# Detection Tool for Dyslexia

Kaushal Suresh
*Computer Science & Engineering*
*PES University*
Bangalore, India
kaushalsureshkr@gmail.com

Kavisha Mathur
*Computer Science & Engineering*
*PES University*
Bangalore, India
kvsha2020@gmail.com

Krishna Bajaj
*Computer Science & Engineering*
*PES University*
Bangalore, India
bajaj.krishna01062002@gmail.com

***Abstract*— Dyslexia is a learning disorder that involves difficulty reading due to problems identifying speech sounds and learning how they relate to letters and words (decoding). This paper aims to diagnose children with dyslexia via methods like handwriting recognition using Convolutional Neural Network (CNN), analysis of electroencephalographic (EEG) signals using regression models and neural networks, and by using a scoring parameter based on the results of questions asked to the subject regarding their language, memory, speed, visual and audio skills. EEG tests were conducted on 12,000 subjects, and were pre-processed before being fed into the model. Screening results from the models built in this research provided an accuracy of 99.6% for EEG dataset and 98.6% for the handwriting dataset. EEG signals, an image of the handwriting of the subject, and the scores of the test are given as input, and a final score is given as output, signifying low, moderate and high chances of dyslexia for the patient. The product shows significant potential, and is an easy-to-use, enabling and quick tool for doctors in diagnosing patients with dyslexia.**

***Keywords*— Machine Learning, dyslexia, handwriting, convolutional neural network, classification***

## I. Introduction

Dyslexia is a disability that involves deficiencies in reading and writing capabilities, but does not affect intellect. Although this condition was commonly known as 'word blindness' in the 1800s, it has now been identified as a condition with a neurological origin and not as a condition to do with lack of vision. According to the National Center for Learning Disabilities (NCLD) dyslexia is among the five most common types of learning disabilities around the world [1]. Children with dyslexia may have difficulty in reading, and some clinical symptoms of dyslexia being mispronunciation of words, slow reading, reading ability below the expected age, difficulty with phonemic awareness, fluency, rhyming comprehension and written expression. Approximately 15% of people in the world are dyslexic [2].

These learning disabilities cannot be completely cured by medications. Treatment of these learning disabilities consists of providing extra training to overcome the symptoms. In developing countries, disabilities like dyslexia are often undiagnosed due to a lack of awareness. Moreover, very few screening and intervention applications have been developed, which use deep learning and machine learning approaches, and none of them address both screening and intervention of dyslexia in one application.

There are many techniques proposed by past research to detect indicators of dyslexia [3]. One of these methods is detection using 'behavioural' symptoms and aspects. This is one of the more conventional methods, and is currently used by most psychologists to diagnose dyslexia. This method assesses whether a person has dyslexia using highly recognised standardised tests on their sensory and neurological skills like language, vocabulary, memory, speed, visual discrimination and audio discrimination.

Another category is the use of brain imaging techniques to portray distinctive brain behaviours, using methods like magneto-encephalography (MEG), magnetic resonance imaging (MRI), electroencephalography (EEG) and positron emission tomography (PET). For dyslexic people, difficulties in reading comprehension will lead to problem in written expression. Hence, handwriting imaging and recognition techniques is another way of diagnosing dyslexia.

The primary objective of this paper is to come up with a tool for doctors to diagnose adults with dyslexia based on the EEG signals of their brain, an image of a paragraph written by them, or by testing and scoring their sensory skills. Section II discusses the literature survey. Section III talks about the dataset collection methods. Section IV explains the methodology used in making the system. Section V gives the system results based on accuracy. Section VI discusses the conclusion.

## II. Literature Review

Year by year, researchers have come out with different approach in detecting dyslexic kids at an early stage by using different approach and method such as through brain behavior and through their handwriting. Brain activity is one of the prominent researches held by researcher in order to detect dyslexia since the problem is based from different functioning of the brain itself. From the previous study in [4], proposed a technique that used Electroencephalography (EEG) average FFT index that focused on writing disorder. The output shows that there are differences in hemispheric activation in brain from the test perform.

A deep learning and machine learning based mobile application, named "Pubudu" was developed for screening and intervention of dyslexia, dysgraphia and dyscalculia supporting local languages by R. Kariyawasam, M. Nadeeshani, T. Hamid, I. Subasinghe, P. Samarasinghe and P. Ratnayake [5]. In "Pubudu", clinical screening and diagnostic procedures recommended were followed by health professionals for screening and intervention. The screening of dyslexia, letter dysgraphia and numeric dysgraphia was carried out using deep neural network and the screening for dyscalculia was carried out using machine learning techniques. Intervention techniques are implemented using gamified environments. System testing was carried out using 50 differently abled children and 50 typical children.

A method of transfer learning of Dyslexia handwriting recognition by using Convolutional Neural Network (CNN) based on famous architecture of handwriting recognition using of LeNet-5, was proposed by M. S. A. B. Rosli, I. S. Isa, S. A. Ramlan, S. N. Sulaiman and M. I. F. Maruzuki [7]. Data augmentation and pre-processing was employed to a total of 138,500 handwriting image dataset before feeding it into network. The hyper-parameter of the model was tuned and analyzed to classify the 3 classes of dyslexic handwriting. The developed CNN model had successfully achieved a remarkable accuracy of 95.34% in classifying 3 classes of dyslexic handwriting. We have also used a convolutional neural network in our project for handwriting recognition.

A thorough review of EEG-based pattern classification frameworks for dyslexia was conducted by H. Perera, M. F. Shiratuddin & K. W. Wong [3] focusing on each framework's (1) data collection, (2) pre-processing, (3) analysis and (4) classification methods. A wide range of inputs as well as classification approaches were experimented for the improvement in EEG-based pattern classification frameworks, and pros and cons of existing EEG-based pattern classification frameworks were inferred. for dyslexia and recommend optimisations through the findings to assist future research. The research also compared multiple classification models such as Linear discriminant analysis, neural networks, Support vector machines, etc. and concludes that SVMs are recommended as a promising classifier to be used in EEG-based pattern classification frameworks for dyslexia.

## III. Dataset Collection

Detection of Dyslexia is mainly done using user study to collect user data. For the detection techniques, psychologists examine behavioral aspects of participants during standardized tests, such as reading and writing, phonological awareness, working memory and handwriting. Dyslexic participants are identified based on their poor score in these tests. Usually these tests are taken to collect various types of data such as text, Eye-movement, MRI scans, EEG scans , and images over a prolonged period of time.

In our study, we have taken the scores of different standardized tests and EEG signal data from 10 college students while they watched MOOC video clips. The students wore a single-channel wireless MindSet that measured activity over the frontal lobe. The MindSet measures the voltage between an electrode resting on the forehead and two electrodes (one ground and one reference) each in contact with an ear. After each session, the student rated his/her confusion level on a scale of 1-7, where one corresponded to the least confusing and seven corresponded to the most confusing. These labels if further normalized into labels of whether the students are confused or not. This label is offered as self-labelled confusion in addition to the predefined label of confusion. Each data point consists of 120+ rows, which is sampled every 0.5 seconds (so each data point is a one minute video). Signals with higher frequency are reported as the mean value during each 0.5 second.

Handwriting data was collected from 3 sources where letter is from NIST Special Database 19 and some datasets for testing is from dyslexic kids of Seberang Jaya primary school, Penang, Malaysia.

The third dataset we have taken are the results of a test a number of people took along with their survey score. It is labelled 0,1 and 2 which tells low, moderate and high chance of having dyslexia. This dataset has been standardised with a set of formulas that our base project has taken. The dataset includes the following attributes- 'Language_vocab', 'Memory''Speed','Visual_discrimination','Audio_Discrimination', 'Survey_Score' and 'Label'

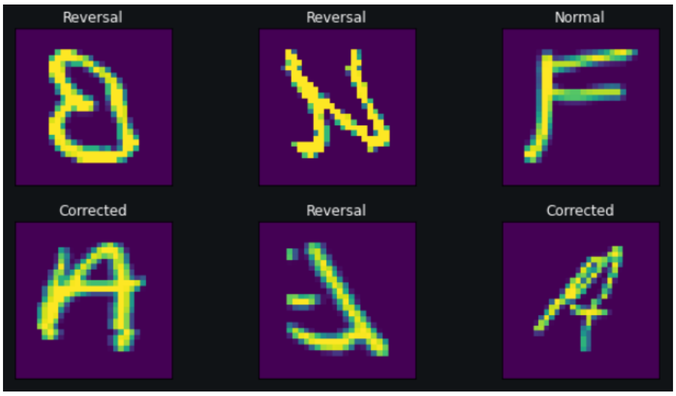


Fig 1 : Handwriting Dataset

| | SubjectID | VideoID | Attention | Mediation | Raw | Delta | Theta | \ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 56.0 | 43.0 | 278.0 | 301963.0 | 90612.0 | |
| 1 | 0.0 | 0.0 | 40.0 | 35.0 | -50.0 | 73787.0 | 28083.0 | |
| 2 | 0.0 | 0.0 | 47.0 | 48.0 | 101.0 | 758353.0 | 383745.0 | |
| 3 | 0.0 | 0.0 | 47.0 | 57.0 | -5.0 | 2012240.0 | 129350.0 | |
| 4 | 0.0 | 0.0 | 44.0 | 53.0 | -8.0 | 1005145.0 | 354328.0 | |

| | Alpha1 | Alpha2 | Beta1 | Beta2 | Gamma1 | Gamma2 | predefinedlabel | \ |
|---|---|---|---|---|---|---|---|---|
| 0 | 33735.0 | 23991.0 | 27946.0 | 45097.0 | 33228.0 | 8293.0 | 0.0 | |
| 1 | 1439.0 | 2240.0 | 2746.0 | 3687.0 | 5293.0 | 2740.0 | 0.0 | |
| 2 | 201999.0 | 62107.0 | 36293.0 | 130536.0 | 57243.0 | 25354.0 | 0.0 | |
| 3 | 61236.0 | 17084.0 | 11488.0 | 62462.0 | 49960.0 | 33932.0 | 0.0 | |
| 4 | 37102.0 | 88881.0 | 45307.0 | 99603.0 | 44790.0 | 29749.0 | 0.0 | |

| | user-definedlabeln |
|---|---|
| 0 | 0.0 |
| 1 | 0.0 |
| 2 | 0.0 |
| 3 | 0.0 |
| 4 | 0.0 |

Fig 2 : EEG Data

| | subject ID | age | ethnicity | gender |
|---|---|---|---|---|
| 0 | 0 | 25 | Han Chinese | M |
| 1 | 1 | 24 | Han Chinese | M |
| 2 | 2 | 31 | English | M |
| 3 | 3 | 28 | Han Chinese | F |
| 4 | 4 | 24 | Bengali | M |

Fig: Demographic profiles of dyslexic patients

## IV. Proposed approach

*Machine Learning Algorithms*

For our handwriting dataset we have applied CNN-The Convolutional Neural Network (CNN or ConvNet) is a subtype of Neural Networks that is mainly used for applications in image and speech recognition. Its built-in convolutional layer reduces the high dimensionality of images without losing its information.

For our next EEG dataset we have used ANN and SVM. Artificial neural network (ANN) is a computational model that consists of several processing elements that receive inputs and deliver outputs based on their predefined activation functions. They are a computational network based on biological neural networks that construct the structure of the human brain. Neural networks perform better for EEG classifications compared to other methods since it can be used to implement boundaries for nonlinear classifications. Nevertheless, to acquire the desired level of accuracy, it is important to choose a suitable number of hidden units, which can become problematic. Having a larger number of hidden units than required results in memorising the training set which causes poor generalisation.

For our next dataset we are choosing the best model amongst SVM(Support Vector Machine), Naive Bayes and Random Forest.
Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well, it is best suited for classification. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.
Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature
Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. This was a type of multilabel classification as we have 3 labels- 0,1 and 2 for low, moderate and high dyslexia.

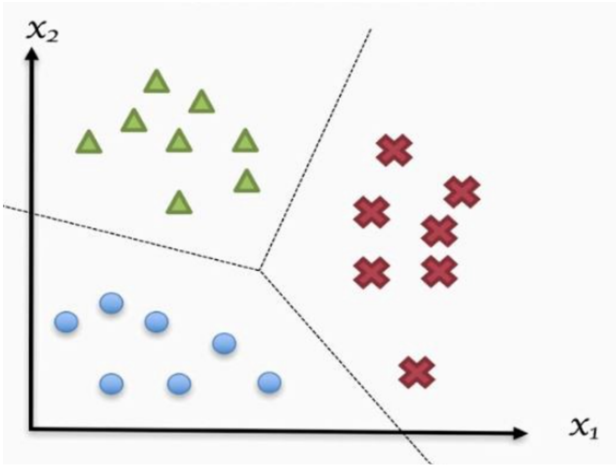


Fig. 3 Multilabel classification

The formula used to conclude the parameters in our dataset were as follows

We have 500 datapoints in both the labelled and unlabelled csv, the preprocessing and data cleaning was already done.
Calculations of Scores -
(Shown as per respective question numbers)
Language_vocab = (Ans.1 + Ans.2 + Ans.3 + Ans.4 + Ans.5 + Ans.6 + Ans.8)/28
Memory = (Ans.2 + Ans.9)/8
Speed = Calculated on the basis of time taken to complete the quiz
Visual_discrimination = (Ans.1 + Ans.3 + Ans.4 + Ans.6)/16
Audio_Discrimination = (Ans.7 + Ans.10)/8
Survey_Score = (Sum of all answers)/80

## V. Experiments and results

The Convolutional Neural Network (CNN) is applied to the application of detection of handwriting as it automatically identifies deep powerful features. In this application, the CNN is implemented using Keras and Tensorflow. There are 4 convolutional layers, 2 max-pooling (MP) layers, 3 fully-connected (FC) layers and an output layer in the CNN. Figure shows the details of the components used in our experiment.
It is not advisable to pass the entire dataset into the neural net at once due to the memory constraint. So, the dataset is divided into a number of batches of size "batch size", and repeatedly iterated over the entire dataset for a certain number of "epochs".

Before training our model on the EEG data, we first combined the EEG scans dataset with the demographics of the patients, in order to not miss out on any important features. Then, the top features are selected using mutual information values, and

the data of these features are split into train and test data, and are then fed into 2 models, an ANN and an SVM model. This is a type of binary classification, as we are predicting the user-defined label to be either 0 (for low chances of dyslexia) or 1 (for high chances of dyslexia).
There are six dense layers in the model, each followed by a batch normalization layer, and a dropout layer to reduce overfitting. The activation function used in each dense layer is 'relu', except for the final layer, where we have used the 'sigmoid' activation layer. The optimizer used is Adamax, and the loss function used is Binary Cross-entropy. On comparing this with the SVM model, we concluded that ANN has higher accuracy. After taking the sample inputs shown in Table VI, the model classified the patient to have higher chances of having dyslexia.

TABLE III

SUMMARY OF PERFORMANCE OF EEG SCANS DATASET

| **Model** | **Accuracy** | **Loss** |
|---|---|---|
| 1.Support Vector Machine | 0.912 | - |
| 2. Artificial Neural Network | 0.997 | 0.021 |
| 3. Random Forest | | |

TABLE IV

USER INPUTS FOR EEG SCANS MODEL

| **Model** | **User Input** |
|---|---|
| Subject ID | 0 |
| Age | 25 |
| Video ID | 1 |
| Attention | 50 |
| Mediation | 69 |
| Raw | 55 |
| Delta | 716119 |
| Theta | 154465 |
| Alpha 1 | 74192 |
| Alpha 2 | 19414 |
| Beta 1 | 29154 |
| Beta 2 | 85313 |
| Gamma 1 | 36150 |
| Gamma 2 | 48516 |

For our standardised datasets, we ran 3 models, SVM, Random Forest and Naive Bayes. We then concluded that SVM model has the best accuracy and ran the same model on our unlabelled dataset. We also take inputs from user in order to classify the person if they are dyslexic (low,moderate and high).
After taking sample user inputs the model classified as this person having moderate dyslexia.

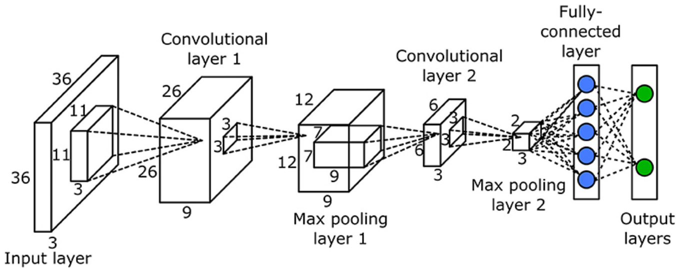


Fig. 4 Model for handwriting dataset

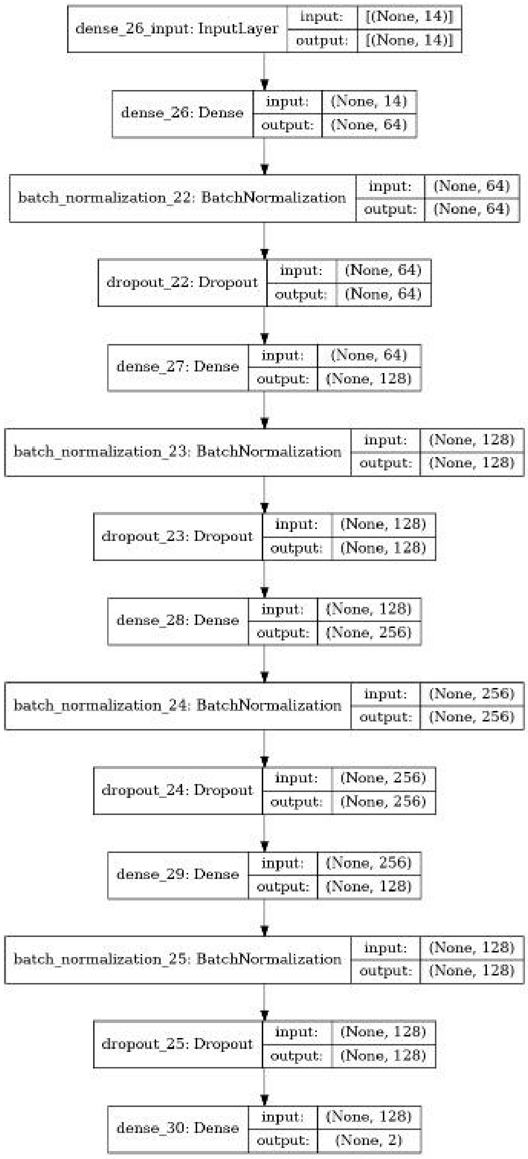


Fig. 4 Model for EEG Dataset

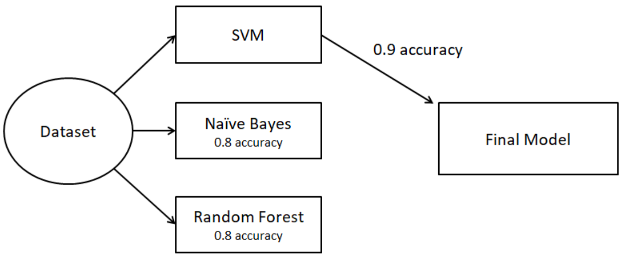


Fig. 5 Best model for standardised dataset

TABLE V

Summary of performance of standardised test dataset

| Model | Accuracy |
|---|---|
| 1.Support Vector Machine | 0.99 |
| 2. Naive Bayes | 0.8 |
| 3.Random Forest | 0.8 |

TABLE VI

User inputs for standardised test model

| Model | User Input |
|---|---|
| Enter name of applicant | XYZ |
| Score for Language Vocabulary test | 0.1 |
| Score of Memory Test | 0.3 |
| Score for Speed Test | 0.5 |
| Score for Visual Discrimination | 0.6 |
| Score for Audio Discrimination | 0.2 |
| Score from Survey | 0.8 |

## VI. Conclusions

Through this paper, we tried to propose a product that doctors in the real world can use, since we are using 3 very important set of parameters to classify a person into dyslexic and not dyslexic. We use the EEG signals, handwriting and the standardised test parameters that are fed in the front end UI/UX page give a set of results to the doctor after computing the parameters. The doctor can thus further form a conclusion based on his/her or domain knowledge.

## Acknowledgment

We would like to acknowledge and express our gratitude to our guide, Mr. Srinivas Katharguppe. Alongside our CSE department which provided us with the necessary resources and tools to execute our project.

## References

[1] "5 Most Common Learning Disabilities -Masters in Special Education Degree Program Guide", Masters-in-special-education.com, 2019.[Online1.Available:https://www.masters-in-specialeducation.com/ lists/5-most-common-learning-disabilities/. [Access ed:31- Oct- 20191.

[2] Stampoltzis, Aglaia, and Stavroula Polychronopoulou. "Dyslexia in Greek higher education: A study of incidence, policy and provision."Journal of Research in Special Educational Needs 8, no. 1 (2008): 37-46.

[3] Perera H, Shiratuddin MF, Wong KW. Review of EEG-based pattern classification frameworks for dyslexia. Brain Inform. 2018 Jun 15;5(2):4. doi: 10.1186/s40708-018-0079-9. PMID: 29904812; PMCID: PMC6094381.

[4] S. N. Farihah, K. Y. Lee, W. Mansor, N. B. Mohamad, Z. Mahmoodin and S. A. Saidi, "EEG average FFT index for dyslexic children with writing disorder," 2015 IEEE Student Symposium in Biomedical Engineering & Sciences (ISSBES), 2015, pp. 118-121, doi: 10.1109/ISSBES.2015.7435880.

[5] R. Kariyawasam, M. Nadeeshani, T. Hamid, I. Subasinghe, P. Samarasinghe and P. Ratnayake, "Pubudu: Deep Learning Based Screening And Intervention of Dyslexia, Dysgraphia And Dyscalculia," 2019 14th Conference on Industrial and Information Systems (ICIIS), 2019, pp. 476-481, doi: 10.1109/ICIIS47346.2019.9063301.

[6] R. Kariyawasam, M. Nadeeshani, T. Hamid, I. Subasinghe and P. Ratnayake, "A Gamified Approach for Screening and Intervention of Dyslexia, Dysgraphia and Dyscalculia," 2019 International Conference on Advancements in Computing (ICAC), 2019, pp. 156-161, doi: 10.1109/ICAC49085.2019.9103336.

[7] M. S. A. B. Rosli, I. S. Isa, S. A. Ramlan, S. N. Sulaiman and M. I. F. Maruzuki, "Development of CNN Transfer Learning for Dyslexia Handwriting Recognition," 2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 2021, pp. 194-199, doi: 10.1109/ICCSCE52189.2021.9530971.

[8] O. L. Usman, R. C. Muniyandi, K. Omar and M. Mohamad, "Advance Machine Learning Methods for Dyslexia Biomarker Detection: A Review of Implementation Details and Challenges," in IEEE Access, vol. 9, pp. 36879-36897, 2021, doi: 10.1109/ACCESS.2021.3062709.