Eye Movement Feature Set and Predictive Model for Dyslexia:
Feature Set and Predictive Model for Dyslexia

Jothi Prabha Appadurai, Vellore Institute of Technology, Chennai, India
Bhargavi R., Vellore Institute of Technology, Chennai, India

ABSTRACT

Dyslexia is a learning disorder that can cause difficulties in reading or writing. Dyslexia is not a visual problem, but many dyslexics have impaired magnocellular system, which causes poor eye control. Eye-trackers are used to track eye movements. This research work proposes a set of significant eye movement features that are used to build a predictive model for dyslexia. Fixation and saccade eye events are detected using the dispersion-threshold and velocity-threshold algorithms. Various machine learning models are experimented. Validation is done on 185 subjects using 10-fold cross-validation. Velocity-based features gave high accuracy compared to statistical and dispersion features. Highest accuracy of 96% was achieved using the hybrid kernel support vector machine-particle swarm optimization model followed by the xtreme gradient boosting model with an accuracy of 95%. The best set of features are the first fixation start time, average fixation saccade duration, the total number of fixations, total number of saccades, and ratio between saccades and fixations.

KEYWORDS
Convolution Neural Networks, Dispersion-Threshold, Ensemble Techniques, Hybrid Kernel SVM–PSO, Hyper-Parameter Tuning, Particle Swarm Optimization, Velocity-Threshold

1. INTRODUCTION

Dyslexia is a kind of neurodevelopmental deficit that causes persistent difficulties in reading and writing. Dyslexia affects around 5-15% of school-going children (Karande & Agarwal, 2017). Dyslexic children tend to have problems in academics such as poor letter recognition, slow reading, poor writing, and spelling. Dyslexics have poor information processing skills causing short term memory loss while reading (Karande & Agarwal, 2017). Dyslexics don’t have poor vision but they tend to have poor structural connectivity between the left visual thalamus (LVT) and middle temporal visual area (V5/MT), a cerebral cortex region responsible for retinotopic movement (Tschentscher et al., 2019).

Eye-tracking to analyze language processing has emerged as a new research trend since the early '90s (1994-2020). Eye trackers are used for detecting gaze points during eye movements. In linguistic studies, they are used to identify saccades, fixations, smooth pursuit eye movements. Saccades are abrupt and quick jumps while reading a text. Saccadic movements move our gaze towards the next word of the passage or text. During saccades, the mind does not process the text and no visual information is registered in the brain. Fixations are gaze points where the visual gaze is stationery at one location.
for some time during which word processing or comprehension occurs (Aryadoust & Ang, 2019). While reading a line or a passage, the eyes of a normal reader does not gaze over the line or passage continuously. The reading process has two types of eye movement events fixations and saccades that occur alternatively. Eye movement events give us information such as the time spent by the reader on a particular word or a line, thereby helping us to know their reading pattern (Wang & Ren, 2019). Eye trackers help to detect eye movement patterns (i.e.) saccades and distance between successive saccades (Brunyé et al., 2019). Exploring and identifying a set of significant features of saccades and fixations along with machine learning algorithms help us in the early prediction of dyslexia. An improved automated computational model for the prediction of dyslexia can be designed using these significant features (Rello & Ballesteros, 2015).

This paper analyzes and identifies a set of eye movement features that contribute more to the prediction of dyslexia using machine learning algorithms. An automated computational model is built using this significant eye movement features. The rest of this paper is organized as follows. Section 2 summarizes the existing methods in the literature for prediction of dyslexia, Section 3 gives a brief description about dataset used in this research, Section 4 explains the methodology and implementation of the proposed model, Section 5 describes the experimental result and analysis, Section 6 shows the summary of results, Section 7 shows the comparative analysis, and Section 8 shows the conclusion and future work.

2. LITERATURE SURVEY

Blythe & Liversedge have observed that there is an abnormality in connectivity between left auditory cortico thalamus pathway and planum temporale. It has been observed that dyslexics have less white matter thereby reducing the strength of connectivity. Connectivity strength plays a major role in reading comprehension and fluency. This work gives novel evidence that there is a correlation between the structure of the brain and reading skills. From this work, it has been observed that dyslexics tend to have different eye movement patterns and reduced reading fluency because of the difference in white matter connectivity (Blythe & Liversedge, 2018). Muller & Kriegstein observed that there is a structural difference between sensory thalamus and left hemisphere- cerebral cortex in dyslexics. The white matter connectivity between the lateral geniculate nucleus (LGN) and the primary visual cortex is tested as its the pathway for vision management. It has been reported that alterations in white matter can cause visual stream dysfunction in dyslexics. They have suggested that cortico thalamic interactions and sub-cortical sensory path give an understanding of dyslexia (Müller & Kriegstein, 2017). Xu P. et al. observed that visual abnormalities can lead to reading problems and there could be numerous reasons that can be addressed using multidisciplinary methods. The coordination of eyes and pupil movements gives us a better understanding of their reading skills. Analysis of fixations and repeated fixations helps in identifying whether he is a normal or slow reader. It has been observed that oculomotor abnormalities have a high impact on reading comprehension (Xu & Xiao, 2015).

Holmqvist K et al. reported that fixation duration is usually 200-250 milliseconds for normal readers. Saccadic jumps occur for about 7 to 9 letters. Saccades are measured in terms of characters of letters. Saccades can be backward and forward movements while reading. These forward and backward saccadic jumps are called regressions which constitute 10 to 15 percentile of reading time. Words that have more than 6 characters have more fixations than small words. Fixation duration and regressions are observed to be more for poor or slow readers (Holmqvist et al., 2011).

Nystrom, M et al. worked on analyzing the pupil movements using eye trackers. The gaze positions are also considered as pupil position and movement alone may not be able to capture eye movement events correctly. Since gaze points are considered, calibration has to be done for every user which is an overhead. Velocity threshold algorithms were used to compute saccades based on the velocity of the movement of the eyes (Nyström & Holmqvist, 2010). Hessel A et al. observed reading comprehension in children while reading texts from their gaze points. The experiment was
carried out on mono-lingual and bi-lingual children. Both groups performed almost the same way. Children having more vocabulary tend to read better and have less re-reading. The amount of time spent on re-reading helps in understanding the reading pattern and comprehending ability of the children (Hessel et al., 2019).

Himmelstoss N et al. investigated on word recognition during reading. They explored how contextual predictions take place during word recognition. Eye movements along with fMRI and EEG are recorded and the impact of neuronal oscillations on reading has been observed. Eye movement along with ERP signals gives insight into understanding the reading process and word recognition (Himmelstoss et al., 2019). Elena Arabadzhiyska et al. identified a new model for computing a saccade landing position from the gaze locations while viewing images. Fast saccadic jumps cause a mismatch between image quality and the gaze positions. To solve this problem, a model is built to predict the endpoint of saccades, and an image for new fixation is given instead of using the actual gaze position. This reduced system latency and also improved the quality of images (Arabadzhiyska et al., 2017).

Maike Schindler et al. investigated the eye and mind hypothesis and its applications in geometry. They studied the role of fixations and saccades in reading the student mental process. Saccades were helpful compared to that of fixations. It has been observed that fixations do not help in understanding the mental process when the subject thinks about something else instead of solving the given problem. Emotional arouses were also easily identified with saccades (Schindler et al., 2019). Reinhard Werth proposed a pseudo word learning experiment to understand the causes of dyslexia. They conducted reading experiments on pseudo words of 3 to 6 characters in length. From the experiment, it has been observed that longer fixations and premature saccades cause different eye movement patterns in dyslexics. Also, few dyslexics do not have the capability of sequencing the characters or letters to form words (Werth, 2019). Tsunehiko Kurokami et al. proposed that dyslexics have reading problems due to nystagmus which is an abnormality in eye movements. Repeated vertical eye movement was observed. It has been reported that there could be a strong neurological path mechanism that causes this abrupt eye movement in dyslexics. No cerebellar dysfunction or vision problem has been observed in the experimented dyslexic subject (Kurokami et al., 2019).

Anne E. Cook et al. have done eye-tracking studies on reading comprehension of passages. Three sets of eight passages were to be read by the subjects. At the end of the passage, a question was asked to check the comprehension skill of the subject. The number of regressions both forward and backward regressions play a major role in comprehending (Cook & Wei, 2019). Marianna Stella et al. examined the differences in eye movements and examined the cognitive factors that cause poor reading comprehension in dyslexics. They analyzed problems with syntactic mistakes and cognitive factors such as working memory and processing speed that causes abnormal eye movements in dyslexics. It has been reported that dyslexics have fewer comprehension skills, take more time to read compared to normal readers (Stella & Engelhardt, 2019). Thomas DW Wilcockson et al. examined the oculomotor inhibitions. They focus on pro-saccade and anti-saccade eye movements to understand about cognitive control. It has been reported that dyslexics have problems with volitional inhibition control which is observed from the antisaccade eye movement event. Most of the literature focuses on fixations for analysis while this work concentrates on saccadic movements too and observed that saccades have significant importance in reading (Wilcockson et al., 2019).

From the literature, it has been observed that there is a structural difference in the neural connectivity of dyslexics when compared to normal subjects that cause reading problems. Fixations and saccades play a major role to predict dyslexics. There is no much eye movement feature analysis done for the prediction of dyslexia. Also, there is no remarkable automated computational model available for the prediction of dyslexia. In this research work, the above problems are addressed as follows. A set of significant eye movement features were computed and an automated computational is built for better prediction of dyslexia. As far as we know, the proposed significant eye movement feature set has not been used earlier for building an automated model to predict dyslexia.
3. DATA COLLECTION

Ober-2, a screen-based eye tracker was used to capture eye gaze points. The eye tracker records the reading time \( t \) in milliseconds (ms). The eye gaze was recorded both horizontally \( (L_x \text{ and } R_x) \) and vertically \( (L_y \text{ and } R_y) \). An infrared headset is worn by subjects while reading. The distance from eyes to screen is around 45 cm. Horizontal and vertical movements of the left and right eye were captured. The participants read a passage of text as per their age complexity and their eye movements were examined while reading. To test their comprehension skills, three questions related to passage are to be answered. The raw eye movement data used in this research was acquired from the research work of (Benfatto et al., 2016). The dataset has eye tracker recordings of 185 subjects whose age is between 9 and 10 years. Among these 185 subjects, 97 are dyslexic and 88 are non-dyslexic. Subjects having low intelligent quotient and mental disorders were excluded. Subjects in the dyslexic group are already assessed by class teachers as dyslexic and have problems in reading. Subjects in the non-dyslexic group have average or good reading skills.

4. METHODOLOGY

This research work proposes an eye movement feature set for the prediction of dyslexia and also builds an automated computational model for the prediction of dyslexia. The proposed prediction model for dyslexia is shown in Figure 1 the raw eye-tracking data is preprocessed to eliminate irrelevant data such as blinks and out-of-range values caused by improper calibration of the eye tracker. The raw data is analyzed for different eye movement events such as fixations and saccades. Various aggregated eye movement features related to these events are generated using basic statistical measures, dispersion, and velocity-based approaches.

Feature selection is done to identify the most relevant features that contribute to the prediction of dyslexia. Feature Selection algorithms experimented are Principal Component Analysis (PCA) and Recursive feature Elimination with Cross-Validation (RFE-CV). Logistic regression (RFE-LR) and Support Vector Machine (RFE-SVM) are used as base estimators for RFE-CV. The best features selected by the feature selection algorithms were used for building a classification model for the prediction of dyslexia.

Machine learning classifiers that were experimented in this research work are Support Vector Machine (SVM), Hybrid Kernel SVM-PSO, Ensemble techniques –boosting and bagging and Deep Learning Convolution Neural Networks (CNN). The dataset is divided in the ratio of 80% for training and 20% for testing. 10 fold cross validation was performed. Also to reduce the variance the process.
is iterated thrice on random splits. The averaged F1 score across the 10 folds was considered as final. F1 score is chosen as the final evaluation metric as the cost of false negatives and false positives are critical while predicting dyslexia. Subjects having dyslexia should not be predicted as non-dyslexic as they cannot go for remediation classes for development and also non-dyslexic subjects should not be predicted as dyslexics as it can cause psychological effects on the subject especially children. Hence the average F1 score is used for prediction.

4.1. Pre-Processing of Data

Preprocessing is done to remove blinks and to handle missing data. Standardization of data is also done to improve the accuracy of the model. Blinks are replaced by zeros and missing values were set to have the mean values of their occurrences.

4.2. Feature Generation and Evaluation

Eye movement features related to fixations and saccades were analyzed and their related features are derived using basic statistic measures and also using dispersion and velocity-based algorithms which computes fixations based on spatial characteristics of the data (Salvucci & Goldberg, 2000). Dispersion-threshold based algorithms (I-DA) computes the fixations based on the distance between successive fixation points. The dispersion threshold is set as 33 pixels. Dispersion algorithms assume that fixations occur as clusters as the user gazes at a particular spread during fixations. The steps of the I-DA algorithm are shown in Figure 2.

In this research work, 52 features related to fixations and saccades were generated from the fixation list computed by using dispersion based algorithms (Salvucci & Goldberg, 2000). Eye movement features derived through dispersion based algorithm are total number of fixations, fixation duration, and repeat fixations, distance between fixations(horizontal), distance between fixations(vertical), total number of saccades, saccade amplitude, scan path length, the ratio of saccades and fixations and duration of the task.

Velocity-threshold based algorithms (I-VA) (Salvucci & Goldberg, 2000) computes the fixations and saccades based on the time and acceleration of movement of eyes while reading. Velocity is measured in terms of time taken for eye movement between gaze points. It assumes that fixations will have low-velocity rates compared to saccades as the user’s eye is stationary during fixation. Two input parameters inter-pixel distance criteria and duration criteria are required for calculating the fixations. The inter-pixel distance criteria and duration criteria are 25 mm and 50 milliseconds respectively. The steps of the I-VA algorithm are shown in Figure 3.
In this research work, 40 features were generated from fixations and saccades computed by velocity-based algorithms. The blink related features were eliminated as it is null and of no use for eye movement analysis. Eye movement features derived through velocity-based algorithm are Fixation start time, Fixation end time, Duration of fixation, Average Fixation points (horizontal), Average Fixation points (vertical), Saccade start time and end time, Saccade duration, No of saccades, No of fixations and No of blinks. The statistical features computed, the machine learning models and feature selection algorithms proposed in this work (Jothi Prabha & Bhargavi, 2019) are implemented and comparative analysis was done while building the classification model for dyslexia. All features were generated for the left and right eye separately. Aggregated features such as maximum fixation or saccade duration, the central tendency of fixations, saccades and regressions were derived using statistical measures such as mean, standard deviation, and median.

5. FEATURE SELECTION

Machine learning classifier performance is mainly dependent on the input features. Irrelevant features and noise can cause the model to have poor performance. When a subset of features is selected instead of the complete feature set, it reduces the training time of the model. It also minimizes the over fitting of data and also reduces the complexity of the model. Feature selection when used to select a right subset of attributes or can give increased classification accuracy.

Feature selection methods are of two types of filter methods and wrapper methods. Filter methods select the subset of features without using machine learning algorithms. They select features based on statistical tests and correlation between the features. Wrapper methods use machine learning models and cross-validation (CV) score to find the best subset of features. Commonly used wrapper methods are forward selection, backward elimination, and recursive feature elimination. This work implements both filter and wrapper methods. Principal Component Analysis (PCA) is a filter method used for transforming features into principal components based on the variance. PCA studies the correlation and variance. PCA transforms the data using linear transformations. Eigenvalues and Eigenvectors of the same direction are identified and represented as principal components (Raiko et al., 2007). The number of principal components is selected based on the scree plot which shows the explained variance of the data. The advantage of PCA is that they are less prone to over fitting compared to wrapper methods. PCA converges at 10 components for statistical and dispersion features. It converges at 5 components for velocity-based features as shown in the scree plots of Figure 4. The convergence point is chosen as the number of components for input to classifiers.

Recursive Feature Elimination (RFE), unlike that of PCA, gives the list of best features. PCA gives us a combination of the features as principal components. Recursive feature elimination with
cross-validation (RFE-CV) is a wrapper method that uses cross-validation to find the optimal number of features to be in the subset. RFE uses a greedy approach to find the subset of features. It starts from the complete feature set and recursively creates a model by storing the best and worst performed models. It gives rank to each feature based on the order of removal (Lin et al., 2018, Huang et al., 2014) They tend to perform better than filter methods but computational complexity is more. The advantage of RFE-CV over RFE is that it uses the cross-validated score to select the optimal number of features. The base estimators chosen in this work are Support Vector Machine (SVM) and Logistic Regression (LR). Figure 5, Figure 6, and Figure 7 show the optimal number of features selected by RFE–CV using Logistic regression (RFE-CV-LR) and RFE-CV using SVM as estimator(RFE-CV-SVM) for statistical, dispersion and velocity features.

Figure 4. Plot a. Scree plot of statistical features, Plot b. Scree plot of dispersion features, Plot c. Scree plot of velocity based features

Figure 5. Statistical features selected by RFE-CV-LR and RFE-CV-SVM based on CV score
The significant set of features selected by feature selection algorithms are the first fixation start time, average fixation saccade duration, the total number of fixations, total number of saccades and ratio between saccades and fixations. These significant features are given as input for classification. It has been observed that velocity-based features gave better FI score and accuracy compared to that of statistical and dispersion based features.

6. CLASSIFICATION

The significant set of features selected by the feature selection algorithms are used to build a predictive model for dyslexia using various machine learning classifiers. SVM, Hybrid Kernel SVM-PSO algorithms are used in this research work as they are widely used in literature for the prediction of diseases. This work also uses State of art ensemble boosting and bagging algorithms and deep learning algorithm Convolutional Neural Networks.

6.1. Support Vector Machine (SVM)

SVM is used for classification and regression problems. In this research work, the basic SVM kernels are used to predict dyslexia. SVM has gained its popularity in different applications such as text and digits recognition and prediction of diseases. SVM tends to work well on smaller datasets. The main objective of SVM is to draw an optimal hyper plane to divide linear separable data. In cases where
the data is not linearly separable, it transforms the data into a new dimension where it can be linearly separated. The kernel functions of SVM are useful for performing these transformations. The hyper plane is drawn with the help of support vectors. Support vectors are data points which are near to the boundary of the separating classes and it is tough to classify. SVM tries to draw maximum size margin with the help of these support vectors (Charoen et al., 2017, Qiu & Lane, 2000)

\[ W^T X + b = 0 \]  

Equation (1) is the function defining the separating hyper plane where \( W \) is a weight vector, \( X \) is input vector and \( b \) is bias. If the data is non-linearly separable in the original dimension, it has to be transformed into a higher-dimensional space which involves more computation time and complexity, especially for high dimensional data. Kernels are similarity functions that reduce the complexity without having to do computations on higher-dimensional space. Basic SVM kernels are linear, polynomial, sigmoid and radial basis function kernel (Mo & Xu, 2010). The equations of linear, polynomial, Radial basis function (RBF) and sigmoid kernels are shown in Equations (2), (3), (4), (5) respectively.

**Linear Kernel**

\[ k(x, z) = x^T z \]  

**Polynomial Kernel**

\[ k(x, z) = (x^T z)^d \]  

**Radial Basis Kernel**

\[ k(x, z) = \exp\left(-\frac{x - z^2}{2\sigma^2}\right) \]  

**Sigmoid Kernel**

\[ k(x, z) = \tanh(x^T z + c) \]  

Here \( x \) and \( z \) are data vectors and \( c \) is constant.

SVM with linear, polynomial, sigmoid and RBF kernels are used for statistical, dispersion and velocity based features. It has been observed that SVM-Linear and SVM-RBF kernels gave better accuracy while predicting dyslexia.

**Hybrid Kernel SVM-Particle Swarm Optimization (SVM-PSO)**

The hybrid kernel SVM-PSO classifier proposed in (Jothi Prabha & Bhargavi, 2019) is used in this research work. This classification model does feature weighting with the help of Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995, Esmin et al., 2015). Feature weighting is done to remove the irrelevant features that do not contribute much for classification. PSO gives a weight for each feature and the dataset is scaled with these weights. Irrelevant features get less weight and relevant features get more weights (Dongoran et al., 2018, Bamakan et al., 2015). The hybrid SVM kernel is a combination of linear and quadratic kernels. It has been observed that there is no one specific kernel to address all types of problems. Combinations of existing kernels to form new kernels have
shown improved performance rather than using a single kernel. The hybrid kernel SVM equation is shown in Equation (6)

\[
\text{Hybrid Kernel SVM: } k(x, z) = x^T z + (x^T z)^2
\]  

(6)

The experimental results show that Hybrid Kernel SVM has given the highest accuracy for all types of eye movement features generated. It also gave high accuracy in par with state of art Xtreme Gradient Boosting algorithm (XGBoost).

6.2. Ensemble Methods – Bagging

Bagging is an ensemble technique that learns homogenous weak learners and aggregates them by using averaging measures. Every learner is learned parallel and independently. Bagging techniques focus on building a model with less variance. In this research work, a random forest classifier is used for the prediction of dyslexia. Random Forest is an improvised version of decision tree algorithm. Random forest generates many decision trees from random subset. Every decision tree identifies the different patterns in the data. For classification problems, majority voting is used to classify a sample to a group or class. Random forest generates complex separating boundaries. The main advantage of random forest is it calculates feature importance and works well on large scale datasets. One more added advantage is there is no need to perform normalization or scaling while using random forest (Bernard et al., 2010). The Random Forest algorithm is presented in Figure 8.

Significant improvements have been observed in classification accuracy while growing random forest. The error rate in a random forest approach is mainly dependent on two features correlation and strength. An increased correlation between trees causes an increase in error rate. The tree which has the lowest error rate is the best classifier. The error rate of the Random forest can be reduced by increasing the strength of every decision tree. Random Forest gave better accuracy for velocity-based features compared to that of dispersion based features and statistical features while predicting dyslexia.

6.3. Ensemble Methods – Boosting

Boosting is an ensemble technique that fits a weak learner model iteratively, builds an ensemble model by updating the training data, considering the pros and cons of the current ensemble model. Learning is done sequentially and the current model depends on the previous one. In this research
work ensemble boosting methods such as Adaptive Boosting (AdaBoost), Light Gradient Boosting Machine (LGBM) and Xtreme Gradient Boosting (XGBoost) are used. AdaBoost is a simple ensemble machine learning algorithm. Its base model is the decision tree. Several sequential weak learning models are fit while trying to improve the errors of the last model. AdaBoost assigns weight to incorrectly predicted observations so that the next trained model makes better predictions than the current model (Li et al., 2008). AdaBoost uses Equation (7) for the initialization of weights, Equation (8) and Equation (9) for calculation of weights and Equation (10) for updating of weights.

\[ S_{t,i} = \frac{1}{N} \quad (7) \]

\[ \beta_j = \frac{E_j}{1 - \beta_j} \quad (8) \]

\[ \alpha = \log \left( \frac{1}{\beta_j} \right) \quad (9) \]

\[ H(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_i h_t(x) \right) \quad (10) \]

Where \( S \) is the dataset, \( N \) is the number of samples, \( t \) is the number of iterations, \( E_j \) is the error rate and \( H(x) \) is the output function and \( \alpha \) is to minimize the exponential error.

Light Gradient Boosting Machine (LGBM) is an implementation of Gradient Boosting Decision Tree which overcomes the efficiency and scalability problems. LGBM is used for ranking and classification tasks. It uses less memory and gives more accuracy and it is capable of handling big data. In many experiments, LGBM has outdated the performance of other boosting classifiers. LGBM is different from other tree-based algorithms in the way the trees are grown. In LGBM the tree grows vertically (i.e.) leaf wise growth. The leaf with maximum delta loss is grown and this helps to reduce the amount of loss compared to level-wise growth (Wang et al., 2017). Figure 9 shows the leaf-wise tree growth in LGBM.

XGBoost is an efficient and simple ensemble machine learning algorithm that can be used for classification and regression. It has an added advantage of in-built regularization which prevents the model from over-fitting. It also supports parallel processing making it faster than GBM. XGBoost is an ensemble of decision trees where the newly built trees overcome the errors of the previously built tree. New trees are constructed until no further improvement is observed in the model (Chen & Guestrin, 2016). Figure 10 shows the learning process of XGBoost classifier.

Hyper parameter optimization was done for XGBoost, LGBM and AdaBoost algorithms for building a classification model for dyslexics. Optimization was done for statistical, dispersion and velocity based features. Two types of commonly used methods for hyper parameter tuning are Grid Search and Randomized search parameter tuning. Randomized Search Cross Validation was used for tuning XGBoost and LGBM parameters. Grid Search Cross-validation was used to tune the parameters of AdaBoost. Hyper parameter optimization improves the efficiency and accuracy of the model as the
machine learning algorithms learn based on the parameters which can be tuned as per the problem. The objective of this tuning is to get the best combination of parameters.

Grid search is a simple method for parameter tuning. Grid search does an exhaustive search to build a model by trying out all possible combinations of parameters that are given in the grid. The best parameters are selected based on the cross-validated score. In this research work, Grid Search is used for optimizing the hyper parameters of AdaBoost. Randomized Search does a random search over the possible values for parameters. The advantage of this method over grid search is that the cost is independent of the number of parameters and adding more parameters does not affect the accuracy.
or efficiency of the model. The parameter values are selected based on the cross-validated score. The hyper parameter tuning does in this research work is shown in Figure 11. Experimental results have shown there is an improvement in the accuracy of the model after hyper parameter optimization. Accuracy has been improved for statistical, dispersion and velocity features after tuning parameters.

6.4. Deep Learning Algorithm: Convolution Neural Networks

Convolution Neural Networks (CNN) is a deep learning algorithm that can be used for the classification of images. It accepts an image as input, assigns bias and learnable weights to different objects in the image so that the images can be differentiated from one another (Dongoran et al., 2018). Pre-Processing cost is less while using CNN compared to other machine learning classifiers. CNN has an inbuilt ability to learn about filters and hence preprocessing made easier and simple. CNN architecture is comparable to the structural connectivity of neurons in the human brain. Neurons respond to stimuli only in some areas of the visual region. CNN is good at capturing the spatial characteristics of images with the help of filters. When large size images are given as input, CNN reduces the image to a form such that no critical features that contribute more to prediction are lost and also easy to process. CNN is good at learning features and is also scalable to huge datasets.

CNN processes several layers in sequence and each layer data is represented from low and high levels. CNN uses the information in these layers to extract features automatically. A basic model of CNN has 4 main layers such as Conv or convolutional layer, rectified linear unit (ReLU), Pooling and fully connected layer. Figure 12 shows the architecture of the CNN model. An image is structured as a matrix of pixels. When an image is given as input CNN predicts the class which the image belongs to and also the probability of the image being in different classes. The four layers have four different functions to perform. CNN uses convolution or Conv layer and feature extraction filters to learn the feature weights from the training data. Each filter detects a feature and is represented as a feature map. The feature map records the spatial location of objects in the image. ReLU introduces non-linearity to help in improving the computational efficiency of the model without affecting the accuracy. A sigmoid or tanh function is usually used in RELU as non-linear functions. The pooling layer aims at reducing the size of the feature maps and also tries to reduce the computational cost of the system. Max pooling layers output the important features and their relative location. In the fully
connected layer, the input is multiplied by weights and bias is added to the output. The output is an n-dimensional vector that gives the probability of the image belonging to a particular class. Here n is the number of classes or categories (Atabay, 2016, Howard et al., 2017).

In this research work, fixations are computed from the gaze points given by the eye tracker. The fixations are converted to images and the scan path is also drawn based on the fixations and saccades computed. The fixations and scan path were given as input images to CNN which further classifies whether the fixation or scan path image is of dyslexic or non-dyslexic. Figure 13 shows the sample fixation and scan path images used in this research work.

The experimental results show less accuracy with the given small dataset. Predictive accuracy was high with fixation images compared to that of scan path images. CNN tends to give higher accuracy in voluminous data and also with high-quality images. Hence experimenting further with large scale
datasets with good image quality and with subjects of different age groups can help us improve the predictive accuracy of the model.

7. EXPERIMENTAL RESULTS AND ANALYSIS

Eye gaze points recorded by the eye tracker are used as input to this research work. The train test split ratio of the dataset is 80:20. The experiments of this research are done using Python 3 framework in Windows 10 -64-bit operating system, with Intel® core i-7-8550U GPU. Several python packages such as sci-kit learn, Keras, matplotlib, pyswarms, seaborn, pandas are used.

Three types of features such as statistical, dispersion and velocity based are generated from the raw eye gaze data. For these 3 different sets of features generated, a significant set of features are selected by using feature selection algorithms PCA and RFE-CV. The optimal number of features selected by these feature selection algorithms is shown in Figure 7, Figure 8 and Figure 9. The three sets of features are given as different inputs for the classification model. 10 fold cross-validation is performed. One out of 10 in each fold is kept aside for testing and the remaining is used for training. The model is trained with the 80% training data and validated with the 20% test data. The performance metrics considered in these experiments are accuracy, F1 score, Precision, and Recall and specificity. All experiments in this works have label-0 for non-dyslexic and label -1 for dyslexics.

Cost analysis is an important measure to be considered while building predictive models for diseases. The cost of falsely predicted as Positive (False Positive) is high as non-dyslexic is predicted as dyslexic. This causes the child to go through unnecessary remediation classes and also may cause a psychological disturbance in kids. The number of correctly predicted as positive (True Positive) should be high as all dyslexics have to be predicted correctly. Precision focuses on correct positive predictions only and recall focuses on missed positive predictions. F1 score is one metric which balances both precision and recall as a single metric. Hence F1 score is considered as final metric for evaluating the classification models. Other metrics such as precision, recall, and specificity are also considered. Equation (11) to Equation (15) shows the computation of these evaluation measures.

\[
Accuracy = \frac{\sum Correctly\_classified\_instances}{\sum All\_instances} \quad (11)
\]

\[
Specificity = \frac{True\_Negative}{False\_Positive + True\_Negative} \quad (12)
\]

\[
Recall = \frac{True\_Positive}{True\_Positive + False\_Negative} \quad (13)
\]

\[
Precision = \frac{True\_Positive}{True\_Positive + False\_Positive} \quad (14)
\]
7.1. Analysis of Statistical Features

Statistical based features are computed by taking aggregations of eye position while reading. Basic statistical measures such as mean, median, standard deviation and mean absolute deviation are used to compute these aggregated features. The accuracy chart for prediction of dyslexia from statistical features using various machine learning classifiers is shown in Figure 14.

The Hybrid Kernel SVM-PSO classifier gave the best accuracy of 92%. SVM linear Kernel gave an accuracy of 90%. SVM-RBF kernel gave an accuracy of 89%. SVM Sigmoid kernel gave an accuracy of 81%. Ensemble boosting algorithm XGBoost gave the highest accuracy of 92%, LGBM classifier gave an accuracy of 90% and AdaBoost gave an accuracy of 91%. Random Forest gave an accuracy of 89. It has been observed that XGBoost and Hybrid Kernel SVM-PSO classifiers gave better predictive accuracy compared to other models. These classifiers have shown the best results irrespective of the feature selection method used.

7.2. Analysis of Dispersion-Based Features

Dispersion features are generated based on the fixation points computed by the dispersion-threshold algorithm (I-DA). The accuracy chart for prediction of dyslexia from dispersion features using various machine learning classifiers is shown in Figure 15.

The Hybrid Kernel SVM-PSO classifier gave the best accuracy of 94%. SVM linear Kernel gave an accuracy of 87%. SVM-RBF kernel gave an accuracy of 86%. SVM Sigmoid kernel gave an accuracy of 92%. Random forest gave an accuracy of 89%. Ensemble boosting algorithm AdaBoost gave the highest accuracy of 93%, LGBM classifier gave an accuracy of 90% and AdaBoost gave an accuracy of 91%. Accuracies achieved for fixation images and scan path images with deep learning algorithm CNN are 87% and 82% respectively. It has been observed that AdaBoost and Hybrid Kernel

$$ F1Score = 2 \times \frac{precision \times recall}{precision + recall} $$

(15)
SVM-PSO classifiers gave better predictive accuracy compared to other models. These classifiers have shown the best results irrespective of the feature selection method used.

### 7.3. Analysis of Velocity-Based Features

Velocity features are computed based on the fixations and saccades detected by the velocity-threshold algorithm (I-VA). The accuracy chart for prediction of dyslexia from velocity features using various machine learning classifiers is shown in Figure 16.

The Hybrid Kernel SVM-PSO classifier gave the best accuracy of 96%. SVM linear Kernel gave an accuracy of 90%. SVM-RBF kernel gave an accuracy of 91%. SVM Sigmoid kernel gave an accuracy of 89%. Random forest gave an accuracy of 93%. Ensemble boosting algorithm XGBoost gave the highest accuracy of 95%. LGBM classifier gave an accuracy of 92% and AdaBoost gave an accuracy of 91%. Accuracies achieved for fixation images and scan path images with deep learning algorithm CNN are 88% and 87% respectively. It has been observed that XGBoost and Hybrid Kernel...
SVM-PSO classifiers gave better predictive accuracy compared to other models. These classifiers have shown the best results irrespective of the feature selection method used.

### 8. SUMMARY OF RESULTS

To summarize, among the three types of features generated, velocity-based features gave the highest accuracy of 96%. Hence velocity based features can be used as a feature set for prediction of dyslexia. Hybrid Kernel SVM-PSO performs well for statistical, dispersion and velocity based features. The accuracies achieved for statistical, dispersion and velocity features are 92%, 94%, and 96% respectively. Hence Hybrid Kernel SVM-PSO classifier can be used to build an automated computational model for prediction of dyslexia.

State of the art ensemble boosting algorithm XGBoost gave good accuracies for statistical (92%) and velocity based features (95%). AdaBoost ensemble boosting algorithm gave good accuracy for dispersion-based features. It has been observed that computational time for ensemble methods is more when compared to Hybrid Kernel SVM-PSO. Deep Learning CNN gave us less accuracy (87%) when compared to SVM and Ensemble models. From the experimental results, it has been observed that over fitting occurs and the training accuracy becomes 100% after a few epochs. CNN tends to perform well on large scale datasets. The dataset used in this research work is small hence it could have caused over fitting of the model.

### 9. COMPARATIVE ANALYSIS

The proposed model is compared with existing methods and models for the prediction of dyslexia. Dyslexia has a wide range of features and there are various methods to classify dyslexics from nondyslexics. Magnetic Resonance Imaging (MRI), functional MRI (fMRI), electroencephalogram signals (EEG), questionnaires and eye movements are some of them (Benfatto et al., 2016). Table 1 shows the findings of these methods. It has been observed that the Hybrid Kernel SVM-PSO classifier with proposed velocity based features gave the highest accuracy when compared to existing methods to predict dyslexia.

### 10. CONCLUSION AND FUTURE WORK

In this research work, a set of significant eye movement features is generated that has a prominent contribution to the prediction of dyslexia. Three sets of features statistic, dispersion and velocity based were generated from the gaze points. Experimental results show that velocity-based features outperform
other sets of features in performance. The best set of features that gave good accuracy are the first fixation start time, average fixation saccade duration, the total number of fixations, total number of saccades and ratio between saccades and fixations. These features gave a high accuracy of 96% with the Hybrid Kernel SVM-PSO model followed by XGBoost classifier which gave 95% accuracy. As future work, dispersion features can be used to predict autism disorder. Dispersion features have scan path regressions and saccade amplitude which are prominent features for the prediction of autism. Deep Learning models can be used on a large dataset for better feature extraction and classification.

ACKNOWLEDGMENT

This work has been partially developed under a grant from Innovation and Entrepreneurship Development Cell (IEDC).
REFERENCES

Arabadzhyska, E., Tursun, O. T., Myszkowski, K., Seidel, H. P., & Didyk, P. (2017). Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics*, 36(4), 1–12. doi:10.1145/3072959.3073642

Aryadoust, V., & Ang, B. H. (2019). *Exploring the frontiers of eye-tracking research in language studies: a novel co-citation scientometric*. Academic Press.

Atabay, H. A. (2016). A convolutional neural network with a new architecture applied on leaf classification. *The IIOAB Journal*, 7(5), 226–331.

Bamakan, S. M. H., Amiri, B., Mirzabagheri, M., & Shi, Y. (2015). A new intrusion detection approach using PSO based multiple criteria linear programming. *Procedia Computer Science*, 55, 231–237. doi:10.1016/j.procs.2015.07.040

Benfatto, M. N., Seimyr, G. Ö., Ygge, J., Pansell, T., Rydberg, A., & Jacobson, C. (2016). Screening for dyslexia using eye-tracking during reading. *PLoS One*, 11(12), e0165508. doi:10.1371/journal.pone.0165508 PMID:27936148

Bernard, S., Heutte, L., & Adam, S. (2010, August). A study of strength and correlation in random forests. In *International Conference on Intelligent Computing* (pp. 186-191). Springer. doi:10.1007/978-3-642-14831-6_25

Blythe, H. I., Kirkby, J. A., & Livesedge, S. P. (2018). Comments on: “What Is Developmental Dyslexia?” *Brain Sci.* 2018, 8, 26. The Relationship between Eye Movements and Reading Difficulties. *Brain Sciences*, 8(6), 100. doi:10.3390/brainsci8060100 PMID:29867069

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2019). A review of eye-tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, 4(1), 1–16. doi:10.1186/s41235-019-0159-2 PMID:30796618

Charoen, P., Kongprawechnon, W., & Tungpimolrut, K. (2017). A hybrid particle swarm optimization-SVM classification for automatic cardiac auscultation. *Songklanakarin Journal of Science and Technology*, 39(2).

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). doi:10.1145/2939672.2939785

Cook, A. E., & Wei, W. (2019). What Can Eye Movements Tell Us about Higher Level Comprehension? *Vision (Basel)*, 3(3), 45. doi:10.3390/vision3030045 PMID:31735846

Dongoran, A., Rahmadani, S., Zarlis, M., & Zakarias, . (2018, March). Feature weighting using particle swarm optimization for learning vector quantization classifier. *Journal of Physics: Conference Series*, 978(1), 012032. doi:10.1088/1742-6596/978/1/012032

Frid, A., & Breznitz, Z. (2012, November). An SVM based algorithm for analysis and discrimination of dyslexic readers from regular readers using ERPs. In 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel (pp. 1-4). IEEE. doi:10.1109/EEEI.2012.6377068

Hessel, A., Nation, K., & Murphy, V. (2019). *Comprehension monitoring during reading: an eye-tracking study with children learning English as an additional language*. Academic Press.

Himmelstoss, N. A., Schuster, S., Hutzler, F., Moran, R., & Hawelka, S. (2019). Co-registration of eye movements and neuroimaging for studying contextual predictions in natural reading. *Language, Cognition and Neuroscience*, 1–18. doi:10.1080/23273798.2019.1616102 PMID:32656295

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye-tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017). *Mobilenets: Efficient convolutional neural networks for mobile vision applications*. arXiv preprint arXiv:1704.04861.
Huang, M. L., Hung, Y. H., Lee, W. M., Li, R. K., & Jiang, B. R. (2014). SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. *The Scientific World Journal, 2014*, 2014. doi:10.1155/2014/795624 PMID:25295306

Iwabuchi, M., Hirabayashi, R., Nakamura, K., & Dim, N. K. (2017). Machine Learning-Based Evaluation of Reading and Writing Difficulties. *Studies in Health Technology and Informatics, 242*, 1001–1004. PMID:28873918

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*. doi:10.1080/03772063.2019.1622461

Jothi Prabha, A., Bhargavi, R., & Ragala, R. (2019). Prediction of dyslexia using support vector machine in distributed environment. *IACSIT International Journal of Engineering and Technology, 2795–2799.*

Karande, S., & Agarwal, A. (2017). Ophthalmic abnormalities in children with dyslexia: A look at current research. *Journal of Postgraduate Medicine, 63*(1), 1. doi:10.4103/0022-3859.198138 PMID:28079041

Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In *Proceedings of ICNN’95-International Conference on Neural Networks* (Vol. 4, pp. 1942-1948). IEEE. doi:10.1109/ICNN.1995.488968

Kurokami, T., Koeda, T., Migita, O., & Hata, K. (2019). Reading disability due to an ocular motor disorder: A case of an adolescent girl with a previous diagnosis of dyslexia. *Brain & Development, 41*(2), 187–190. doi:10.1016/j.braindev.2018.09.003 PMID:30266219

Li, X., Wang, L., & Sung, E. (2008). AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence, 21*(5), 785–795. doi:10.1016/j.engappai.2007.07.001

Lin, X., Li, C., Zhang, Y., Su, B., Fan, M., & Wei, H. (2018). Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. *Molecules (Basel, Switzerland), 23*(1), 52. doi:10.3390/molecules23010052 PMID:29278382

Mo, Y., & Xu, S. (2010, June). Application of SVM based on hybrid kernel function in heart disease diagnoses. In *2010 International conference on intelligent computing and cognitive informatics* (pp. 462-465). IEEE. doi:10.1109/ICICCI.2010.96

Müller-Axt, C., Anwander, A., & von Kriegstein, K. (2017). Altered structural connectivity of the left visual thalamus in developmental dyslexia. *Current Biology, 27*(23), 3692–3698. doi:10.1016/j.cub.2017.10.034 PMID:29153326

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eye-tracking data. *Behavior Research Methods, 42*(1), 188–204. doi:10.3758/BRM.42.1.188 PMID:20160299

Prabha, A. J., & Bhargavi, R. (2019). Prediction of Dyslexia Using Machine Learning—A Research Travelogue. In *Proceedings of the Third International Conference on Microelectronics, Computing and Communication Systems* (pp. 23-34). Springer.

Qiu, S., & Lane, T. (2008). A framework for multiple kernel support vector regression and its applications to siRNA efficacy prediction. *IEEE/ACM Transactions on Computational Biology and BioInformatics, 6*(2), 190–199. PMID:19407344

Raiko, T., Ilin, A., & Karhunen, J. (2007, November). Principal component analysis for sparse high-dimensional data. In *International Conference on Neural Information Processing* (pp. 566-575). Springer.

Rello, L., & Ballesteros, M. (2015, May). Detecting readers with dyslexia using machine learning with eye-tracking measures. In *Proceedings of the 12th Web for All Conference* (pp. 1-8). doi:10.1145/2745555.2746644

Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye-tracking research & applications* (pp. 71-78). ACM. doi:10.1145/355017.355028

Schindler, M., & Lilenthal, A. J. (2019). Domain-specific interpretation of eye-tracking data: Towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics, 101*(1), 123–139. doi:10.1007/s10649-019-9878-z
Jothi Prabha Appadurai obtained her Bachelor's degree in Information Technology from Bharathidasan University in the year 2002 and Master's degree in Software Engineering from Jawaharlal Nehru Technological University in the year 2012. She has 6 years of Industry experience as a Software Engineer at Bank of America Continuum solutions, Hyderabad and 9 years of Academic experience as Associate Professor in CSE Department at Jyothishmathi Institute of Technology and Science. Currently, she is a full time Research scholar at School of Computing Science and Engineering, VIT University, Chennai Campus, India. Her current research interests are Machine Learning and Data Analytics.

Bhargavi Rentachintala is presently working as Associate Professor in the School of Computing Science and Engineering, VIT University, Chennai Campus, India. She has more than 20 years of Industry, Academic and Research experience. She received her M.Tech and Ph.D degrees from IIT Madras and Anna University respectively. Her research interests include Complex Event Processing, Machine learning, in Healthcare and Data Science. She has authored and published several research papers in IEEE/ACM/Springer international conferences and refereed Journals. She also authored chapters in highly reputed research reference books.

Stella, M., & Engelhardt, P. E. (2019). Syntactic ambiguity resolution in dyslexia: An examination of cognitive factors underlying eye movement differences and comprehension failures. *Dyslexia (Chichester, England)*, 25(2), 115–141. doi:10.1002/dys.1613 PMID:30990960

Tamboer, V., Vorst, H. C. M., Ghebreab, S., & Scholte, H. S. (2016). Machine learning and dyslexia: Classification of individual structural neuro-imaging scans of students with and without dyslexia. *NeuroImage. Clinical, 11*, 508–514. doi:10.1016/j.nicl.2016.03.014 PMID:27114899

Tsentscher, N., Ruisinger, A., Blank, H., Díaz, B., & von Kriegstein, K. (2019). Reduced structural connectivity between left auditory thalamus and the motion-sensitive planum temporale in developmental dyslexia. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 39*(9), 1720–1732. doi:10.1523/JNEUROSCI.1435-18.2018 PMID:30643025

Wang, D., Zhang, Y., & Zhao, Y. (2017, October). LightGBM: an effective miRNA classification method in breast cancer patients. In *Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics* (pp. 7-11). doi:10.1145/3155077.3155079

Wang, X., Zhao, X., & Ren, J. (2019). A new type of eye movement model based on recurrent neural networks for simulating the gaze behavior of human reading. *Complexity, 2019*, 2019. doi:10.1155/2019/8641074

Werth, R. (2019). What causes dyslexia? Identifying the causes of dyslexia and effective compensatory therapy. *Restorative Neurology and Neuroscience*, (Preprint), 1-18.

Wilcockson, T. D., Mardanbegi, D., Sawyer, P., Gellersen, H., Xia, B., & Crawford, T. J. (2019). Oculomotor and inhibitory control in dyslexia. *Frontiers in Systems Neuroscience, 12*, 66. doi:10.3389/fnsys.2018.00066 PMID:30687026

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam-based eye tracking. arXiv preprint arXiv:1504.06755.