algorithms on the obtained data set is also given.

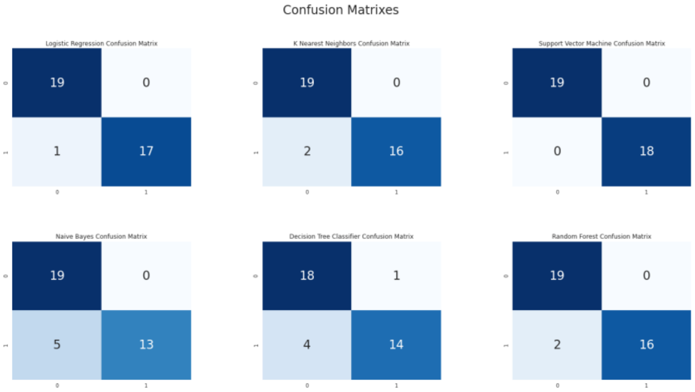


Figure 6 : Confusion Matrix

6. Conclusion and Future work

In this paper, different classifiers to predict the risk of dyslexia is implemented, from the eye tracking data collected. The classifiers used for comparison are Logistic Regression, Gaussian NB, Decision Tree, SVC, KNN and naïve bayes. To increase the performance of the classifiers, different kinds of features will be extracted from the raw eye tracking data by using various spatial and temporal algorithms like I-DT and I-HMM. The number of metrics extracted from the eye tracking data also can be increased so as to get better classification accuracy. The better purpose of this research is to increase the classification accuracy of the classifiers.

References

[1]. Tiffanye McCoy-Thomas, "Eye Tracking and Learning Predictability",Journal of International Education and Practice, Volume 02, Issue 04,(2019),pp 11-18

[2]. Jothi Prabha A, Bhargavi R, Et al, "Predictive Model for Dyslexia from Eye Fixation Events", International Journal of Engineering and Advanced Technology(IJEAT), Volume-9, Issue-1S3, (2019), pp 235-240

[3]. Suraj Shrestha and Pietro Murano, "An Algorithm for automatically detecting Dyslexia on the fly", International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 3, (2018)

[4]. Masooda Modak, Ketan Ghotane, Et al, "Detection of Dyslexia using Eye Tracking Measures", International Journal of Innovative Technology and Exploring Engineering(IJITEE), Volume-8, Issue-6S4, (2019),pp 1011 - 1014

[5]. Ying Xu, Huailong Li Et al, "Knowledge Reprsentation and Design about Expert System for Learning-Disability Check for Children" , IEEE, 2011, doi: 10.1109/ITiME.2011.6132174, pp. 567-570.

[6]. T-K Wu, S-C Huang Et al, "Customizing Asynchronous Parallel Pattern Search Algorithm to improve ANN Classifier for Learning Disabilities Students Identification", Seventh International Conference on Natural Computation, (2011), doi: 10.1109/ICNC.2011.6022322, pp. 1639-1643.

[7]. Oussama Tahan, Farah Barake, “A Gaming Environment to Train Teachers Diagnose Children Learing Disabilities”, IEEE, Published in 2018 14th International Computer Engineering Conference(ICENCO), doi: 10.1109/ICENCO.2018.8636127,pp 48-51.

[8]. Parikh, Saurin. “Eye Gaze Feature Classification for Predicting Levels of Learning.” (2018).AIED 2018 workshop proceedings.

[9].Mattias Nilsson Benfatto, Gustaf Oqvist Et al, “Screening for Dyslexia usng Eye Tracking during Reading”. PLos ONE 11(12): e0165508, doi:10.1371/journal.pone.0165508, (2016).

[10]. Joachim Bingel, Maria Barrett, Et al, “Predicting misreadings from gaze in children with reading difficulties”, Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, (2018), doi = "10.18653/v1/W18-0503", pp 24-34

[11]. Asvestopoulou, T., Manousaki, V., Psistakis, A., Smyrnakis, I., Andreadakis, V., Aslanides, I.M., & Papadopouli, M. (2019). “DysLexML: Screening Tool for Dyslexia Using Machine Learning”, ArXiv, abs/1903.06274,pp 121-127

[12]. Katie Spoon, David Crandall, “Towards Detecting Dyslexia in children’s handwriting using neural networks”, https://aiforsocialgood.github.io/icml2019/accepted/track1/pdfs/43_aisg_icml2019.pdf, 2019

[13] Selvi.H, Saravanan M.S, “A study of dyslexia using different machine learning algorithm with data mining techniques”, International Journal of Engineering & Technology, doi: 10.14419/ijet.v7i4.14545,(2018), pp 3406-3411

[14] Dario D. Salvucci & Joseph H.Goldberg, “Identifying Fixations and Saccades in Eye-Tracking Protocols, Proceedings of the Eye Tracking Research and Application Symposium” , New York: ACM Press, pp. 71-78.