Int. J. Sci. R. Tech., 2025 2(3)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

Advance Machine Learning Methods for Dyslexia Biomarker Detection

Dr. K. N. Sivabalan[1], Nandhika G.[2], Anupriya S.[2], Divya M.[2]

[1]*Professor, Department of Information Technology, Sri Krishna College of Engineering and Technology*
[2]*Department of Information Technology, Sri Krishna College of Engineering and Technology*

ABSTRACT

Dyslexia is a neurological disease that greatly impairs reading and comprehension ability, notably in school-aged children, with a higher prevalence in boys. This syndrome can result in poor academic achievement and long-term effects on self-esteem. To improve dyslexia detection, this study investigates the application of the Random Forest method, a machine learning technique noted for its durability and accuracy. The Random Forest algorithm is used to classify persons with dyslexia using a dataset of brain and behavioral markers. The goal is to uncover patterns and biomarkers that can aid in early detection. This strategy addresses major issues in dyslexia diagnosis, such as the necessity for interpretable biomarkers and the risk of overfitting, by taking advantage of Random Forests' ensemble nature to improve model reliability and generalization. The results of this study show that the Random Forest algorithm has the capacity to detect dyslexia with clinically significant accuracy, making it a promising tool for early intervention and support.

Keywords: Dyslexia Detection, Feature Extraction, Diagnosing Dyslexia

INTRODUCTION

Dyslexia, a learning condition that impairs reading and language processing, is typically overlooked until severe difficulties occur in academic or daily tasks. Early diagnosis is critical for offering effective therapies, and traditional methods are based on subjective assessments. However, with the advent of deep learning, prediction algorithms can now analyses patterns in voice, text, and visual data to detect dyslexia indicators more precisely and earlier than ever before. These models, which use techniques like as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, can interpret complicated multimodal data, making them a possible solution for enhancing dyslexia prediction and interventions.

1.1 Dyslexia Detection

Dyslexia identification is detecting the cognitive and linguistic patterns that distinguish dyslexics from non-dyslexics, with a focus on phonological processing, reading fluency, and visual attention. Traditionally, this has been accomplished through educational exams, such as reading tests and psychological examinations. However, recent advances in artificial intelligence, including deep learning, have improved the accuracy and efficiency of dyslexia identification. Deep learning models can detect subtle dyslexic characteristics by analyzing patterns in spoken language, written text, and even eye movements during reading activities. These models, particularly those that employ Convolutional Neural Networks (CNNs) for visual data such as eye-tracking and Long Short-Term Memory (LSTM) networks for audio or sequential data, can reveal intricate, non-linear associations that would not be obvious through manual analysis. This shift towards automated, data-driven detection paves the path for earlier and more personalized interventions, which may help to alleviate the academic and emotional issues associated with dyslexia.

1.2 Feature Extraction

Feature extraction in dyslexia detection is an important procedure that involves finding and separating the most significant qualities from raw

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

data, such as speech recordings, text samples, or eye-tracking metrics, in order to increase predictive model performance. In deep learning, neural networks usually handle this procedure automatically. In speech analysis, characteristics such as pronunciation errors, hesitation patterns, and phoneme articulation may be retrieved using Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models. Linguistic elements in text data, such as spelling errors, reading fluency, and sentence structure, are important indications of dyslexia and can be detected using Natural Language Processing (NLP) approaches. Similarly, Convolutional Neural Networks (CNNs) may extract visual elements from eye-tracking data, including gaze duration, fixation spots, and irregular saccades, which are frequently related with reading difficulties. Deep learning models are better suited to detect subtle dyslexia symptoms by focusing on these important aspects, which improves predicted accuracy and allows for early intervention.

1.3 Diagnosing Dyslexia

Diagnosing dyslexia entails a thorough evaluation of an individual's reading abilities, phonological skills, and cognitive functions to determine the presence of this learning condition. Traditionally, diagnosis has been based on standardized tests and observational assessments administered by educational psychologists or specialists, with a focus on the individual's performance in reading, spelling, and writing. However, technological improvements have provided novel approaches to dyslexia diagnosis, such as the use of machine learning and deep learning algorithms. These models use a variety of data sources, including speech patterns, text analysis, and eye-tracking metrics, to detect specific behavioral and linguistic characteristics associated with dyslexia. By processing vast datasets, these algorithms can find intricate links that traditional assessments may miss, resulting in a more objective and accurate diagnosis. This current technique not only facilitates early detection of dyslexia, but it also allows for individualized interventions that target each individual's unique needs, resulting in superior educational achievements.

LITERATURE REVIEW

Shahriar Kaisar [1] et al. proposed in this system Developmental dyslexia is a learning problem that primarily affects children during their early childhood. Dyslexic children struggle to read, spell, and write words, despite having normal or above-average intelligence. As a result, dyslexic youngsters often experience negative emotions such as low self-esteem, frustration, and rage. As a result, early detection of dyslexia is critical for supporting dyslexic children from the beginning. Researchers have proposed a variety of methods for detecting developmental dyslexia, including game-based approaches, reading and writing exams, facial image capture and analysis, eye tracking, magnetic reasoning imaging (MRI), and electroencephalography (EEG) scans. This overview paper examines current advances to detecting dyslexia using machine learning approaches and identifies promising areas for future research. The term 'Dyslexia' derives from the Greek language and implies difficulty with words. Dyslexia is a sort of specific learning problem (SLD) in which a person struggles to read, spell, and write fluently, despite having normal or above-average intellect. According to the Australian Dyslexia Association (ADA), Dyslexia affects 10% of the Australian population, whereas the figure rises to 20% in other English-speaking countries such as Canada and the United Kingdom. Dyslexic youngsters frequently experience anger, frustration, and low self-esteem. As a result, it is critical to recognize and address dyslexic children's learning challenges as early as possible. Ulrike Kuhl [2] et al. proposed in this system. Developmental dyslexia is a neurodevelopmental learning condition characterized by a substantial loss in literacy skills. However, it is unclear whether current neurobiological theories of dyslexia account for putative deficiency predispositions or the results of limited reading experience. We used functional and structural magnetic resonance imaging techniques to follow 32 children from preliterate to school age across time. Children were divided into 16 dyslexic and 16 control groups based on standardized and age-normed reading and spelling tests conducted at school age. This longitudinal approach enabled us to distinguish probable neurobiological predispositions to dyslexia from the effects of individual differences in literacy experience. In our sample, auditory brain gyrification and abnormal downstream connections

within the speech processing system can predict the issue even before literacy learning occurs. These findings support the theory that dyslexia may result from an abnormal maturation of the speech network prior to literacy education. Developmental dyslexia (DD) is a distinct neurodevelopmental learning condition. Despite normal intellectual abilities, affected individuals struggle greatly with literacy acquisition, resulting in significant educational disadvantages throughout life. Bhavana Srivastava [3] et al. proposed in this system. The internet has been a source of knowledge for decades. The pool of information cannot exist without the network of networks. The internet has many practical applications in business, social, and educational settings. E-learning is currently one of the most beneficial apps on the Internet. E-learning has advanced in a variety of sectors, including adaptive e-learning systems. The combination of computer science and psycholinguistics has done an excellent job in offering technical solutions to students. However, learning difficulties on the e-learning platform continue to require more research. As a result, this research presents a personalized evaluation strategy for alphabet learning using learning objects for children who have dyslexia. The cognitive propensity of dyslexic learners was determined using an evaluation approach. This article investigates the cognitive capacity of dyslexic learners and develops a personalized e-learning platform to address their alphabetical difficulties. The means of getting knowledge via the web has been a crucial mechanism of learning for many years, but it is now recognized as a vital source of information. E-learning not only helps to overcome geographical barriers, but it also aids in the reduction of linguistic barriers. Personalized e-learning has become even more useful. When the system delivers learning content based on learner preferences, the e-learning platform becomes more personalized. For example, for the same subject, learners can choose between a video lecture and an audio lecture. One pupil learns through video, the other through audio. There are several alternative techniques for the machine to learn what the learner desires for obtaining knowledge and understanding of alphabet structures for dyslexic students. For the evaluation, we visited special education institutions for dyslexia in Patna, India, and interviewed the teachers. Experts in these domains have given this research positive feedback based on their studies and observations. However, more learning objects and a feedback mechanism are required to facilitate large-scale experimentation. Jothi Prabha Appadurai [4] et al. proposed in this system. Dyslexia is a learning problem that can make it difficult to read and write. Dyslexia is not a visual disorder; however, many dyslexics have a damaged magnocellular system, resulting in poor eye control. Eye-trackers are devices that track eye movements. This study suggests a set of relevant eye movement variables for use in developing a predictive model for dyslexia. Fixation and saccade eye events are recognized using dispersion-threshold and velocity-threshold methods, respectively. Several machine learning models are tested. Validation is performed on 185 subjects with 10-fold cross-validation. Velocity-based features outperformed statistical and dispersion characteristics in terms of accuracy. The hybrid kernel support vector machine-particle swarm optimizations model achieved the highest accuracy (96%), followed by the extreme gradient boosting model (95%). The optimal set of features are the first fixation start time, average fixation saccade duration, total number of fixations, total number of saccades, and saccade-fixation ratio. Dyslexia is a type of neurodevelopmental disorder that causes chronic difficulties with reading and writing. Sofia Zahia [5] et al. proposed in this system. Dyslexia is a neurological illness that impairs the learning of those who suffer from it, primarily youngsters, and creates difficulties with reading and writing. When dyslexia goes undetected, it causes fear and irritation in the affected youngsters as well as their family circles. Without early intervention, children may enter high school with significant achievement discrepancies. As a result, early detection and intervention services for dyslexic students are critical and advised to help children build good self-esteem and achieve their full academic potential. This research proposes a new technique to the automatic detection of children with dyslexia using functional magnetic resonance. Imaging.: Our proposed method consists of a series of preprocessing processes for retrieving the brain activation zones during three separate reading tasks. The fMRI scans were converted to Nifti volumes, head motion was adjusted, and normalization and smoothing operations were conducted in order to combine all of the subject brains into a single model that allows for voxel

comparisons between subjects. Then, using Statistical Parametric Maps (SPMs), 165 3D volumes encompassing 55 children's brain activation were generated. The classification of these volumes was handled using three parallel 3D Convolutional Neural Networks (3D CNN), each corresponding to a brain activation during one reading task, and concatenated in the last two dense layers, forming a single architecture devoted to performing optimized detection of dyslexic brain activity.

Iii.Existing System

Dyslexia is a neurological condition that causes hurdles and difficulties in the learning process, particularly reading. People with dyslexia typically have poor reading, writing, spelling, and fluency skills. However, these obstacles are unrelated to their IQ. An early diagnosis of this issue will enable dyslexic children to improve their abilities through the use of appropriate tools and specialized software. Machine learning and deep learning algorithms have been used to identify dyslexia using diverse datasets obtained from medical and educational organizations. This review paper examines the predictive effectiveness of deep learning models for dyslexia and summarizes the issues that researchers confront when using deep learning models for categorization and diagnosis. Using the PRISMA procedure, 19 articles were examined and analyzed, with an emphasis on data gathering, preprocessing, feature extraction, and model performance. The goal of this review was to help researchers develop a predictive model for dyslexia based on available dyslexia-related datasets. The report illustrated some of the problems that researchers face in this subject and must overcome.

Iv.Proposed System

The suggested method uses the Random Forest algorithm to improve dyslexia detection by analyzing a large CSV dataset that includes neurological and behavioral characteristics. This method tries to find and classify essential dyslexic biomarkers, with the goal of enhancing early detection accuracy and reliability. Using Random Forests' ensemble learning capabilities, the system overcomes issues such as the necessity for interpretable biomarkers, data privacy problems, and overfitting. The ultimate goal is to create a machine learning model with clinically relevant accuracy, thereby providing an effective tool for early diagnosis and assistance for people at risk of dyslexia.

A. Load Data

The Load Data program imports a CSV dataset comprising neurological and behavioral markers related with dyslexia. This stage entails reading the input into an appropriate data structure for analysis, ensuring that all important information is available for subsequent processing. During this step, data integrity checks are performed to uncover any inconsistencies or missing values that could have an impact on the analysis's quality. The module also includes capabilities for visualizing data distribution, which provides insights into the dataset's features and guides subsequent processing.

B. Data Pre-Processing

The Data Preprocessing module cleans and prepares the dataset for analysis. This includes dealing with missing values, outliers, and noise, which can bias results. Normalizations and standardization techniques are used to ensure that features are on the same scale, which is required for machine learning algorithms to perform effectively. In addition, categorical variables are encoded, and superfluous features are deleted to minimize dimensionality and improve model efficiency. This module lays the groundwork for dependable feature extraction and model training by ensuring that the data is in optimal condition.

C. Feature Extraction

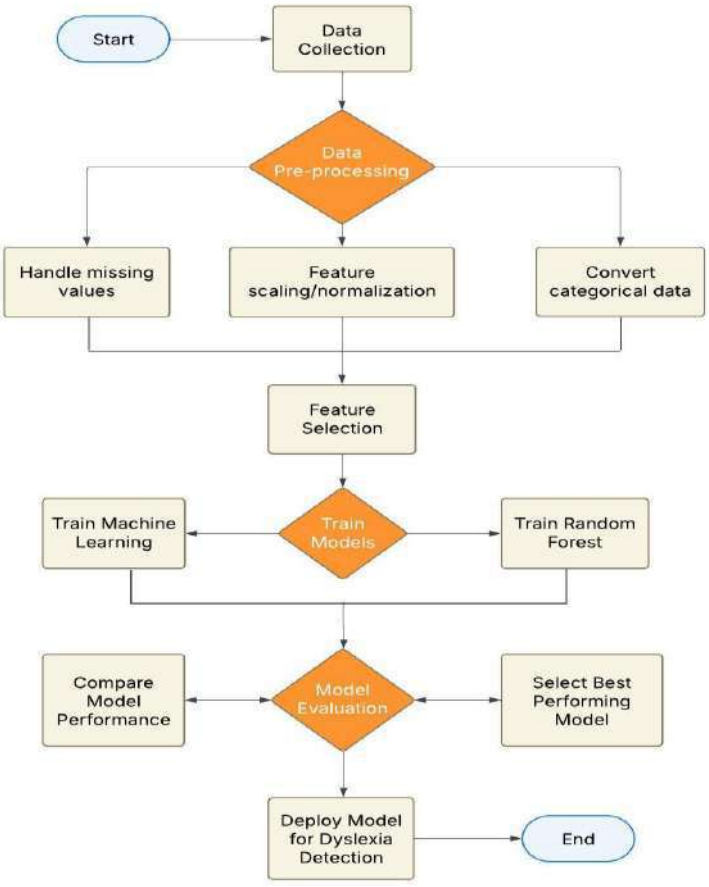
The Feature Extraction module seeks to discover and pick the most relevant features from the pre-processed dataset that can accurately represent the underlying patterns associated with dyslexia. This procedure may use statistical approaches like correlation analysis or principal component analysis (PCA) to minimize redundancy and improve interpretability. This module improves the model's performance by focusing on crucial biomarkers, allowing for a better understanding of the traits that distinguish people with dyslexia from those without. The chosen features are then prepared for input into the machine learning model, ensuring that only the most informative data is used in future training.

D.Training and Testing

The Training and Testing module is responsible for developing and evaluating the Random Forest model. The dataset is divided into training and testing subsets, allowing for a thorough evaluation of the model's prediction ability. During the training phase, the Random Forest algorithm learns from the training data, detecting complex patterns and correlations between the selected features. The testing phase involves applying the trained model to previously unseen testing data to assess its accuracy and generalization performance. This thorough data split ensures that the model is not biassed and can accurately diagnose individuals with dyslexia using the extracted features.

F. Model Evaluation

The Model Evaluation module is crucial for determining the performance of the trained Random Forest model. This module uses a variety of metrics, including accuracy, precision, recall, and F1-score, to assess the model's success in classifying dyslexia. Furthermore, confusion matrices are used to visualize classification results, showing regions where the model excels and where improvements are required. This rigorous evaluation method not only confirms the model's clinical relevance, but also leads ongoing modification and optimizations, guaranteeing that the system is a dependable tool for early intervention and assistance for dyslexic persons.



System Flow Diagram

Algorithm Details

The suggested system uses the Random Forest algorithm, a powerful ensemble learning method, to improve dyslexia detection by categorizing individuals based on neurological and behavioral characteristics. During training, Random Forest constructs several decision trees and outputs the class that represents the mode of the classifications from individual trees. This ensemble strategy decreases the risk of overfitting by building each tree from a random subset of the data, resulting in a more generalized and reliable model. The algorithm's ability to handle high-dimensional data, together with its inherent feature relevance rating, makes it ideal for discovering critical biomarkers related with dyslexia. By pooling the

outputs of several trees, Random Forest improves prediction accuracy and model stability, making it an effective tool for medical diagnosis tasks like as dyslexia detection. Furthermore, its interpretability enables researchers to determine which traits (or biomarkers) contribute the most significantly to categorization, offering useful insights for therapeutic applications. The implementation of Random Forest in this approach achieves a compromise between model complexity and interpretability, resulting in more accurate and trustworthy dyslexia diagnoses.

RESULT ANALYSIS

The result analysis phase critically assesses the dyslexia detection tool's performance and effectiveness, with a focus on the Random Forest model's outputs. This research begins with a thorough examination of the classification metrics, such as accuracy, precision, recall, and F1-score, which together provide information on the model's predictive capabilities. A confusion matrix is used to visually represent the model's performance across several classes, highlighting areas of success and identifying potential misclassifications. Furthermore, feature relevance scores are used to determine which biomarkers make the most significant contributions to the model's predictions, providing vital insights into the underlying traits linked with dyslexia. The research also compares the model's outputs to existing benchmarks and clinical standards to ensure their relevance and trustworthiness. Furthermore, qualitative feedback from end-users, such as physicians and researchers, is used to determine the tool's practical usability in real-world contexts. This extensive assessment of findings not only supports the model's usefulness, but also serves as a direction for future improvements and enhancements, ultimately contributing to better early intervention techniques for dyslexic individuals.

Figure1 Comparison Table

Algorithm	Accuracy
Existing	85
Proposed	90

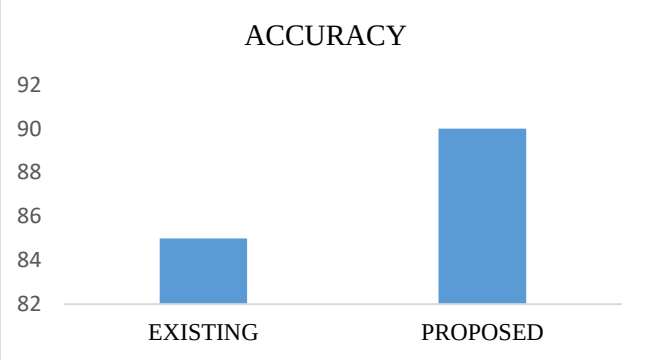


Figure 1 Comparison Graph

CONCUSION

Finally, the suggested dyslexia diagnosis tool, which uses the Random Forest algorithm, shows great promise for improving early identification and intervention options for dyslexic individuals. The system successfully detects critical biomarkers linked with dyslexia using a comprehensive strategy that includes data loading, preprocessing, feature extraction, model training, and evaluation, yielding clinically useful results. The rigorous testing and deployment phases ensure the model's dependability and usefulness for physicians and researchers, while the extensive result analysis demonstrates its prediction accuracy and interpretability. Overall, this tool not only helps to further our understanding of dyslexia, but it also acts as a vital resource for timely support and intervention, with the ultimate goal of improving educational outcomes and self-esteem for affected persons. Future work can expand on this foundation by including larger datasets and investigating sophisticated machine learning techniques, hence improving the system's skills in dyslexia detection and support.

FUTURE WORK

Future work on the dyslexia detection tool might concentrate on a few important areas to improve its functionality and impact. One approach is to broaden

the dataset to include a more diverse population, including different age groups, nationalities, and socioeconomic backgrounds, which can improve the model's generalizability and accuracy across demographics. Furthermore, incorporating more advanced machine learning approaches, such as deep learning algorithms, could improve the model's predictive accuracy by detecting complex patterns in the data. Collaborating with educational institutions and healthcare professionals to perform longitudinal studies will also provide useful information about the tool's long-term usefulness in real-world contexts. Furthermore, incorporating user feedback into iterative design modifications helps improve system usability and functionality, ensuring that it matches the changing needs of doctors and researchers. Finally, investigating the integration of this tool with existing educational frameworks and interventions would allow for a more comprehensive approach to supporting individuals with dyslexia, ultimately resulting in improved educational outcomes and general well-being.

REFERENCE

1. S. Kaisar, "Detecting developmental dyslexia with machine learning: A survey." ICT Exp., vol. 6, no. 3, pp. 181–184, 2020.
2. U. Kuhl, N. E. Neef, I. Kraft, G. Schaadt, L. Dörr, J. Brauer, I. Czepezauer, B. Müller, A. Wilcke, H. Kirsten, F. Emmrich, J. Boltze, A. D. Friederici, and M. A. Skeide. "The emergence of dyslexia in the developing brain." Neuro Image, Vol. 211, May 2020, Art. No. 116633.
3. B. Srivastava & M. T. U. Haider, "Personalised assessment approach for alphabets learning with learning objects in an e-learning environment for dyslexia," J. King Saud University-Compute. Inf. Sci., vol. 32, no. 7, pp. 809-817, 2020.
4. A. Jothi Prabha and R. Bhargavi developed a predictive model for dyslexia based on fixations and saccadic eye movement events. Compute. Methods Programs Biomed., volume 195, October 2020.
5. Dragas, A.S.; Politi-Georgios, S. ICTs as a new method for dyslexia screening: A current perspective. International Journal of Online Biomedical Engineering 2019, 15: 46-60.
6. Prabha, A.J. and Bhargavi, R. Predicting Dyslexia using Machine Learning: A Research Travelogue. Res. Travel. Lecture Notes Electra. Eng. 2019, 556, 1-8.
7. Spoon, K; Crandall, D.; Siek, K. A Neural Network Approach to Detecting Dyslexia in Children's Handwriting. In Proceedings of the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, CA, USA, 9-15 June 2019, pp. 1–5.
8. Isa, I. S.; Rahimi, W. N. S.; Ramlan, S. A.; Sulaiman, S. N. Automated Dyslexia Symptom Detection Using Handwriting Images for Primary School Students. Procedia Computer Science 2019.
9. Shorten, C. and Khoshgoftaar, T. M. A survey about Image Data Augmentation for Deep Learning. J. Big Data 2019, 6:60.
10. Perera, H.; Sirajuddin, M. F.; Wong, K. W. A review of EEG-based pattern classification systems for dyslexia. Brain Inform

HOW TO CITE: Dr. K. N. Sivabalan, Nandhika G., Anupriya S., Divya M., Advance Machine Learning Methods for Dyslexia Biomarker Detection, Int. J. Sci. R. Tech., 2025, 2 (3), 594-600. https://doi.org/10.5281/zenodo.15101904