INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

**Received:** 23-01-2023

**Accepted:** 28-07-2023

**Published:** 08-09-2023

**Citation:** Vanitha G, Kasthuri M (2023) Performance Analysis of Feature Selection and Classification Methods for Predicting Dyslexia. Indian Journal of Science and Technology 16(33): 2663-2669. https://doi.org/10.17485/IJST/v16i33.165

* **Corresponding author**.

vanithaguna11@gmail.com

**Funding:** None

**Competing Interests:** None

Published By Indian Society for Education and Environment (iSee)

**ISSN**
Print: 0974-6846
Electronic: 0974-5645

# Performance Analysis of Feature Selection and Classification Methods for Predicting Dyslexia

**G Vanitha**[1]*, **M Kasthuri**[2]

**1** Research Scholar/Associate Professor, Department of Information Technology, Bishop Heber College, Affiliated Bharathidasan University, Tiruchirappalli, 620 024, Tamil Nadu, India
**2** Associate Professor, Department of Computer Applications, Bishop Heber College, Affiliated Bharathidasan University, Tiruchirappalli, 620 024, Tamil Nadu, India

## Abstract

**Objectives:** This study aims to select efficient and relevant features to detect Dyslexia with better accuracy using various Machine Learning (ML) models. **Methods:** A benchmark online gamified test dataset was used. Dyslexia from Kaggle used which contains 196 features. The dataset is divided as training and testing with 80-20%. Information Gain (IG), Principal Components Analysis (PCA), and Correlation Attribute Evaluation (CAE) are used to select relevant features. The performances of the selected features are evaluated using ML Classifiers models such as C 4.5, Random Forest (RF), Decision Table (DT), Logistic Regression (LR), and Support Vector Machine (SVM). **Findings:** Our feature selection method IG selects 192, PCA selects 195, and CAE selects 186 features out of 196 features. The selected features are tested with various above-mentioned ML classifier models. This study shows CAE with the LR classifier model well suited for select relevant features with 89.8% of accuracy. **Novelty:** This study presents a CAE feature selection approach with LR classifier approximately greater than 1.5 % accuracy of the existing approach of MIG, K-Best Features, and Recursive Feature Elimination in Random Forest. The proposed technique achieved improvement in accuracy.

**Keywords:** Machine Learning; Feature selection; Classification; Dyslexia

## 1 Introduction

According to the International Dyslexia Association, around the globe, it is estimated that 1 out of 10 people have dyslexia [1]. There are approximately 780 million people with dyslexia worldwide [2]. In India; research indicates that 10 to 15% of children may suffer some type of dyslexia [3]. Healthcare data of Dyslexia patients in the form of electronic medical records such as magnetic resonance imaging (MRI), electroencephalography (EEG), computed tomography (CT) scans, eye movements, and scans that are obtained from the patients [4]. Machine learning methods have been successfully implemented in the detection of dyslexia from eye movements, with promising results [5,6].

Identifying dyslexia in Chinese children using character dictation has limitations; they use manual coding of word dictation errors, which is time-consuming and requires more human involvement. To improve the efficiency and precision of prediction, they need to use handwriting features[7]. In[8] they suggested various techniques for dyslexia detection and research gaps that mainly focus on machine learning. They are classified as applications covering different machine learning-based approaches, image processing techniques, game-based approaches, and different assessment and assistive tools to identify dyslexic people. Machine Learning Algorithm Usage: Existing literature uses SVM at 29.2%, followed by Naive Bayes and K-nearest neighbor at 12.5% each. Mostly, feature selection techniques are used to predict dyslexia. The principal component analysis, SIFT (Scale-Invariant Feature Transform), Discrete Wavelet Transform, and statistical approaches such as Pandas, LASSO, and RFE-SVM (Recursive Feature Elimination) are used. When possible improvements or research gaps are identified, the dataset needs to be increased for improved accuracy. Language and native country of screening tools only a few language data sets are available, like Hebrew, Spanish, Malay, Chinese, English, etc. The most important thing about this survey is that it was done in India. Research has been done on the existing datasets from other countries. No such dataset has been designed by native Indians.

In[9] DysLexML is proposed as a screening tool for dyslexia that uses various ML classifiers, such as K-means, SVM, and Naive Bayes, and reasonably estimates their performance using data collected in the field study of RADAR (Rapid Assessment of Difficulties and Abnormalities in Reading). This approach gave the authors of DysLexML general (non-word-specific) features of 35 and word-specific. They evaluated their tool using LASSO regression with five-fold cross-validation to identify the dominant features. Evaluation based on RADAR uses Leave One Out Cross Validation (LOOCV), an appropriate choice given the relatively small size of the dataset. Comparatively evaluating DysLexML, with SVM and LASSO ($\lambda$1SE), it outperforms RADAR: 97.10% vs. 94.2% for the baseline text. For the easy text, RADAR reports an 87.9% correct classification, while DysLexML, with K-means with k equal to 2, exhibits an accuracy of 89.39%.

In[10], they proposed a data-driven classification of Dyslexia using machine learning. The authors used an eye-tracking data set collected from the natural reading of 48 young adults. In this approach, a set of 67 features containing saccade, glissade, fixation-related measures, and the reading speed were used. They extracted the features using MATLAB R2011b. To detect participants with dyslexic reading patterns, they used the ML method of a linear support vector machine. They used a recursive feature elimination method for feature selection, and they also measured hyperparameter optimization, with regular and nested cross-validation. They evaluated their models using 10-fold cross-validation and compared their results; the overall best model achieved 90.1% classification accuracy, while the best nested model achieved 75.75% accuracy.

In[11] proposed detection of dyslexia with machine learning, Random Forest (RF) is used to select the most important eye movement features to be used as input to a Support Vector Machine classifier. The eSeek dataset is collected from the Department of Psychology at the University of Jyvaskyla. They evaluated their model using five-fold cross-validation[12]. In order to conserve computational time and increase the number of unpredicted dyslectic cases in each fold. In the case of RF, the algorithm also calculates the feature importance for each model created in the cross-validation folds. They used a hybrid method that was capable of reliably identifying dyslexic readers and also provided insight into the data used. The SVM classifier using the most relevant eye movement features selected using RF achieved an accuracy of 89.7% and a recall score of 84.8%.

In[13] proposed treatment plans were proposed for improving dyslexic children's cognitive skills using electroencephalogram (EEG) patterns. They used a treatment consisting of Transcranial Direct Current Stimulation (tDCS) and occupational therapy using the Brainwave SAFARI software. They collected data from the EEG signals of 16 dyslexic children recorded during the eyes-closed resting state before and after treatment. An optimal subset of features extracted from recorded EEG signals was determined using principal Component Analysis (PCA) in conjunction with the Sequential Floating Forward Selection (SFFS) algorithm. The most discriminative subset of features could classify the data with an accuracy of 92% with an SVM classifier.

The main aim of feature selection is to remove irrelevant or redundant features and get a meaningful dataset. Feature selection gives an efficient way to solve the problem of removing irrelevant and redundant data, and it can facilitate a better understanding of the learning model or data, reduce computation time, and improve learning accuracy. In this study, we discuss supervised, unsupervised, and semi-supervised feature selection methods. The techniques for feature selection are broadly classified into supervised techniques and unsupervised techniques. Supervised techniques can be used for labeled data, whereas unsupervised techniques can be used for unlabeled data. Supervised methods may be divided into filter methods, wrapper methods, and embedded methods.

In this paper, we will analyze the performance of various feature selection methods for predicting that dyslexia is a learning disability. Data collected from an online gamified test is used to analyze and predict the risk of dyslexia. The dataset used for this work is taken from the online Kaggle repository[14,15].

Key contributions to this research are listed as follows:

- This paper demonstrates Information Gain, Principal Component Analysis, Correlation Attribute Evaluation feature selection techniques with different classifier models for dyslexia datasets.

- This paper proposes the best feature selection and classifier model for the datasets used.

The remaining sections of the paper are organized as follows: Section 2 focuses on the proposed methodology. Section 3 discusses the results obtained from the proposed work. Section 4 draws attention to the conclusion of the present study.

## 2 Methodology

In this section, we provide a detailed overview of the methodology and the proposed work for efficiently predicting the risk of dyslexia using feature selection methods and classification algorithms. The proposed methodology involves various steps; the preprocessed data is applied for various feature selection methods such as IG, PCA and CAE to select the relevant features to identify the disorder. Next step is to divide the dataset into training and testing. We use several machine learning algorithms such as C4.5, Random Forest, Decision Table, Logistic Regression and SVM. This approach not only predict the risk of dyslexia but also to identify children managing strategies and learning methods more accurately and efficiently. The trained model is used to predict dyslexia for each and every child, and the Accuracy of the model is evaluated using various performance metrics such as Precision, Recall, F -measure and ROC area. An overview of the suggested framework for dyslexia prediction is shown in Figure 1 [16]

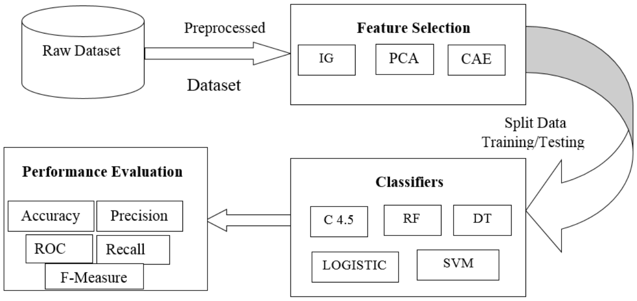


**Fig 1.** Overview of the proposed model for dyslexia prediction

### 2.1 Feature Selection

#### 2.1.1 Information Gain

Relevant features are selected based on feature selection. In this research, Information Gain (IG), Principal Component Analysis and Correlation Attribute Evaluation methods were used [17]. Different number of principal components to decrease the computational complexity [18] . Information gain is calculated using equation 1.

$$Information\ Gain\ (T,A) = 1 - E(V) \tag{1}$$

Where T- Target

A- Column (Attribute)

V- Each value in A

Where E(V) can be calculated using equation 2.

$$E(V) = \sum_{I=1}^{n} -pi\ \ log2\ \ pi \tag{2}$$

#### 2.1.2 Principal Component Analysis

Principal Component Analysis is to reduce the dimensionality of a large dataset. First step to compute the variance, co -variance, Eigenvectors and Eigenvalues.

$$Standard\ Deviation^2 = \sum_{i=1}^{n} \frac{(Xi - X)^2}{(n-1)} \tag{3}$$

$$Co - Variance\ COV\ (A,B) = \sum_{i=1}^{n} \frac{\left(Ai - \bar{A}\right)\left(Bi - \bar{B}\right)}{(n-1)} \tag{4}$$

Vectors ‘Z’ having same direction as AZ are called Eigen vectors of Z.

$$AZ = \lambda Z \tag{5}$$

Where Eigen value $\lambda$ can be calculated using equation 5. $\lambda$ is called Eigenvalue of A. Eigen values measure the variation explained by each PC. Eigenvectors weights to compute the uncorrelated PC. We obtained the new feature set after the removal of less or unimportant features from the dataset with the help of Principal Component Analysis.

### *2.1.3 Correlation Attribute Evaluation*

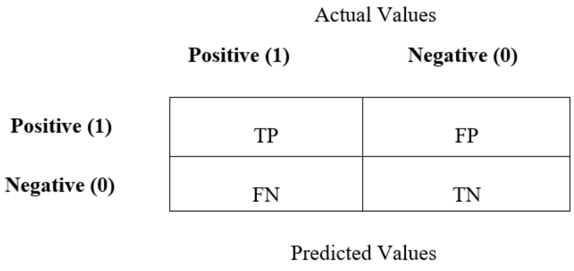
CAE evaluates the attributes with respect to the target variable. Pearson’s correlation method is used to measure the correlation between each attribute and target class attribute.

$$correlation\ (X,Y) = \frac{COV\ (X,Y)}{Standard\ Deviation\,(X,Y)} \tag{6}$$

After preprocessed and selected attributes are given,(18) the classifier models such as Random Forest (RF), C4.5, Decision Table (DT), Logistic, Naïve Bayes, and Support Vector Machine (SVM) are discussed. These classifiers are evaluated and produce different performance measures.

### *2.1.4 Performance measures*

The confusion matrix is a performance measure for Machine Learning classification problems to predict the values. The output can be two or more classes. Four different combinations of actual and predicted values of the confusion matrix are shown in Figure 2 .



| | Actual Values | |
|---|---|---|
| | **Positive (1)** | **Negative (0)** |
| **Positive (1)** | TP | FP |
| **Negative (0)** | FN | TN |

Predicted Values

**Fig 2.** Confusion matrix

It is really useful for calculating Accuracy, Recall, Precision, Specificity and very importantly ROC curves.

The evaluation metrics can be explained below: predictive accuracy is the proportion of correctly classified outcomes either true positive or true negative.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{7}$$

Precision is the percentage of positively classified outcomes that are correctly classified.

$$Precision = \frac{(TP)}{(TP+FP)} \tag{8}$$

Recall is the percentage of positive cases that are correctly classified.

$$Recall = \frac{TP}{(TP+FN)} \tag{9}$$

F-Measure is for all positive data sets containing at least one pair of non-equal values, that is the harmonic mean of precision and recall.

$$F_{Measure} = \frac{(2 \times Preciosn \times Recall)}{(Precison + Recall)} \tag{10}$$

A ROC curve shows the performance of a classification model at all classification thresholds. The ROC area has

The x-axis is 1 – specificity

$$False\ Positive\ Fraction\ = \frac{FP}{(FP+TN)} \tag{11}$$

The y-axis is sensitivity

$$True\ Positive\ Fraction = \frac{TP}{(TP+FN)} \tag{12}$$

## 3 Results and Discussion

The dataset that is used in this study is ”Predicting Risk of Dyslexia” from the Kaggle repository. This work was approved by the Carnegie Mellon University Institutional Review Board and was conducted by the authors, Luz Rello et al. The aim of this study was to understand whether the risk of dyslexia can be predicted or not by administering questions through an online game that is very easy to implement. Hence, this dataset was considered for the proposed work. This dataset, consisting of a total of 3644 participants, was involved in this study between the ages of 7 and 17 years. From this group, 392 were dyslexic, and the remaining was in the control group. All the tests were conducted in Spanish.

Four features related to demography were collected from the participants. They are Gender, Native Lang (to understand whether their native language is Spanish or not), language subject (a binary value of yes or no to understand whether the participant has failed any language subject in school), and age. Apart from the four demographic features, answers related to 32 questions were collected from the participants through games. Question numbers 1–21 had questions related to auditory and visual discrimination. Question numbers 22–29 focused on correcting words and sentences. Question nos. 30–32 was used for checking sequential visual and auditory working memory. For each of the questions 1–32, six responses were recorded. There are no clicks, a number of correct answers (hits), the sum of hits (score), accuracy (calculated as the number of hits/number of clicks), and miss rate (calculated as the number of misses/number of clicks). So, a total of 4 demography-related features plus 32 questions x 6 responses for each question = 192. The total number of features was 196. There were 3644 participants. In this dataset, there are no missing values. The output variable is a column called Dyslexia," which has a categorical value of Yes or No, indicating the presence or absence of dyslexia.

We used five different supervised Machine Learning algorithms, all of which are available and implementable with WEKA. They are C4.5, Random Forest, Decision Table, Logistic Regression, and SVM. We applied the five Machine Learning algorithms to each of the datasets using 10-fold cross-validations to evaluate their performance. The measures we used to evaluate the performance of each model were predictive Accuracy, Precision, Recall, F-Measure and ROC (Receiver Operating Characteristic) area. These measures are based on the confusion matrix, which is a 2x2 matrix comparing the model's predicted class values to the actual class values. In the first quadrant, we have true positives (TP), which is the number of students with dyslexia who are correctly classified. Next, we have false positives (FP), or the students without dyslexia who were incorrectly classified as having dyslexia. Following this are false negatives (FN), or students who have dyslexia but are not classified correctly by the model. Finally, true negatives (TN), which are patients without dyslexia that are correctly classified,

The dyslexia dataset was transferred from the Kaggle repository. In the previous study, features were selected by using feature mapping through PCA and feature normalization. In this research, Information Gain, Principal Component Analysis and Correlation Attribute Evaluation feature selection techniques were used. The selected features are compared with different classifier models such as RF, Decision Table, SVM, Logistic, etc. Table 1 shows the selected attributes through various feature selection techniques.

**Table 1.** Selected Attributes

| **Feature Selection Algorithms** | **Total number of features: 196** |
|---|---|
| | **Number of features selected** |
| Information Gain | 192 |
| Principal Component Analysis | 195 |
| Correlation Attribute Evaluation | 186 |

An analysis of the existing research in this area reveals that the few existing prior ML studies use the least useful set of data useful to reading with reference to only fixations and saccades of eye movements[19].

**Table 2.** Performance Metrics of Feature selection Methods and Machine Learning Algorithms

| Classifiers | Performance Measures in % | | | | |
|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | F-Measure | ROC area |
| C 4.5 | 86.8 | 37.6 | 34.2 | 35.8 | 54.7 |
| Random Forest | 89.5 | 81.0 | 04.3 | 08.2 | 86.2 |
| Decision Table | 89.2 | 51.6 | 08.4 | 14.5 | 74.7 |
| **Logistic** | **89.8** | 54.2 | 34.7 | 42.3 | 84.2 |
| SVM | 89.4 | 88.9 | 02.0 | 04.0 | 51.0 |

Table 2 shows the various performance measures of feature selection methods in data mining algorithms. The most accurate method was built using Logistic, which achieved an accuracy of 89.8%, precision of 54.2%, recall of 34.7%, F-measure of 42.3%, and ROC area of 84.2% compared with other methods.

In this research paper, several Machine Learning (ML) algorithms were performed with the aim of predicting the risk of dyslexia. This study included a total of 3644 participants, ranging in age from 7 to 17 years. Only 392 of the total 3644 were dyslexics, with the rest being control subjects. In the research, all the tests were conducted in the Spanish language. Four features related to the demography were collected from the participants. ML has been applied to many fields, including identifying and diagnosing it, medical imaging diagnosis, health, and neuroscience. Research on the neurology of eye movements and EEG signals, functional MRI scans has been a revolutionary field of cognitive neuroscience.

The existing work[(20)] using various classifiers such as Gaussian Naive Bayes, AdaBoost, and the Ridge shows an accuracy of 89.5% with the regressor feature selection method for the prediction of hypothyroid disease.[(21)] Shows 89% accuracy with Mutual Information Gain (MIG), K-Best Features, and Recursive Feature Elimination (RFE) with a voting ensemble approach classifier and[(22)] shows 89.7% accuracy with an ensemble classifier model such as Random Forest (RF) or the SVM model for dyslexia detection[(23)] . The accuracy of J48 is 80%, compared to RF's 76% and SVM's 88%, for predicting autism spectrum disorder. In our work, the accuracy of benchmark datasets obtained through FS methods is evaluated using C4.5, RF, Decision Table, Logistics, and SVM classifier models. The FS results gained and accomplished in WEKA are shown in Table 3.

## 4 Conclusion

This study focuses on selected features from different filter and wrapper methods individually. In this study, feature selection is used to select related features for predicting dyslexia. Among the three-feature selection methods, CAE selects186 features out of 196, IG selects 192 features out of 196 and PCA selects 195 features out of 196. The IG and PCA drops important features, the obtained features which are not useful for further classification. It is concluded that IG, PCA, and Correlation Attribute Evaluation feature selection methods select optimal features such that 186 features from 196 for dyslexia give 89.8% accuracy. Also, a single method does not give better results, but a collective approach can help improve efficiency and accuracy. In future, one or more filter and wrapper methods are combined to produce better accuracy in the classification process.

## References

1) Dyslexia at a glance . 2014. Available from: https://dyslexiaida.org/dyslexia-at-a-glance/.
2) Dyslexia association of India . 2023. Available from: https://www.dyslexiaindia.org.in/what-dyslexia2.html/.
3) Zauderer S. Dyslexia statistics & facts: How many people have Dyslexia? *Crossrivertherapycom Cross River Therapy*. 2023. Available from: https://www.crossrivertherapy.com/research/dyslexia-statistics.
4) Hmimdi AEE, Ward LM, Palpanas T, Kapoula Z. Predicting Dyslexia and Reading Speed in Adolescents from Eye Movements in Reading and Non-Reading Tasks: A Machine Learning Approach. *Brain Sciences*. 2021;11(10):1337. Available from: http://dx.doi.org/10.3390/brainsci11101337.
5) Alqahtani ND, Alzahrani B, Ramzan MS. Deep Learning Applications for Dyslexia Prediction. *Applied Sciences*. 2023;13(5):2804. Available from: http://dx.doi.org/10.3390/app13052804.
6) Prabha J, Bhargavi A. Prediction of dyslexia from eye movements using machine learning. *IETE Journal of Research*. 2022;68(2):814–837. Available from: http://dx.doi.org/10.1080/03772063.2019.1622461.
7) Man K, Lee S, Liu HW, Tong SX. Identifying Chinese children with dyslexia using machine learning with character dictation. *Scientific Studies of Reading*. 2023;27(1):82–100. Available from: http://dx.doi.org/10.1080/10888438.2022.2088373.
8) Jan TG, Khan SM. A Systematic Review of Research Dimensions Towards Dyslexia Screening Using Machine Learning. *Journal of The Institution of Engineers (India): Series B*. 2023;104(2):511–522. Available from: http://dx.doi.org/10.1007/s40031-023-00853-8.
9) Asvestopoulou T, Manousaki V, Psistakis A, Smyrnakis I, Andreadakis V, Aslanides IM. Screening tool for dyslexia using machine learning. *ARXIV*. 2019. Available from: http://dx.doi.org/10.48550/ARXIV.1903.06274.
10) Szalma J, Weiss B. Data-Driven Classification of Dyslexia Using Eye-Movement Correlates of Natural Reading. In: ACM Symposium on Eye Tracking Research and Applications;vol. 2020. ACM. 2020. Available from: https://doi.org/10.1145/3379156.3391379.

11) Raatikainen P, Hautala J, Loberg O, Kärkkäinen T, Leppänen P, Nieminen P. Detection of developmental dyslexia with machine learning using eye movement data. *Array*. 2021;12:100087. Available from: http://dx.doi.org/10.1016/j.array.2021.100087.

12) Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*. 2020;143:106839. Available from: https://doi.org/10.1016/j.csda.2019.106839.

13) Oliaee A, Mohebbi M, Shirani S, Rostami R. Extraction of discriminative features from EEG signals of dyslexic children; before and after the treatment. *Cognitive Neurodynamics*. 2022;16(6):1249–1259. Available from: http://dx.doi.org/10.1007/s11571-022-09794-2.

14) Rello L, Baeza-Yates R, Ali A, Bigham JP, Serra MP. Predicting risk of dyslexia with an online gamified test. *PLOS ONE*. 2020;15(12):e0241687. Available from: http://dx.doi.org/10.1371/journal.pone.0241687.

15) Dataset. Kaggle Repository. . Available from: https://doi.org/10.34740/kaggle/dsv/1617514.

16) Alyasiri OM, Cheah NN, Abasi AK, Al-Janabi OM. Wrapper and Hybrid Feature Selection Methods Using Metaheuristic Algorithms for English Text Classification: A Systematic Review. *IEEE Access*. 2022;10:39833–39852. Available from: http://dx.doi.org/10.1109/access.2022.3165814.

17) Shankar S, Ashokkumar G, Vinayakumar P, Ghosh R, Mansoor U, S WW. An embedded-based weighted feature selection algorithm for classifying web document. *Wireless Communications and Mobile Computing*. 2020;p. 1–10. Available from: https://doi.org/10.1155/2020/8879054.

18) N A, MB R, HM EH, and Rashid M AI. An efficient machine learning-based feature optimization model for the detection of dyslexia. 2023. Available from: http://dx.doi.org/10.1155/2022/8491753.

19) Vani Chakraborty, Sundaram M. Machine learning algorithms for prediction of dyslexia using eye movement. *Journal of Physics: Conference Series*. 2020;1427(1):012012. Available from: https://doi.org/10.1088/1742-6596/1427/1/012012.

20) Devi MS, Kumar VD, Brezulianu A, Geman O, Arif M. A Novel Blunge Calibration Intelligent Feature Classification Model for the Prediction of Hypothyroid Disease. PMCID. 2023. Available from: https://doi.org/10.3390/s23031128.

21) Jan TG, Khan SM. An effective feature selection and classification technique based on ensemble learning for dyslexia detection. 2023. Available from: https://doi.org/10.1007/978-981-19-1844-5_32.

22) Raatikainen P, Hautala J, Loberg O, Kärkkäinen T, Leppänen P, Nieminen P. Detection of developmental dyslexia with machine learning using eye movement data. *Array*. 2021;12:100087. Available from: http://dx.doi.org/10.1016/j.array.2021.100087.

23) Radzi SFM, Hassan MS, Radzi MAHM. Comparison of classification algorithms for predicting autistic spectrum disorder using WEKA modeler. *BMC Medical Informatics and Decision Making*. 2022;22(1):306. Available from: http://dx.doi.org/10.1186/s12911-022-02050-x.