Doi : 10.48047/jocaaa.33.6.1227

AN ENSEMBLE LEARNING APPROACH FOR DYSLEXIA PREDICTION

Dr. K. Sunil Kumar [1]Myakala Manoj Kumar[2], K R Anudeep Laxmikanth[3] and Suneetha Madduluri[4]

[1234] Department of Electronics and Communication, National Institute of Technology Warangal, India,
sunil.veena10@gmail.com,mk23ecr1r14@student.nitw.ac.in,kralk1989@gmail.com, suneethasivakrishna@gmail.com.

Abstract. A neurodevelopmental disorder Dyslexia has a global impact on millions of children with sloppy word comprehension and subpar reading abilities. Which affects their academic, emotional, and social well-being. Early detection of dyslexia can be highly beneficial for dyslexic children as their learning needs can be properly addressed. Machine learning methods have been implemented to recognize dyslexia with various datasets acquired from medical and educational organization. An innovative approach to early dyslexia detection By employing ensemble machine learning method, such as Logistic Regression, Random Forest, Nearest Neighbors (KNN), and a Support Vector Machine (SVM).In this research indicates that all models exhibit commendable performance, with Logistic Regression achieving the highest accuracy 89.59%, followed closely by KNN and Random Forest. Likewise, SVM demonstrates a commendable accuracy rate of 86.58%. With an impressive 91.50 percent accuracy, the ensemble model, which outperforms individual models and demonstrates the efficacy of ensemble method.

Keywords: Dyslexia, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM).

1 Introduction

Most Dyslexia, a neurological disorder characterized by difficulties in word comprehension and reading abilities, presents significant challenges for affected individuals, particularly children [1]. With approximately 10% of the population affected by dyslexia [2], early detection becomes imperative for effective intervention and support. However, identifying individuals at risk of dyslexia during the early years of schooling remains a daunting task, necessitating innovative approaches for diagnosis and intervention [3].

Researchers have explored various methods for diagnosing dyslexia, leveraging data from diverse sources such as reading and writing assessments, online word games, and neuroimaging techniques like MRI and EEG scans. Among these methods, machine learning techniques have emerged as promising tools for dyslexia detection due to their ability to offer improved accuracy and predictive outcomes [4]. Despite the growing interest in dyslexia detection, there is a notable gap in the literature concerning machine learning approaches specifically tailored for this purpose [3, 5, 6].

Machine learning, as the name suggests, enables systems or machines to learn from data patterns using algorithms and models without explicit programming. By analyzing vast datasets quickly and effectively, machine learning algorithms can discern intricate patterns and make predictions beyond human capability. Ensemble machine learning, a subset of machine learning, combines multiple models or algorithms to enhance problem-solving efficiency, accuracy, and robustness. Through techniques like bagging, boosting, and stacking, ensemble learning harnesses the collective intelligence of diverse models to improve outcomes for complex tasks [6].

Doi : 10.48047/jocaaa.33.6.1227

Machine learning, as the name suggests, enables systems or machines to learn from data patterns using algorithms and models without explicit programming. By analyzing vast datasets quickly and effectively, machine learning algorithms can discern intricate patterns and make predictions beyond human capability. Ensemble machine learning, a subset of machine learning, combines multiple models or algorithms to enhance problem-solving efficiency, accuracy, and robustness. Through techniques like bagging, boosting, and stacking, ensemble learning harnesses the collective intelligence of diverse models to improve outcomes for complex tasks [6].

The literature review encompasses existing research on machine learning methods for dyslexia biomarker detection, emphasizing implementation details and challenges. Data collection, preparation, and pre-processing are crucial steps in machine learning model development, ensuring the quality and relevance of training data. Feature extraction and selection techniques further refine the dataset, extracting pertinent information for classification models. Model training and classification involve the utilization of diverse machine learning algorithms to predict dyslexia risk based on input data[6].

2 ML for Dyslexia Detection

All Four phases are involved in utilizing machine learning approaches to diagnose developmental dyslexia: (i) collecting data; (ii) preprocessing, , and feature selection; (iii) feature extraction (iv) training and classification.A schematic illustration of the steps that are covered here can be found in Fig. 1.

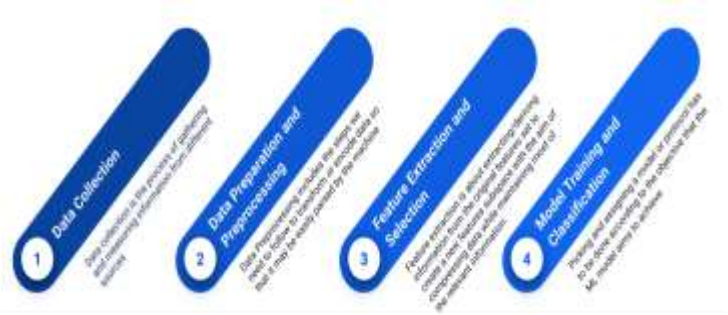

Figure 1:. Design Flow For ML algorithm

Every User research is performed as the initial stage in dyslexia detection in order to collect user data. In the context of traditional dyslexia detection methods, behavioral dimensions of participants' performance on standardized tests including reading and writing, phonological awareness, and working memory are assessed by psychologists. Dyslexic individuals are identified according to their subpar performance on said assessments.As the symptoms differ among individuals, these methods are frequently ineffective and time-consuming for large groups of participants. Scientists employ machine learning methodologies as a result, as they are frequently inexpensive and require less time. Reading [7–12,14,16–18], writing and typing [15], handwriting [19], and a web-based game [13] are among the many assessments utilized to gather a variety of data types, including text [7,12,13,15], eye movement [8,9,17], MRI scans [10,11], EEG scans [14,15,18], and images [16,19]. The participants' ages ranged from seven to sixty two years, and their native tongues included Malay, English, Greek, Spanish, Hebrew, Swedish, Mandarin, French, German, Polish, and Spanish.In light of the fact that the majority of these studies were carried out exclusively in a particular language, a language-independent game-based test [20] would be a more suitable option. While [20] em-

Doi : 10.48047/jocaaa.33.6.1227

ploys a language-independent data capture approach, no machine learning algorithm was utilized. Certain methodologies also necessitate the implementation of specialized equipment in a laboratory setting, including an EEG headset [15], a customized camera [17], an infrared ocular reflection system [9], an MRI scanner [10], and an eye tracker [8], in order to gather EEG or MRI scan data. While these methodologies attain superior precision, they are costly, limited in scope to a subset of users, and potentially induce atypical behavior among participants during testing or observation. Computer-based reading or writing assessments, in conjunction with game-based methodologies, may offer greater advantages in this context. In light of the increasing prevalence of smart mobile devices, an app-based data acquisition method will additionally facilitate the expansion of the target user demographic.

2.2. Preprocessing

Wherever Prior to its application in machine learning methodologies, the gathered data must undergo preprocessing and filtration. In the first step, the data must be transformed into either a quantitative (numbers) or qualitative (textual categories) format. In order to ac-accomplish this, EEG scan data must be transformed into high pass and low pass filters. In this instance, various wavelet transform techniques are implemented; for instance, Fred et al. [14] implemented discrete wavelet transform. Additionally, some of the studies employed manual preprocessing and feature extraction [12,16], whereas others utilized PANDA [10] and Free Surfer [11]. Preprocessing is performed with the objective of identifying pertinent attributes and eliminating invalid values. And successively eliminates a number of features in iterations. When the computational complexity increases in proportion to the number of features, feature selection becomes a critical undertaking. Comparative performance analyses of various feature selection techniques, however, have not been included in prior research. Same line.

2.3 Feature Selection

Wherever prior to its application in machine learning methodologies, the gathered data must undergo preprocessing, during which pertinent features are identified and designated a range of values. Consider both numeric and categorical values. The quantity of features exhibited variation among investigations, ranging from 12 to 226 [8, 13]. Identifying the set of prominent characteristics that are most significant in determining the object's class is the following step. A few studies employed manual selection [14] in order to accomplish this, whereas others utilized methods including least absolute shrinkage and selection operator (LASSO) [17] SVM-RFE as well [9]. By effectively executing regularization and variable selection operations concurrently, LASSO can be utilized to enhance both accuracy and interpretability. These models are appropriate for regression purposes. Conversely, SVMRFE selects features based on their significance in facilitating the separation of classes by SVM classifiers. This method commences with an exhaustive feature set.

2.4. Training and Classification

Following the process of feature selection, machine learning algorithms are employed to perform system training and classification. Separated into training and testing subsets is the dataset. The majority of previous research employed 10-fold cross-validation, in which the dataset is divided into ten equal parts. Of these parts, nine were used to train the algorithm, while the remaining set was utilized to evaluate its performance [8, 11,13]. However, alternative splits such as fivefold [17], leave-one-out-cross-validation (LOOCV) [10, 11,18], and 70–30 [12] were also discussed. The utilization of supervised classification algorithms is limited to testing objectives, as the training dataset already comprises class information such as dyslexia or non-dyslexia. Support vector machine (SVM), Naive Bayes, Logistic regression, neural network, K-Nearest Neighbour (K-NN), and Linear Discriminant Analysis (LDA) were the prevailing machine learning algorithms utilized in prior research to categories participants.SVM was the algorithm most frequent-

Doi : 10.48047/jocaaa.33.6.1227

ly implemented in numerous investigations. Identifying dyslexic and non-dyslexic users is essentially a binary classification problem; there-fore, SVM is anticipated to perform well when the feature space is sparse and the number of dimensions exceeds the number of samples. Nevertheless, SVM interpretation is a difficult endeavor, and its performance degrades as the amount of noise in the dataset increases. A method like logistic regression, on the other hand, is simpler to implement and comprehend, and it is anticipated to yield an excellent solution for binary problems classification-related issues. In general, the choice of suitable classification methods would largely be determined by the data itself. Therefore, research should demonstrate the outcomes of various machine learning models through comparative performance analyses, as opposed to simply reporting the performance of a single model. Furthermore, the implementation of ensemble methods may prove advantageous in attaining enhanced performances.

For performance evaluation, MATLAB, WEKA, and Python-based tools were utilized in the prior literature. In this particular instance, the efficacy of dyslexia detection methods utilizing machine learning approaches was assessed utilizing a variety of metrics. The aforementioned metrics comprise the area under the receiver operating characteristic (ROC) curve, mean square error (MSE), positive predictive value (PPV), negative predictive value (NPV), accuracy, sensitivity, specificity, and recall. Sensitivity and specificity quantify the proportion of correctly identified dyslexic and non-dyslexic users, respectively, whereas accuracy ranks the number of accurately classified objects in relation to the total number of objects. Recall is the proportion of the total number of dyslexic users that were accurately identified, whereas recall, or positive predictive value, denotes the proportion of dyslexic users that were accurately identified relative to the total number of dyslexic users identified. In various studies [15], EEG-based methods attained an accuracy ranging from 60% to 80%, whereas MRI scan-based methods attained an accuracy of 83.61% [10] and game-based techniques attained an accuracy of 80.24% [8].

1. Dyslexia Detection using Ensemble Based ML

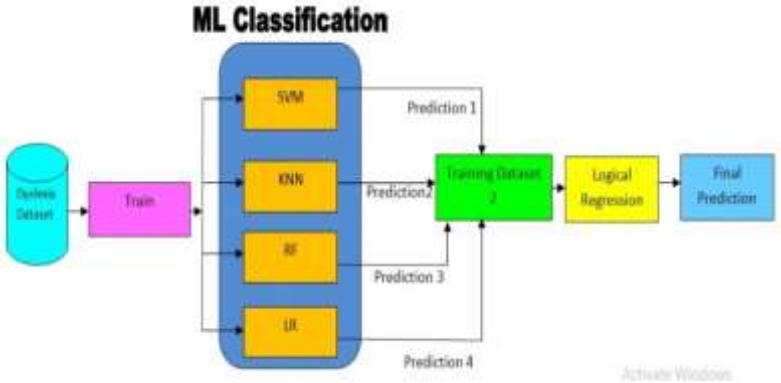

Figure 2: Proposed ensemble Model

The Fig. 2 describes the proposed ensemble model for prediction of dyslexia. This section explains the details of implementation of the proposed model. Ensemble machine learning mixes various models or algorithms for a communal approach to problem-solving. Ensemble approaches outperform individual models in terms of performance, accuracy, and robustness by taking advantage of the characteristics of various models. Predictions are combined using bagging, boosting, random forests, and stacking. The benefits of ensemble learning include higher accuracy, robustness, and the capacity to handle complicated problems. Ensemble learning is adaptable and applicable to a variety of machine learning techniques. It manages noisy data and lessens over fitting. In simpler words, Ensemble machine learning uses numerous models' collective intelligence to improve outcomes for machine learning tasks.

Doi : 10.48047/jocaaa.33.6.1227

Despite significant progress in dyslexia detection, research gaps persist, particularly in the development of ensemble machine learning models tailored for heterogeneous datasets and specialized deep learning architectures. Addressing these gaps holds the potential to significantly impact dyslexia diagnosis and intervention, empowering individuals with dyslexia to achieve their full potential academically, socially, and economically.

In light of the pressing need for accurate and timely dyslexia detection, this study aims to develop machine learning models capable of identifying dyslexia risk with high accuracy. Leveraging ensemble machine learning techniques, the study seeks to enhance the performance of baseline models like Random Forest and explore stacked models that combine the strengths of diverse machine learning approaches. By advancing the state-of-the-art in dyslexia detection, this research endeavors to make meaningful contributions to the field, ultimately improving outcomes for individuals affected by dyslexia. We have implemented the Ensemble machine Learning-based model Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and the K-Nearest Neighbors (KNN) compared them with the chosen Baseline Model.

2. Experimental Results

The dataset used in the study consists of 196 features for each participant. Here's a breakdown of the features:

1: Gender (Binary): Indicates the gender of the participant (Female or Male).
2: Native language (Binary): Specifies whether the participant's native language is Spanish (Yes) or if they are bilingual with Spanish being one of the languages (No).
3: Language subject (Binary): Indicates whether the participant has failed a language subject at school at least once (Yes) or if they have never failed that subject (No).
: Age: Represents the age of the participant, ranging from 7 to 17 years old.

The remaining features (5-196) are performance features derived from the participants' interaction while playing a test consisting of 32 questions. These features are divided into different categories based on the type of task and the skills targeted:[21]

5-28: Alphabetic Awareness, Phonological Awareness, and Visual discrimination and categorization tasks related to questions 1 to 4. [21]
29-58: Phonological Awareness, Syllabic Awareness, and Auditory Discrimination and Categorization tasks related to questions 5 to 9. [21]
59-82: Lexical Awareness, Auditory Working Memory, and Auditory Discrimination and Categorization tasks related to questions 10 to 13. [21]
83-106: Visual Discrimination and Categorization, and Executive Functions tasks related to questions 14 to 17. [21]
107-130: Spelling tasks related to questions 18 to 21. [21]
131-142: Visual Working Memory, Sequential Auditory Working Memory, and Auditory Discrimination and Categorization tasks related to questions 22 and 23. [21] 143-148: Lexical, Phonological, and Orthographic Awareness tasks related to questions 24 and 25. [21]
149-154: Morphological and Semantic Awareness tasks related to question 24. [21] 155-160: Syntactic Awareness tasks related to question 25. [21]
161-172: Phonological, Lexical, and Orthographic Awareness tasks related to question 26. [21]
173-178: Rearranging letters and syllables tasks related to questions 27 and 28[21]. 179-184: Word separation tasks related to question 29. [21]

Doi : 10.48047/jocaaa.33.6.1227

185-196: Sequential Visual Working Memory tasks related to question 30. [21]

Additionally, the dataset includes dictation tasks for questions 31 and 32, involving writing words (working memory, phonological awareness and sequential auditory) and (orthographic awareness, auditory working memory and lexical), respectively. These features were used to assess participants' performance and target various cognitive skills and abilities related to language acquisition and processing.

Tab 1: Accuracy Comparison of different models

Algo-rithms	Accu-racy	Re-marks
RF Model	87.9%	Baseline Model
SVM Model	86.6%	
LR Model	89.5%	
KNN Model	87.4%	
Stacked Model	91.3%	Proposed Model

Tab.1 gives Accuracy Comparison of different models LR achieved the highest accuracy score of 89.59%, followed closely by KNN and RF with accuracies of 87.40% and 87.95%, respectively. SVM also demonstrated respectable performance with an accuracy of 86.58%. The stacked model, formed by combining the predictions of Logistic Regression, KNN, Random Forest, SVM, using a meta-learner, achieved a notable accuracy of 91.50%. This indicates that combining the strengths of multiple models through stacking can lead to improved predictive performance compared to individual models. We have used Area under Precision Recall curve and Area under ROC curve performance metrics to compare the performance of those models.

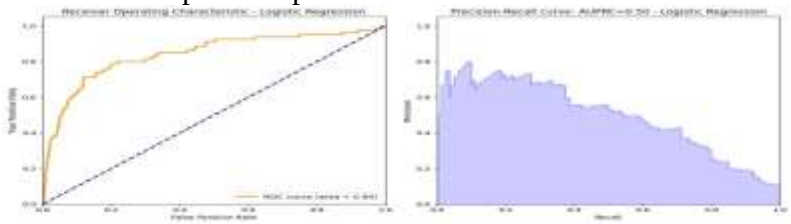

Figure: 4 ROC LR

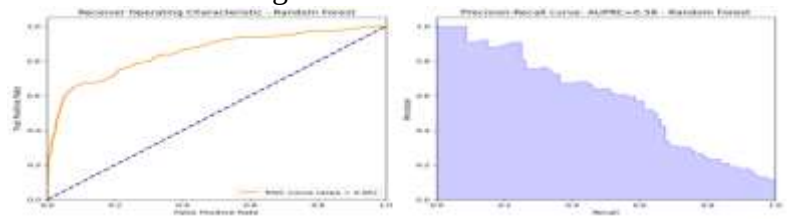

Figure: 5 ROC RF

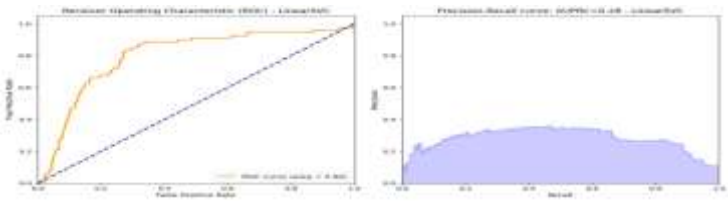

Figure: 6 ROC Linear SVC

Doi : 10.48047/jocaaa.33.6.1227

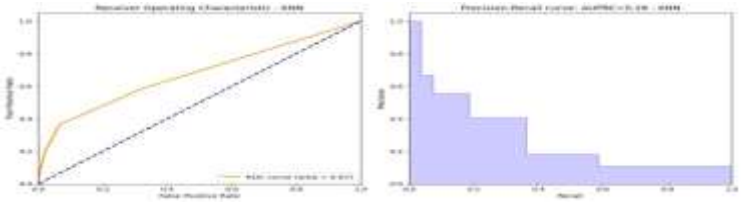

Figure: 7 ROC KNN

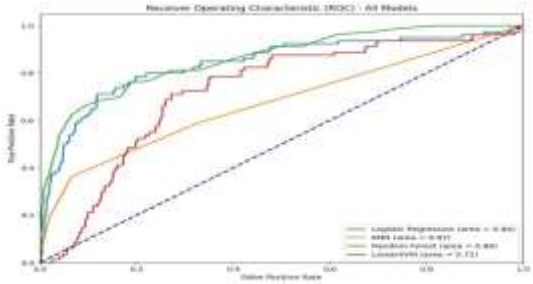

Figure: 8 ROC all Models without stack model

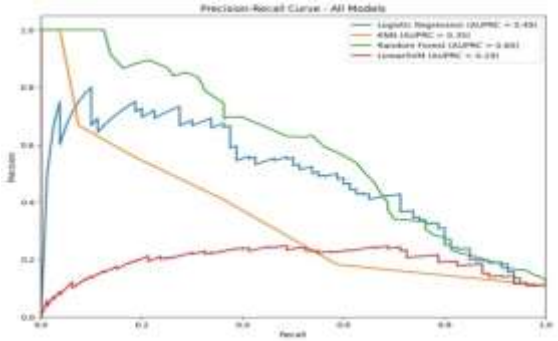

Figure: 9. PRC all models without stack model (It plots the Precision vs Recall)

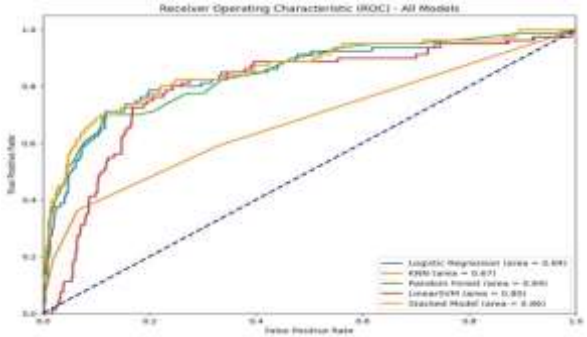

Figure: 10 ROC all Models with stack model It plots the true positive rate (sensitivity) vs. false positive rate(specificity) . More the area under theROC curve the better the model is. Stacked model has higher AUROC than Random Forest Model

Doi : 10.48047/jocaaa.33.6.1227

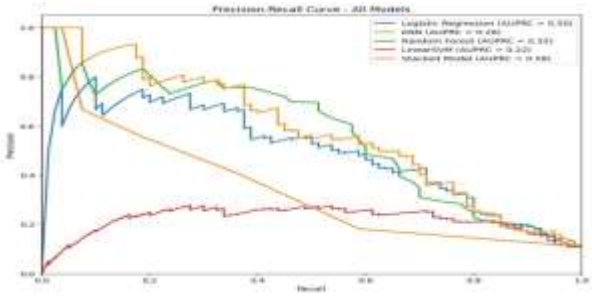

Figure:11 PRC all Models with stack model ((It plots the Precision vs Recall. More the area under the PRC curve the better the model is. Stacked model has higher AUPRC than Random Forest Model))

3. Conclusion

In this paper, based on the accuracy scores obtained from various classification algorithms for the task of classifying Dyslexia using the given dataset, several conclusions can be drawn. Among the individual models, LR achieved the highest accuracy score of 89.59%, followed closely by KNN and RF with accuracies of 87.40% and 87.95%, respectively .SVM also demonstrated respectable performance with an accuracy of 86.58%.

The stacked model, formed by combining the predictions of Logistic Regression, KNN, Random Forest, SVM, using a meta-learner, achieved a notable accuracy of 91.50%. This indicates that combining the strengths of multiple models through stacking can lead to improved predictive performance compared to individual models. The high accuracy scores obtained by Logistic Regression, KNN, Random Forest, and SVM suggest that these algorithms are well-suited for the classification of Dyslexia using the given dataset. the results demonstrate the effectiveness of Logistic Regression, KNN, Random Forest, and SVM in accurately classifying Dyslexia. Furthermore, the stacked model, leveraging the diverse predictions of multiple algorithms, yielded even higher accuracy, highlighting the potential of ensemble methods in improving classification performance.

Acknowledgments

The authors would like to acknowledge this work was supported by the Science and Engineering Research Board, a Statutory Body Established through an Act of Parliament: SERB Act 2008Government of India. (SERB Project File No: EEQ/2023/006970.

References

[1] What is dyslexia? 2020, https://dyslexiaassociation.org.au/what-is-dyslexia/, accessed on 28 February 2020.

[2] Dyslexia in Australia, 2020, https://dyslexiaassociation.org.au/dyslexia -in-australia/, accessed on 28 February 2020.

[3] S.S.A. Hamid, N. Admodisastro, A. Kamaruddin, A study of computerbasedlearning model for students with dyslexia, in: 2015 9thMalaysian Software Engineering Conference (MySEC), IEEE, 2015, pp. 284–289.

[4] Dyslexia, 2020, https://www.open.edu/openlearn/education-development/education/understanding-dyslexia/content-section-1.7.2, accessed on28 February 2020.

Doi : 10.48047/jocaaa.33.6.1227

[5] H. Perera, M.F. Shiratuddin, K.W. Wong, Review of eeg-based patternclassification frameworks for dyslexia, Brain Inform. 5 (2) (2018) 4.
[6] H. Perera, M.F. Shiratuddin, K.W. Wong, Review of the role ofmodern computational technologies in the detection of dyslexia, in:Information Science and Applications (ICISA) 2016, Springer, 2016,pp. 1465–1475.
[7] Y. Lakretz, G. Chechik, N. Friedmann, M. Rosen-Zvi, Probabilisticgraphical models of dyslexia, in: Proceedings of the 21th ACMSIGKDD International Conference on Knowledge Discovery and DataMining, 2015, pp. 1919–1928.
[8] L. Rello, M. Ballesteros, Detecting readers with dyslexia using machine learning with eye tracking measures, in: Proceedings of the 12th Web for All Conference, 2015, pp. 1–8.
[9] M.N. Benfatto, G.Ö. Seimyr, J. Ygge, T. Pansell, A. Rydberg, C.Jacobson, Screening for dyslexia using eye tracking during reading,PLoS One 11 (12) (2016).
[10] Z. Cui, Z. Xia, M. Su, H. Shu, G. Gong, Disrupted white matter connectivity underlying developmental dyslexia: a machine learningapproach, Hum. Brain Mapp. 37 (4) (2016) 1443–1458.
[11] P. Płoński, W. Gradkowski, I. Altarelli, K. Monzalvo, M. vanErmingen-Marbach, M. Grande, S. Heim, A. Marchewka, P. Bogorodzki,F. Ramus, et al., Multi-parameter machine learning approach to the neuroanatomical basis of developmental dyslexia, Hum. Brain Mapp. 38 (2) (2017) 900–908.
[12] R.U. Khan, J.L.A. Cheng, O.Y. Bee, Machine learning and dyslexia: Diagnostic and classification system (dcs) for kids with learning disabilities, Int. J. Eng. Technol. 7 (3.18) (2018) 97–100.
[13] L. Rello, E. Romero, M. Rauschenberger, A. Ali, K. Williams, J.P.Bigham, N.C. White, Screening dyslexia for English using hci measures and machine learning, in: Proceedings of the 2018 International Conference on Digital Health, 2018, pp. 80–84.
[14] A. Frid, L.M. Manevitz, Features and machine learning for correlating and classifying between brain areas and dyslexia, 2018, arXiv preprint arXiv:1812.10622.
[15] H. Perera, M.F. Shiratuddin, K.W. Wong, K. Fullarton, Eeg signal analysis of writing and typing between adults with dyslexia and normal controls, Int. J. Interact. Multimedia Artif. Intell. 5 (1) (2018) 62.
[16] S.S.A. Hamid, N. Admodisastro, N. Manshor, A. Kamaruddin, A.A.A.Ghani, Dyslexia adaptive learning model: student engagement prediction using machine learning approach, in: International Conference on Soft Computing and Data Mining, Springer, 2018, pp. 372–384.
[17] T. Asvestopoulou, V. Manousaki, A. Psistakis, I. Smyrnakis, V.Andreadakis, I.M. Aslanides, M. Papadopouli, Dyslexml: Screening tool for dyslexia using machine learning, 2019, arXiv preprint arXiv:1903.06274.
[18] Z. Rezvani, M. Zare, G. Žarić, M. Bonte, J. Tijms, M. Van der Molen,G.F. González, Machine learning classification of dyslexic children based on eeg local network features, 2019, pp. 1–23, bioRxiv.
[19] K. Spoon, D. Crandall, K. Siek, Towards detecting dyslexia in children's handwriting using neural networks, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 1–5.
[20] M. Rauschenberger, L. Rello, R. Baeza-Yates, J.P. Bigham, Towards language independent detection of dyslexia with a web-based game, in: Proceedings of the Internet of Accessible Things, 2018, pp. 1–10.
[21] L. Rello, R. Baeza-Yates, A. Ali, J. P. Bigham, and M. Serra, "Predicting risk of dyslexia with an online gamified test," PLOS ONE, vol. 15, no. 12, p. e0241687, Dec. 2020, doi: 10.1371/journal.pone.0241687.

Doi : 10.48047/jocaaa.33.6.1227