IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 3, September 2024, pp. 3147~3156
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i3.pp3147-3156

An ensemble features aware machine learning model for detection and staging of dyslexia

Sailaja Mulakaluri[1,2], Girisha Gowdra Shivappa[2]
[1]Department of Computer Science, St. Francis De Sales College, Bengaluru, India
[2]Department of Computer Science and Engineering, Dayananda Sagar University, Bengaluru, India

Article Info

Article history:

Received Nov 23, 2023
Revised Feb 21, 2024
Accepted Mar 13, 2024

Keywords:

Deep learning
Dyslexia
Data imbalance
Feature importance
Learning disorder
Machine learning

ABSTRACT

Dyslexia is a specific learning disorder (SLD) which may affect young child's cognitive skills, text comprehension, reading-writing and also problem-solving abilities. To diagnose and identify dyslexia, the testing scale tool has been proposed using artificial intelligence technique. The proposed tool allows the student who is suspected to have dyslexia to take up quiz and perform certain task based on the type of learning impairments. After completion of the test, resultant data is provided as input to the proposed ensemble feature aware machine-learning (EFAM) XGBoost (XGB) model. Based on the student assessment score and time taken by children, the EFAM-XGB algorithm predicts dyslexia. The proposed EFAM-XGB is used to develop an integrated and user-friendly tool that is highly accurate in identifying reading disorders even with presence of realistic imbalanced dataset and suggest the most appropriate instructional activities to parents and teachers. The EFAM-XGB-based dyslexia detection method achieves very good accuracy of 98.7% for dyslexia dataset; thus, attain better performance in comparison with existing machine learning (ML)-based methodologies.

Corresponding Author:

Sailaja Mulakaluri
Department of Computer Science and Engineering, Dayananda Sagar University
Bengaluru, India
Email: sailaja.mulakaluri07@gmail.com

1. INTRODUCTION

Dyslexia, a specific learning disorder (SLD), is a condition characterized by neurobiological factors that impact individuals worldwide, affecting approximately 5-15% of the overall worldwide population [1]. Individuals diagnosed with dyslexia experience challenges in the areas of writing and reading, which are not influenced by factors such as intelligence, native language, socioeconomic status, or educational background. Moreover, individuals who possess knowledge of their dyslexia diagnosis have the potential to acquire and implement various coping strategies aimed at mitigating the adverse impacts associated with this condition [2], [3]. Nevertheless, it has been observed that individuals diagnosed with dyslexia tend to experience academic challenges if they do not receive adequate assistance. According to recent data, a significant proportion of individuals, specifically 35%, discontinue their education prematurely. Furthermore, it has been projected that just a small fraction, just over two percent, of individuals diagnosed with dyslexia successfully attain an undergraduate degree [4].

Identifying dyslexia poses a significant challenge, particularly in the context of Indian native languages characterized by transparent orthographies. In languages characterized by shallow orthographies, the relationship between graphemes (letters) and phonemes (sounds) tends to exhibit a higher level of consistency compared to spoken languages using deep orthographies, like English. Consequently, individuals with dyslexia encounter greater difficulties in acquiring reading skills within the context of English [5], [6].

Due to the difficulties in diagnosing dyslexia in languages with clear orthographies and the less serious nature of its symptoms, dyslexia is therefore referred to as a "hidden disability" [6]. The present diagnostic and screening procedures necessitate the involvement of trained individuals who administer an extensive in-person assessment [7], [8]. This assessment entails the measurement of various performance indicators associated with writing and reading abilities, such as speed of reading (expressed in words per minute), reading mistakes, writing mistakes, reading vocabulary, pseudo-word reading, linguistic fluency, and comprehension of texts.

Machine learning (ML) techniques have gained significant attention and application in the field of dyslexia. These algorithms [9], [10] play a vital role in identifying, predicting, and intervening in dyslexia by analyzing patterns, identifying relevant features, and making data-driven predictions. ML is utilized in various ways within the context of dyslexia, such as diagnosing and screening individuals based on cognitive assessments and educational records [11], [12]. Aditionally, it contributes to the development of assistive technologies like speech recognition and visual processing tools that assist individuals with reading, writing, and other learning difficulties [13]. Overall, ML shows great potential in enhancing the understanding, diagnosis, intervention, and support for individuals with specific learning disabilities, aiming to improve their learning outcomes and overall quality of life.

ML techniques applied to dyslexia encounter challenges that affect their effectiveness. Two main challenges are data imbalance and feature importance [14]. Imbalanced datasets, where one class is significantly more prevalent than the other, can lead to biased models favoring the majority class and performing poorly in identifying dyslexia. Techniques like oversampling, under-sampling, or synthetic minority over-sampling technique (SMOTE) [15] address this issue. Identifying relevant features for accurate models is challenging due to the complex nature of dyslexia. Various data sources, such as cognitive assessments and behavioral observations, need careful consideration. Feature selection techniques like recursive feature elimination (RFE) [16] or permutation importance help determine crucial features. Additional challenges include the heterogeneity of dyslexia, requiring specialized models, the interpretability of models for comprehension, and limited and diverse data availability. Collaboration among stakeholders and efforts in dataset collection, feature selection, and model transparency are crucial for advancing ML applications in dyslexia and supporting individuals with learning difficulties. To solve all the issues mentioned above, this work introduces a novel ensemble feature aware machine-learning XGBoost (EFAM-XGB) mechanism that gives equal importance to both positive (i.e., correct) and negative (i.e., wrong) dyslexia prediction. Then, a novel multi-level K-fold cross validation is introduced to select effective features with presence of imbalanced data. The significance of using proposed EFAM-XGB for detecting dyslexia among young student is given as:

– The model introduced a classification methodology that can address both binary and multi-label classification problems.
– The work introduced an effective weight optimization mechanism to give ideal weight optimization process for both correct and wrongly classified labels.
– The model presents a feature selection mechanism by introducing new cross validation function under presence of imbalanced data.
– The EFAM-XGB attains a very good accuracy, precision, specificity, sensitivity, and F1-score in comparison with ML technique like decision tree (DT), support vector machine grid search (SVM-GS), random forest grid search (RF-GS), and XGBoost (XGB).
– The EFAM-XGB is very efficient in detecting dyslexia disabilities among young kids in comparison with existing methods.

The manuscript organization. In section 2, different existing methodologies pertaining to detecting dyslexia among young students using technology and artificial intelligence technique have been studied and limitations have been identified. Section 3 provides a methodology for detecting dyslexia using ensemble-based learning mechanism. Section 4 presents the result achieved using proposed dyslexia detection using EFAM-XGB. The last section discusses the significance of result improvement and scope of the proposed work.

2. LITERATURE SURVEY

This section provides survey of various existing ML and deep learning (DL) method presented for detecting SLD among young student. Kohn *et al.* have done an in-depth study on the Calcularis 2.0 [17], [18], which is used for identifying the dyscalculia SLD in the student and providing a solution to the dyscalculia SLD. After the study, they have noted that after the identification of the dyscalculia, the students are trained using the Calcularis 2.0 for twelve weeks to train the student understand the mathematic numbers and arithmetic expression. It was also noted that the Calcularis 2.0 can be helped for the students having dyslexia. The results for this work show that, if a student opts this program, then there is a probability that the student may correct his dyscalculia and dyslexia issue within three months. Babu *et al.* [19] has tried to help the guardians and parents whose children are suffering from SLD's. In this work, they have developed a web-application which helps these children to have fast-recovery. The work provided a stimulating training environment where the disability problems can be overcome by making them a regular part of daily life.

Dhamal and Mehrotra [20] main focus was to predict the SLD attaining highest accuracy. Hence in this work they have evaluated multiple ML methods such as DT, support vector machine (SVM), K-nearest neighbor (KNN), naïve Bayes (NB), logistic regression (LR), and gradient boosting (GB) for predicting the learning disability. They utilized a hospital dataset comprising of 630 individuals having sixteen attributes. The results show that the prediction accuracy for predicting the SLD is achieved more by the random forest (RF), DT, and GB. Kunhoth *et al*. [21] has presented an image dataset for the prediction of the dysgraphia. In this work, they have evaluated the image dataset using ML and DL techniques. In this study, they used handwritten picture data to diagnose dysgraphia utilizing a transfer learning process that included feature extraction as well as fine-tuning. They have also used an ensemble-learning strategy by training a set of deep convolutional neural network (CNN) classifiers that are customized for the task of recognizing handwriting. They additionally employed a feature-fusion technique, which involves the merging of elements that are unique to the task of handwriting. To properly categorize regular as well as dysgraphia handwritten images, in this work they have extracted features using multiple handwritten tasks and have generated classifier for the ML methods. They have evaluated the dataset using RF, SVM, and AdaBoost (AB). Vilasini *et al.* [22] have proposed a DL method, CNN, for the detection of the SLD. Their main focus was to help the pre and primary school students suffering from the SLD. In this work, they used the students handwriting to predict whether the student is suffering from any kind of disability. The presented work was evaluated using a vision transformer model. The results show that the CNN predicts accurately but fails as the dataset only consists of handwritten images. Hence, this model is not efficient.

Hewapathirana *et al.* [23] provided mobile application to detect the SLD in a given individual. In this work, they have used DL (CNN) and ML (RF, SVM) methods for evaluating the individual. The results show that the CNN attained 99 percent for detecting letter dysgraphia, 99 percent for lexical dyscalculia, 92 percent for verbal dyscalculia. Further, the results for ML methods show that it attained 97 percent for number dysgraphia and 98 percent for practognostic and operational dyscalculia. The results indicate that by utilizing DL and ML methods, a high accuracy for the prediction of the SLD can be achieved. Modak *et al.* [24] evaluated a learning management system (LMS) which detects the SLD students and non-SLD students [24]. This work mainly focused on the dyslexia students. In this work, they have used natural language processing and ML (SVM, LR) to analyze the student for detecting whether the student is suffering from dyslexia or not. The proposed work has been evaluated using one dataset having two classes. The results show that the LR method attains better accuracy for the prediction of the SLD (dyslexia). From the literature survey, it can be said that each research work focuses on different work. Moreover, most of the researchers have used ML [25], [26] and DL method [27], [28] for predicting the SLD. Also, most of the methods have not focused on addressing the data imbalance issue and concept drift issues in their work. Hence, to solve this, in this work we present a methodology called EFAM-XGB in the next section.

3. METHODOLOGY

Here, a methodology has been introduced for the detection of dyslexia in a group of students. Consider a dataset which consists of various students having dyslexia students and non-dyslexia students. Let the dataset be described using E. From this the overall dataset can be represented using (1).

$$E = \{(a_1, b_1), (a_2, b_2), \dots, (a_m, b_m)\} \tag{1}$$

where, a_j is used for defining all the characteristics of the dyslexia student and non-dyslexia student having n-dimensionality vector. In a_j, the j is given as $j = 1,2,3, \dots, m$ for defining the overall size of the dataset E. Furthermore, b_j is represented as $b_j \in \{-1,1\}$ for defining the output for each characteristic of a_j. The main focus of this work is to construct a detection technique $\hat{G}$ which will predict whether the student is suffering from dyslexia using a EFAM-XGB. To predict whether the student is suffering from dyslexia or not, the dyslexia-level G is defined using (2).

$$G: A \to B \tag{2}$$

3.1. Architecture

The architecture of the presented work, i.e., the detection and prediction technique $\hat{G}$ for dyslexia has been given in Figure 1. The proposed architecture is divided into 5 phases. In the first phase, preprocessing of the overall dataset is performed. In the second phase, this work detects whether the data is classified as multi-label or binary. In the third phase, the dataset is trained using the proposed EFAM-XGB. In the fourth phase, an optimization process takes place where K-fold is used if the data is imbalanced. Also, in the fourth phase, a best parameter is selected for training the imbalanced data which will help to predict the dyslexia students more accurately having minimal misclassification. In the last phase, i.e., phase 5, the EFAM-XGB verifies whether any drift exists in the detection and prediction technique $\hat{G}$. If any drift is detected, the EFAM-XGB is trained again using the same dataset. The complete process of each phase has been explained in the subsequent sections.

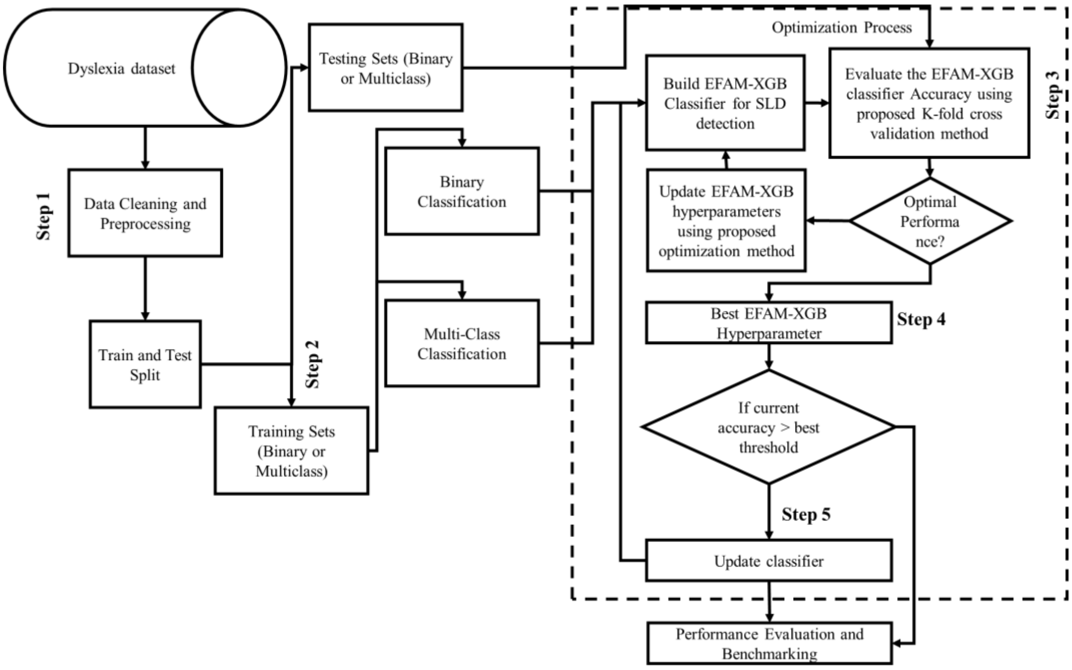

Figure 1. Proposed EFAM-XGB model for dyslexia Identification

3.2. Extreme gradient boosting ensemble prediction model

To summarize a collection of classification-rules generated using a tree-like structure (XGB) obtained through the given dataset, DT is a popular classification approach. There are three main elements that make up DT: the root-node, which represents the entire dataset; the decision-nodes, which describe splitting's as well as tests performed on every attribute; then the leaf-nodes, which describe the result of each classification. To establish better decisions, the DT algorithm repeatedly subdivides the initial training dataset into subgroups having higher characteristic values. Moreover, pruning is a technique used in DT to reduce over-fitting by removing some of the branches off of decision-nodes. Since more branches in a tree provide better information for making decisions, the highest possible tree level is a crucial hyper-parameter for controlling the computational complexities which has been considered in this work.

Many ensemble methods have already been established to improve their methods efficiency by combining many DT. Some examples of these methods are the RF, ensemble tree (ET), and extreme-XGB methods. In RF, several DT are combined into one using the bagging method. In the same way as RF constructs DTs using every sample, ET [29] randomly selects set of features to be utilized in its tree-based ensemble-learning method. Another difference between RF as well as ET is that RF improves DT splitting while ET generates splitting randomly. For improved performance and speed, many researchers turn toward the XGB method, a tree-based ensemble which utilizes gradient-descent and boosting to combine fundamental DTs [30]. The XGB is a sophisticated gradient-tree boosting-based set up capable of completely managing massive ML workloads. It has dominated Kaggle contests mainly because of its superior predicting ability and lightning-fast training time. The aim behind this approach is to build a tree by repeatedly adding nodes and separating characteristics. Every time a tree is added, the XGB method learns an entirely novel function that corresponds to the previously projected residual. Consider z_j as an input to the XGB method, z_j as true-label and a_j be the "raw prediction" before the sigmoid function, then, according to [29], [30], the objective function of the XGB model is defined as (3).

$$M^{(u)} = \sum_{j=1}^{o} m\left(z_j, A_j^{(u-1)} + g_u(y_j)\right) + \rho(g_u) + d \tag{3}$$

where, $m(.,.)$ is used for defining the loss-function, u is used for denoting the overall tree, ρ is used a penalizing function to represent the methods complexity, $\rho(g_u)$ is used for denoting the penalty-regularization function and d represents the constant. Further, the Taylor's second-order expansion is given as (4).

$$g(y + \Delta y) \approx g(y) + g'(y)\Delta y + \frac{1}{2} g''(y)\Delta y^2 \tag{4}$$

By using (3) and (4), the (5) is obtained:

$$M^{(u)} \approx \sum_{j=1}^{o} \left[m\left(z_j + A_j^{(u-1)}\right) + h_j g_u(y_j) + \frac{1}{2} i_j \left(g_u(y_j)\right)^2\right] + \rho(g_u) + d \tag{5}$$

where, h_j and i_j is evaluated by using (6) and (7):

$$h_j = \frac{\partial M}{\partial a_j} \tag{6}$$

$$i_j = \frac{\partial^2 M}{\partial a_j^2}. \tag{7}$$

Further by discarding the constant variables from the (5), the (8) is obtained which simplifies the objective-tree u.

$$M^{(u)} \approx \sum_{j=1}^{o} \left[h_j g_j(y_j) + \frac{1}{2} i_j \left(g_u(y_j)\right)^2\right] + \rho(g_u) \tag{8}$$

Moreover, the XGB method cannot be fitted without the h_j and i_j in the objective-tree. Hence, both the variables are important. For the dataset which has been classified as binary, the XGB method loss-function is defined cross entropy loss (CEL). This is defined as (9).

$$M = -\sum_{j=1}^{o}\left[z_j \log(\hat{z}_j) + (1 - z_j)\log(1 - \hat{z}_j)\right] \tag{9}$$

where, z_j is evaluated using (10).

$$\hat{z}_j = \frac{1}{[1+\exp(-a_j)]} \tag{10}$$

Hence, in the case where binary classified dataset exists, sigmoid is used as the activation function. From this, the (11) is obtained.

$$\frac{\partial \hat{z}_j}{\partial a_j} = \hat{z}_j(1 - \hat{z}_j) \tag{11}$$

3.3. Ensemble classifier performance optimization

Many classification algorithms solely concentrate on reducing the loss-function, regardless of whether or not a characteristic or scenario was correctly classified. The fundamental concept behind EFAM-XGB is to provide higher weight to samples that are positive during the training process by increasing the quantity of weight assigned for the errors generated through the samples which are positive of an incorrectly classified class within the methods loss-function. In this EFAM method, only the misclassification scenarios have been considered. Consider $O_{00} = O_{11}$, $O_{10} = b(b > 0)$, and $O_{01} = 1$, hence, using this, the loss-function having the weigh factor can be represented using (12).

$$M_b = -\sum_{j=1}^{o}\left[b z_j \log(\hat{z}_j) + (1 - z_j)\log(1 - \hat{z}_j)\right] \tag{12}$$

where, b is used for representing the optimization factor. Moreover, false negatives (FN) are more likely to incur further losses when b is above 1 and false positives (FP) are more likely to incur additional losses when b is below 1. Hence, to solve this, the first-order derivative of h_j and i_j in (6) and (7) is taken which is given using (13) and (14).

$$h_j = \frac{\partial M_b}{\partial^2 a_j} = \hat{z}_j\left(1 - z_j + bz_j\right) - bz_j \tag{13}$$

$$i_j = \frac{\partial M_b}{\partial^2 a_j^2} = \hat{z}_j\left(1 - \hat{z}_j\right)\left(1 - z_j + bz_j\right) \tag{14}$$

whenever, there exists data imbalance, the EFAM method accuracy can be impacted hence, to solve this, a novel K-fold cross-validation (K-CV) has been presented in the next section.

3.4. Feature optimization for imbalanced data

This study improves the prediction method utilized by the industry-standard XGB by modifying the feature selection method. Improving CV to produce the smallest possible validation error benefits the feature selection method. To maximize the accuracy of the prediction methods, in this work a method called K-CV has been employed, in which the dataset is arbitrarily split across K subsets having equal-sizes. Following that, K-1 are utilized for building the dyslexia predictive method, while the remaining data is utilized to maximize the accuracy of the dyslexia predictive method predictions. Finally, the CV error is optimized by taking the average of the prediction errors for each possible value of K. Then, the characteristics having the most weight are prioritized, then the dyslexia predictive method having the lowest CV error is selected, all from a grid of l suitable outcomes. In order to choose features efficiently, the suggested CV approach involves two stages. The primary characteristics are chosen from feature subsets throughout the first stage. The chosen characteristics from the first stage are then used to build an accurate dyslexia performance predictive method in the subsequent stage. The existing single K-CV error used in the existing works can be represented using (15).

$$CV(\sigma) = \frac{1}{M}\sum_{k=1}^{K}\sum_{j\in G_{-k}} P\left(b_j, \hat{g}_\sigma^{-k(j)}\left(y_j, \sigma\right)\right) \tag{15}$$

Nevertheless, the (15) fails to detect the feature which impacts the accuracy of the prediction method. Hence, to solve this, this proposed work introduces a novel CV which selects the important features having higher importance which affects the accuracy of the prediction method is given (16).

$$CV(\sigma) = \frac{1}{SM}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{j\in G_{-k}} P\left(b_j, \hat{g}_\sigma^{-k(j)}\left(y_j, \sigma\right)\right) \tag{16}$$

For selecting the ideal $\hat{\sigma}$ and to optimize the dyslexia prediction method given in (16), the $\hat{\sigma}$ is evaluated using (17).

$$\hat{\sigma} = \underset{\sigma\in\{\sigma_1,\ldots,\sigma_l\}}{\arg\min}\ CV_s(\sigma) \tag{17}$$

Moreover, the M in (16) has been defined to denoted the training-size of the dataset, $P(\cdot)$ represents the loss-function and $\hat{g}_\sigma^{-k(j)}(\cdot)$ is the function which is utilized for evaluating the coefficients. To build the most accurate dyslexia predictive method, we iteratively use (16), optimizing the error in training in the initial stage before passing the parameter values forward towards the second stage in order to learn and incorporate the feature's important characteristic within the method. By minimizing the objective-function utilizing the gradient descent method, an optimized solution for a given feature can be attained using optimization. Using the ranking algorithm $r(\cdot)$ given in (18), the most relevant feature is chosen for the dyslexia predictive method.

$$r(a) = \begin{cases} 0 \ \ if\ n_j\ is\ not\ selected \\ 1 \ \ if\ n_j\ is\ selected\ as\ optimal\ prediction\ model\ j = 1,2,3,\ldots,n \end{cases} \tag{18}$$

Further, the subset of features is built using (19).

$$F_s = \{r(n_1), r(n_1), \ldots, r(n_n)\} \tag{19}$$

Furthermore, we derive the optimal feature with the highest score over all possible K-folds instances using (20).

$$F_{s_k} = \{r(n_1), r(n_1), \ldots, r(n_n)\} \tag{20}$$

Finally, for the K feature subsets with the highest score, we calculate the frequency with which a certain feature was chosen using (21).

$$F_{s_{final}}=\{f_s(p_1),f_s(n_2),\ldots,f_s(n_n)\} \quad (21)$$

where, $f_s(\cdot)$ represents a scenario where the n^{th} feature may get selected or not. This can be defined using (22).

$$F_s(a)=\begin{cases} 0 \;\; if\ q_j\ is\ chosen\ lesser\ than\ \frac{K}{2}\ times, j=1,2,3,\ldots,n \\ 1 \;\; if\ q_j\ is\ chosen\ greater\ or\ equal\ to\ \frac{K}{2}\ times, j=1,2,3,\ldots,n \end{cases} \quad (22)$$

where, the (22) is utilized for generating the n' selected features subset, where, the n^{th} is used to define how much a respective feature has been chosen for prediction. In order to construct a reliable dyslexia predictive method, we begin by selecting a subset of the dyslexia training data based on certain features. K-folds are constructed by performing S iterations, with S being the number of times through which randomness is reduced throughout the training process. In the second stage, a subset of features is chosen with the goal of lowering variance. When compared to the state-of-the-art ML-based dyslexia predictive methods, the presented EFAM-based dyslexia predictive method provides vast improvements in overall prediction accuracy.

4. RESULT AND DISCUSSION

This section studies the performance achieved using proposed EFAM-XGB based dyslexia predictive method over standard XGB-based classification methods. Further, the model is compared with existing dyslexia predictive methods [25], [28]. The proposed and other existing dyslexia predictive methods were implemented using Anaconda Python framework. The accuracies, sensitivity, specificity, precision, and F-measure are metrics used for validating the classification algorithm performance. The specificity for the predictive method is evaluated using (23).

$$\text{Specificity}=\frac{\text{TN}}{\text{TN+FP}} \quad (23)$$

where, TP denotes the true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative. Further, the sensitivity for the predictive method is evaluated using (24).

$$\frac{\text{Sensitivity}}{\text{Recall}}=\frac{\text{TP}}{\text{TP+FN}} \quad (24)$$

The accuracy for the predictive method is evaluated using (25).

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (25)$$

The precision for the predictive method is evaluated using (26).

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \quad (26)$$

The F-measure for the predictive method is evaluated using (27).

$$\text{F}-\text{measure}=\frac{2*\text{Precision}*\text{Sensitivity}}{\text{Precision}*\text{Sensitivity}} \quad (27)$$

4.1. Dataset construction

Experiments were carried out using dyslexia [28] for this work. The dyslexia dataset [28] consists of various columns which describe the language vocabulary, speed, memory, visual-discrimination, audio-discrimination, survey-score and label of each participant. Each participant has taken a quiz having multiple question. From this quiz the speed, vocabulary, memory, audio and discrimination for every participant has been evaluated. For each correct answer they were given score for the respective section. As an additional measure, a survey-score' is determined using another survey. Based on this analysis, a 'Label' value, between 0 and 2, is generated. The student has a low, moderate, or high likelihood to suffer from dyslexia, as indicated by the numbers 0, 1, and 2, respectively.

4.2. Dyslexia performance study

The results for the performance study for the dyslexia classification in terms of accuracy, precision, recall, F1-score, and specificity have been given in Figures 2 to 6, respectively. In this section, the proposed EFAM-XGB has been compared with the existing DT, SVM-GS, RF-GS, and XGB. The accuracy, precision, recall, F1-score and specificity have been evaluated for all the models as shown in Table 1. The results show that the EFAM-XGB shows better performance in comparison to the existing models. The Figure 7(a) shows the important features which have been considered by EFAM-XGB and Figure 7(b) shows the important features which have been considered by XGB model.

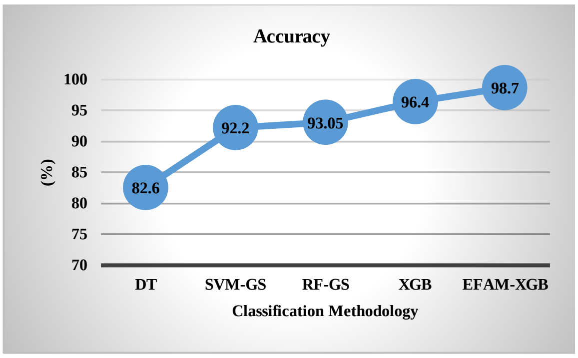

Figure 2. Accuracy performance for dyslexia classification

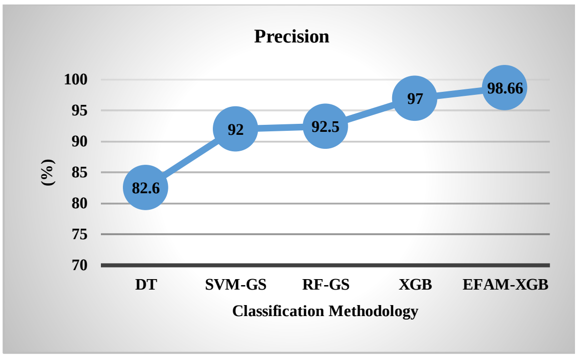

Figure 3. Precision performance for dyslexia classification

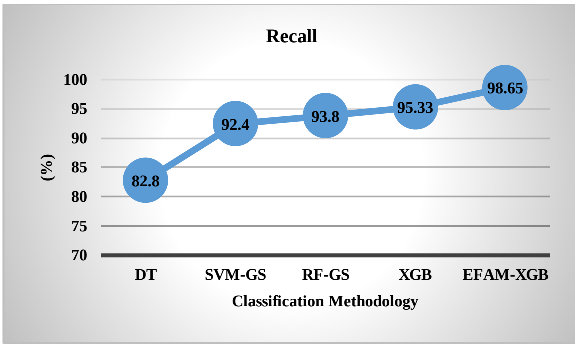

Figure 4. Recall performance for dyslexia classification

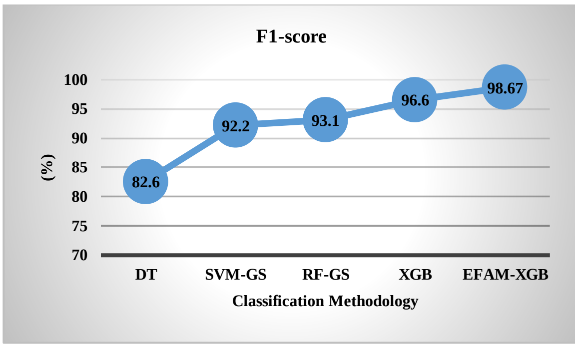

Figure 5. F1-score performance for dyslexia classification

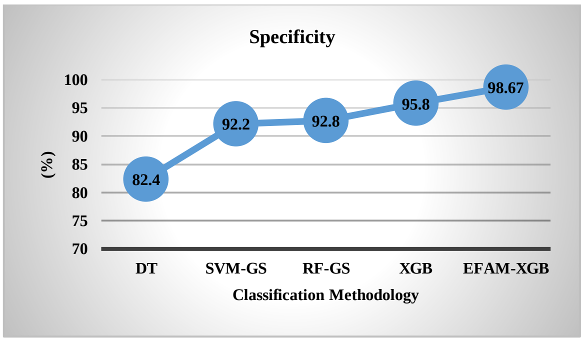

Figure 6. Specificity performance for dyslexia classification

Table 1. dyslexia classification performance study

	Accuracy	Precision	Recall	F-score	Specificity
DT	82.6	82.6	82.8	82.6	82.4
SVM-GS	92.2	92.0	92.4	92.2	92.2
RF-GS	93.05	92.5	93.8	93.1	92.8
XGB	96.4	97.0	95.33	96.6	95.8
EFAM-XGB	98.7	98.66	98.65	98.67	98.67

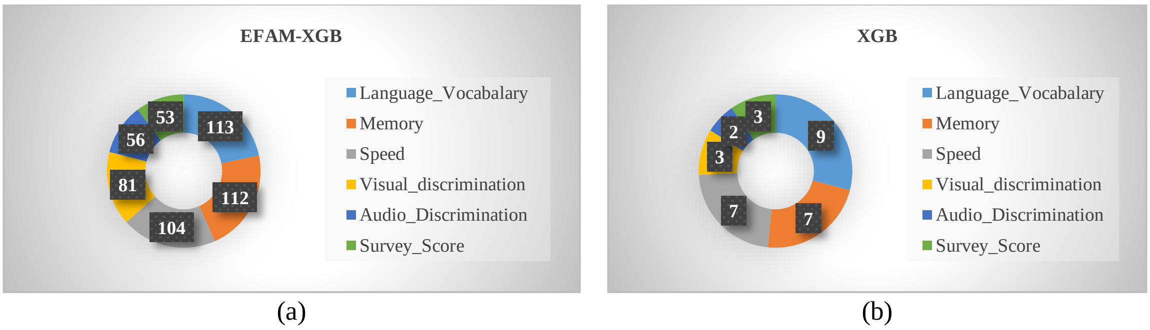

Figure 7. Feature importance analysis using: (a) EFAM-XGB and (b) XGB for performance dyslexia classification

5. CONCLUSION

The prediction of dyslexia using a quiz or by taking a test of student using multiple questions is very challenging tasks. In the recent years the ML and DL techniques have shown better performance for the prediction of the dyslexia. Nevertheless, these techniques fail to provide better performance accuracy when there exists concept drift in the dataset and the data is imbalanced. Hence, to solve this in this work a EFAM-XGB has been presented. This solves the data imbalance issues and concept drift issues while predicting whether the student is suffering from dyslexia or not. To solve this issue, in this work, a novel weight method has been presented. For the selection of the best features, in this work a K-CV method has been presented. The presented EFAM-XGB has been evaluated using dyslexia dataset. The results show that the EFAM-XGB provides better performance to predict whether the student is suffering from dyslexia even when there exists model drift issue due to data imbalance. Also, the proposed EFAM-XGB considers more features for providing accurate prediction when compared with the previous methods. From all the results, it can be seen that the proposed EFAM-XGB predicts dyslexia with higher accuracy. For, the future work, the proposed EFAM-XGB can be evaluated using other dyslexia's such as dysgraphia and dyscalculia. Also, the error during the training process for the multi-label classification can be reduced.

REFERENCES

[1] A. Kohli, S. Sharma, and S. Padhy, "Specific learning disabilities: Issues that remain unanswered," *Indian Journal of Psychological Medicine*, vol. 40, no. 5, pp. 399–405, 2018, doi: 10.4103/IJPSYM.IJPSYM_86_18.

[2] J. M. Lodge, G. Kennedy, L. Lockyer, A. Arguel, and M. Pachman, "Understanding difficulties and resulting confusion in learning: An integrative review," *Frontiers in Education*, vol. 3, Jun. 2018, doi: 10.3389/feduc.2018.00049.

[3] Y.-C. Yu, K. Shyntassov, A. Zewge, and L. Gabel, "Classification predictive modeling of dyslexia," in *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, IEEE, Mar. 2022, pp. 177–181, doi: 10.1109/CISS53076.2022.9751182.

[4] N. Giri *et al.*, "Detection of dyscalculia using machine learning," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, Jun. 2020, pp. 1–6, doi: 10.1109/ICCES48766.2020.9137871.

[5] J. Kunhoth, S. Al Maadeed, M. Saleh, and Y. Akbari, "Machine learning methods for dysgraphia screening with online handwriting features," in *2022 International Conference on Computer and Applications (ICCA)*, IEEE, Dec. 2022, pp. 1–6, doi: 10.1109/ICCA56443.2022.10039584.

[6] B. Dhanalakshmi, R. Dhanagopal, D. Raguraman, and T. Thamdapani, "Improving cognitive learning of children with dyspraxia using selection based mid-air gestures in Athynos game," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, Dec. 2020, pp. 231–237, doi: 10.1109/ICISS49785.2020.9316070.

[7] A. K. Dubey, A. Saxena, A. Kumar, and A. Rai, "Attention deficit hyperactivity disorder (ADHD) prediction system," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, IEEE, Apr. 2022, pp. 1321–1325, doi: 10.1109/ICSCDS53736.2022.9760921.

[8] M. B. Mohammed, L. Salsabil, M. Shahriar, S. S. Tanaaz, and A. Fahmin, "Identification of autism spectrum disorder through feature selection-based machine learning," in *2021 24th International Conference on Computer and Information Technology (ICCIT)*, IEEE, Dec. 2021, pp. 1–6, doi: 10.1109/ICCIT54785.2021.9689805.

[9] M. Orsoni *et al.*, "Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile," *Heliyon*, vol. 9, no. 3, pp. e14506–e14506, Mar. 2023, doi: 10.1016/j.heliyon.2023.e14506.

[10] D. Arumugam, K. Govindaraju, and A. K. Tamilarasan, "AIIoT-based smart framework for screening specific learning disabilities," in *Machine Learning for Critical Internet of Medical Things*, Cham: Springer International Publishing, 2022, pp. 103–124, doi: 10.1007/978-3-030-80928-7_5.

[11] C.-F. Tsai *et al.*, "A machine-learning-based assessment method for early-stage neurocognitive impairment by an immersive virtual supermarket," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2124–2132, 2021, doi: 10.1109/TNSRE.2021.3118918.

[12] G. M. Wiest, K. P. Rosales, L. Looney, E. H. Wong, and D. J. Wiest, "Utilizing cognitive training to improve working memory, attention, and impulsivity in school-aged children with ADHD and SLD," *Brain Sciences*, vol. 12, no. 2, Jan. 2022, doi: 10.3390/brainsci12020141.

[13] H. Munir, B. Vogel, and A. Jacobsson, "Artificial intelligence and machine learning approaches in digital education: A systematic revision," *Information*, vol. 13, no. 4, Apr. 2022, doi: 10.3390/info13040203.

[14] S. Kaisar and A. Chowdhury, "Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests," *ICT Express*, vol. 8, no. 4, pp. 563–568, Dec. 2022, doi: 10.1016/j.icte.2022.02.011.

[15] W. Satriaji and R. Kusumaningrum, “Effect of synthetic minority oversampling technique (SMOTE), feature representation, and classification algorithm on imbalanced sentiment analysis,” in *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, Oct. 2018, pp. 1–5, doi: 10.1109/ICICOS.2018.8621648.
[16] B. N. Nakkas, “Feature selection and SMOTE based recommendation for Parkinson’s imbalanced dataset prediction problem,” in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, IEEE, May 2022, pp. 1–4, doi: 10.1109/SIU55565.2022.9864672.
[17] J. Kohn *et al.*, “Efficacy of a computer-based learning program in children with developmental dyscalculia. what influences individual responsiveness?,” *Frontiers in Psychology*, vol. 11, Jul. 2020, doi: 10.3389/fpsyg.2020.01115.
[18] T. Käser *et al.*, “Design and evaluation of the computer-based training program Calcularis for enhancing numerical cognition,” *Frontiers in Psychology*, vol. 4, Aug. 2013, doi: 10.3389/fpsyg.2013.00489.
[19] M. Babu *et al.*, “Web based assessment and training model for dyslexia dyscalculia dysgraphia dyspraxia ADHD & autism,” in *4th International Conference on Inventive Research in Computing Applications, ICIRCA 2022 - Proceedings*, IEEE, Sep. 2022, pp. 1753–1757, doi: 10.1109/ICIRCA54612.2022.9985715.
[20] P. Dhamal and S. Mehrotra, “Finding learning disorder based on machine learning,” *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3852544.
[21] J. Kunhoth, S. Al Maadeed, M. Saleh, and Y. Akbari, “CNN feature and classifier fusion on novel transformed image dataset for dysgraphia diagnosis in children,” *Expert Systems with Applications*, vol. 231, 2023, doi: 10.1016/j.eswa.2023.120740.
[22] V. Vilasini, B. B. Rekha, V. Sandeep, and V. C. Venkatesh, “Deep learning techniques to detect learning disabilities among children using handwriting,” in *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, IEEE, Aug. 2022, pp. 1710–1717, doi: 10.1109/ICICICT54557.2022.9917890.
[23] C. Hewapathirana, K. Abeysinghe, P. Maheshani, P. Liyanage, J. Krishara, and S. Thelijjagoda, “A mobile-based screening and refinement system to identify the risk of dyscalculia and dysgraphia learning disabilities in primary school students,” in *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, IEEE, Aug. 2021, pp. 287–292, doi: 10.1109/ICIAfS52090.2021.9605998.
[24] M. Modak, O. Warade, G. Saiprasad, and S. Shekhar, “Machine learning based learning disability detection using LMS,” in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, IEEE, Oct. 2020, pp. 414–419, doi: 10.1109/ICCCA49541.2020.9250761.
[25] A. J. Prabha and R. Bhargavi, “Prediction of dyslexia using machine learning—A research travelogue,” in *Lecture Notes in Electrical Engineering*, Springer Singapore, 2019, pp. 23–34, doi: 10.1007/978-981-13-7091-5_3.
[26] P. Drotár and M. Dobeš, “Dysgraphia detection through machine learning,” *Scientific Reports*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-78611-9.
[27] F. Masood, W. U. Khan, K. Ullah, A. Khan, F. H. Alghamedy, and H. Aljuaid, “A hybrid CNN-LSTM random forest model for dysgraphia classification from hand-written characters with uniform/normal distribution,” *Applied Sciences*, vol. 13, no. 7, Mar. 2023, doi: 10.3390/app13074275.
[28] N. D. Alqahtani, B. Alzahrani, and M. S. Ramzan, “Deep learning applications for dyslexia prediction,” *Applied Sciences*, vol. 13, no. 5, 2023, doi: 10.3390/app13052804.
[29] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA: ACM, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[30] D. Nielsen, “Tree boosting with XGBoost why does XGBoost win ‘every’ machine learning competition?,” M.Sc. Thesis, Department of Mathematical Science, Norwegian University of Scence Technology, Trondheim, Norwegia, 2016.

BIOGRAPHIES OF AUTHORS

Sailaja Mulakaluri is a dedicated professional with a strong background in Computer Science and Applications. Currently serving as an Assistant Professor in the Department of Computer Science and Applications at St. Francis de Sales College, she is also actively engaged in research as a Research Scholar in the Department of Computer Science and Engineering at Dayananda Sagar University. She has made significant contributions to the field of education and technology, particularly in the area of identifying and addressing learning difficulties in children. Her research work reflects a keen interest in applying machine learning and data mining techniques to develop frameworks and tools for diagnosing and managing specific learning disabilities. She can be contacted at email: sailaja.m@gmail.com.

Girisha Gowdra Shivappa is working as Professor in the Department of Computer Science & Engineering and has done his PhD from VTU, Belagavi in the domain of Multimedia Data Mining. He has total of 24 years of academic’s experience. He has 25 publications in his credit including Eleven international peer reviewed journals and eight international conference papers. He has given invited lectures on advanced microprocessor and block chain technology in various Engineering Colleges. He was involved in syllabus framing for the AMIETE courses. He is currently guiding five Ph.D. students. He has worked as coordinator in various departmental and institute levels works including, mentoring, organizing conferences/workshops/FDP/Technical talks, ISO, NBA & NBA accreditation. His areas of research interest include data mining, data science, and block chain. He can be contacted at email: girisha-cse@dsu.edu.in.