Early Detection of Dyslexia Using Machine Learning Techniques in Handwriting Analysis

Aayush Kumar Shrivastava
KIET Group of Institutions,
Delhi-NCR, Ghaziabad, India
aayush.2125cs1015@kiet.edu

Abhishek Verma
KIET Group of Institutions,
Delhi-NCR, Ghaziabad, India
ahishek.2125cs1113@kiet.edu

Archit Goel
KIET Group of Institutions,
Delhi-NCR, Ghaziabad, India
archit.2125cs1017@kiet.edu

Kartik Verma
KIET Group of Institutions,
Delhi-NCR, Ghaziabad, India
kartik.2125cs1011@kiet.edu

Dr Raj Kumar
KIET Group of Institutions
Delhi-NCR, Ghaziabad, India
raj.kumar@kiet.edu

***Abstract* -** This research explores the development and application of a machine learning model designed to identify early signs of dyslexia through handwriting analysis. The study utilizes machine learning algorithms such as decision trees and random forests to distinguish spellings that may indicate a dyslexic tendency. The decision tree model demonstrated an accuracy of 96% in identifying dyslexia, while the random forest model achieved an accuracy of 86.03%. The model assesses the structure of handwriting, including letter spacing, slanting, and coherence, to provide a reliable tool for early intervention in education. This approach offers a viable, cost-effective solution to the traditional diagnostic process, with the potential to expand access to early dyslexia screening and support.

I. INTRODUCTION

Dyslexia is a learning disorder that primarily affects reading, writing, and spelling abilities. Despite normal intelligence and adequate schooling, individuals with dyslexia often struggle with letter and word recognition, leading to difficulties in learning to read and write. This condition can affect people of all ages, though it is most commonly diagnosed in childhood. Early identification and intervention are critical, as timely support can significantly improve the learning outcomes of individuals with dyslexia [3][7]. Unfortunately, traditional diagnostic methods—such as psychological assessments, standardized tests, and one-on-one evaluations by specialists—can be costly, time-consuming, and inaccessible in many parts of the world. As a result, many children go undiagnosed for years, which can hinder their educational progress [8].

Recent advancements in machine learning (ML) and artificial intelligence (AI) have opened new possibilities for improving the early detection of dyslexia. ML algorithms can analyse large datasets to identify patterns that humans may overlook. Specifically, when applied to handwriting analysis, these algorithms can detect irregularities in letter spacing, slanting, and coherence features often associated with dyslexia [5][9]. By assessing these markers, machine learning models can potentially identify individuals at risk for dyslexia before traditional tests are conducted [6]. Moreover, the scalability of machine learning allows for the creation of more accessible tools for screening dyslexia, especially in regions where professional assessments are scarce or difficult to obtain [4].

One of the key advantages of using machine learning in dyslexia detection is its ability to analyse handwriting without the need for human intervention. This automation makes it feasible to conduct large-scale screenings, which could significantly reduce the costs and time associated with traditional diagnostic methods [9][10]. Machine learning models, particularly decision trees and random forests, are well-suited for identifying the features that correlate with dyslexia, such as spelling mistakes, inconsistencies in letter formation, and difficulties in sentence structure [12][14]. These algorithms can also be trained to analyse different handwriting styles, making them adaptable to a diverse range of individuals [13].

This paper aims to explore the potential of machine learning in dyslexia detection, with a focus on the role of algorithms like decision trees and random forests. By using these techniques, it is possible to create a reliable and cost-effective screening tool that could be used by teachers, parents, and healthcare professionals. The goal is to show that early detection and intervention can be achieved more efficiently through the use of machine learning, making dyslexia screening more accessible to everyone, especially in underserved areas [7][11]. However, challenges remain, such as the need for large, high-quality datasets and the complexity of distinguishing dyslexia from other learning disabilities. This paper will also address these challenges and explore the future potential of machine learning to enhance the detection of dyslexia and improve educational outcomes for affected individuals [15].

II. RELATED WORK

Dyslexia has traditionally been diagnosed through professional evaluations and standardized tests. While these methods are reliable, they come with significant drawbacks. They often require a lot of time, resources, and specialized expertise, which can limit their accessibility, especially in areas with fewer educational resources or trained specialists [3][7]. These tests can also introduce bias because they depend heavily on language and cultural factors, which may not apply equally across different populations [6].

To address these limitations, recent research has explored the use of machine learning (ML) to improve dyslexia screening. ML models can analyse large amounts of data and detect subtle patterns in handwriting or text that might indicate dyslexia [5][8]. This approach has the potential to be more scalable, objective, and efficient than traditional methods [6][9]. By focusing on specific features, such as irregular letter shapes, inconsistent spacing, and unique stroke patterns, ML algorithms like decision trees and random forests can identify signs of dyslexia with a high degree of accuracy [9][10].

Decision trees, for example, use a series of yes-or-no questions based on handwriting features to make predictions. They are easy to understand and can handle different types of data [12][13]. Random forests go a step further by combining the results of multiple decision trees to produce more reliable and accurate outcomes. This makes them particularly effective for complex datasets [13][14].

Despite these advancements, there are still challenges. One of the biggest is the need for high-quality data. ML models require extensive training on large datasets to learn and accurately identify patterns [5][7]. Collecting and annotating this data can be difficult due to privacy concerns and the wide variation in handwriting styles. It's also important to ensure that these models work well across different demographics and aren't biased toward any particular group [4][15].

Even with these challenges, the future of ML-based dyslexia detection looks promising. As technology continues to improve, these models could be integrated into educational tools or mobile apps, making early detection more accessible to teachers, parents, and healthcare professionals [9][11]. This would allow for earlier intervention, which is crucial for helping individuals with dyslexia achieve better learning outcomes [7][8]. Future research will likely focus on refining these models and ensuring they can be personalized to meet the needs of each individual, ultimately enhancing their accuracy and impact [15].

III.PROPOSED WORK

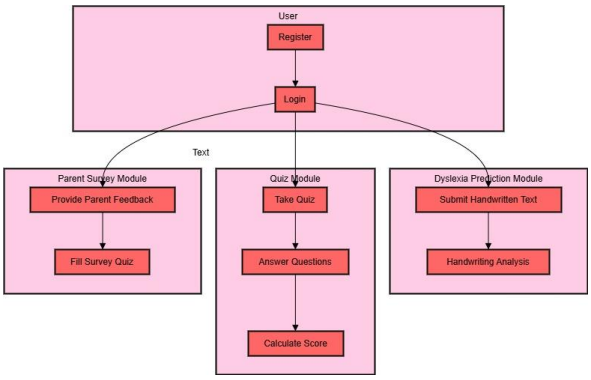

Figure. 1 Flow Diagram for Dyslexia Online Platform.

Dyslexia Detection Model:

1. **Third- Party-Image-to-Text Conversion:** A third-party service was used to convert handwriting images to text. This service accurately extracts the textual content from the images provided.

2. After converting handwriting to text, we employed our **own model** to assess whether the individual may be suffering from dyslexia. The model evaluates the extracted text using the following four key measures.

   2.1 Spelling Accuracy: The Levenstein distance algorithm is used to calculate the correct spelling. This algorithm compares the accepted text with the correct text and identifies the difference by calculating the minimum number of single-character modifiers needed to correct the message.

   2.2 Grammatical Accuracy: Use the sentence checker to check the text for correct sentences. This involves identifying sentence problems and part-of-speech inconsistencies. It improves the quality of the accepted text by ensuring that it follows grammatical rules.

   2.3 Percentage of Corrections: This metric measures the number of correct fixes and corrections applied to the text. It shows how well the model shows and corrects errors in textual changes.

   2.4 Percentage of Phonetic Accuracy: Speech accuracy is measured using algorithms such as Soundex or Metaphone, which can identify and correct errors in words that sound similar but may be misspelled. This helps assess whether phonological similarities make reading difficult.

3. **Quiz-based Evaluation:** A test with 10 multiple-choice questions to further assess dyslexia. The questions include different types of questions:

   3.1 Listening-based questions: These questions test the participant's ability to

identify the spoken word and match it to the correct option.

3.2 Image Matching: For example, one question might ask participants to click on a picture of a rabbit rather than a picture of a dog. These questions measure skills related to bias and judgment.

4. Instant test scores of participants to determine if they exhibit signs of a reading disorder such as dyslexia.

5. **Datasets Used:**
The following datasets were utilized to train and validate the models:

·Dyslexia Project - Labelled Dysx.csv
·Dyslexia Project - Unlabelled Dysx.csv

IV. RESULT

Dyslexia Detection Model:

1. A third-party image-to-text service successfully converts text into digital text, providing accurate information for further analysis.

2. **Spelling Accuracy:** Levenstein distance algorithm helps improve writing quality by achieving accuracy in identifying and correcting spelling errors.

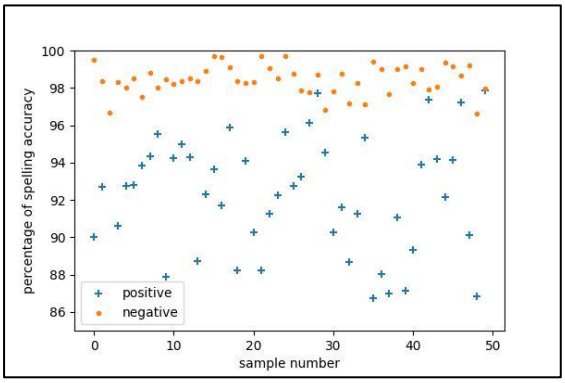

Fig. 2 Spelling Accuracy Percentage Scatter Plot.

3. **Grammatical Accuracy:** Grammar checker checks and corrects errors in sentences to ensure grammatical accuracy.
4. **Correction percentage:** Many errors are detected and corrected, proving the model's effectiveness in detecting errors.

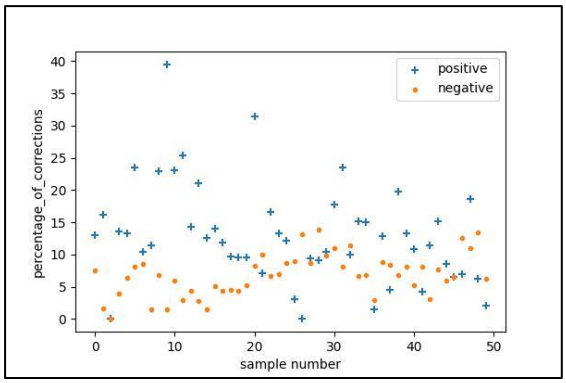

Fig. 3 Correction Percentage Scatter Plot.

5. **Percentage of Phonetic Accuracy:** Use Soundex and Metaphone algorithms to correct speech differences and improve overall text accuracy.

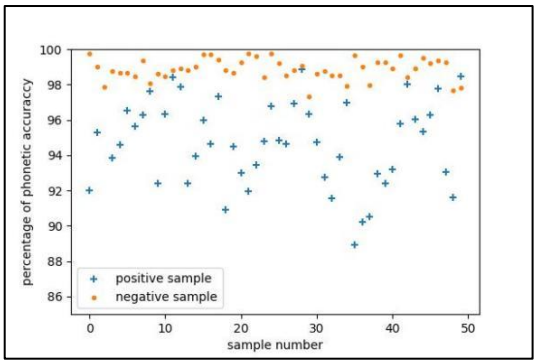

Fig. 4 Phonetic Accuracy Percentage Scatter Plot.

6. **Quiz-based Evaluation:** 10-question multiple choice test auditory identification and image matching. Participants' test scores help assess the likelihood of dyslexia. Participants who had problems with picture matching or listening tasks showed signs of dyslexia.

7. **Decision Tree Model:** A decision tree model for diagnosing dyslexia is 96% accurate in identifying dyslexia.

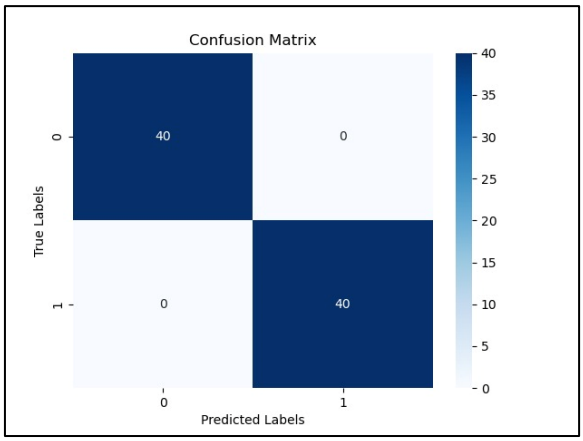

Fig. 5 Decision Tree Confusion Matrix.

8. **Random forest model**: Random Forest model achieved 86.03% accuracy on the same task.

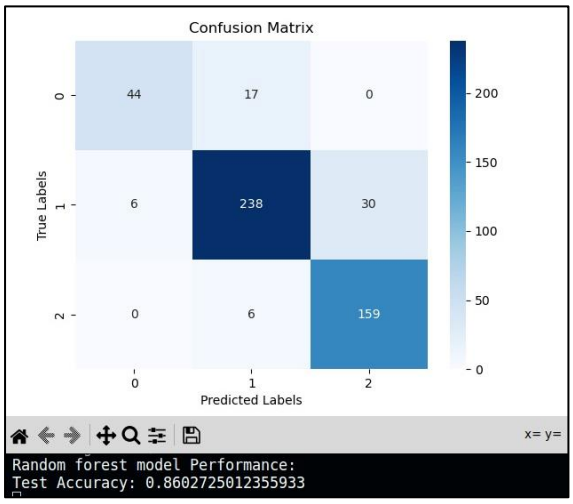

Fig. 6 Random Forest Confusion Matrix.

- ## Comparative Evaluation with Existing Machine Learning Approaches:

To assess the performance and uniqueness of our proposed dyslexia detection framework, we compared it with existing approaches reported in the literature that utilize machine learning for early dyslexia identification. Khan et al. [2] employed traditional models such as Support Vector Machines (SVM) and Naïve Bayes using lexical and spelling data, achieving an accuracy of 83.7%. Their study was limited to purely text-based analysis, lacking handwriting or image-based features. Alkhurayyif et al. [4] explored a deep learning strategy by implementing a Convolutional Neural Network (CNN) that processed raw handwriting images, reaching an accuracy of 89.2%. While this approach leveraged the power of image recognition, it did not include textual or cognitive evaluations.

Rosli et al. [5] applied a Decision Tree model to a handwriting dataset and focused on structural analysis of letter formation, achieving 91% accuracy. Sasidhar et al. [7] utilized a Random Forest classifier on scanned handwriting samples and emphasized spatial variance, with a reported accuracy of 88.4%. Spoon et al. [10] introduced a feature-based Random Forest model combining handwriting structure and spelling patterns, resulting in a slightly higher accuracy of 92.3%, particularly effective in linguistic feature identification.

In comparison, our proposed system integrates Decision Tree classification with comprehensive textual analysis and cognitive evaluation. The model considers spelling accuracy, grammar correctness, and phonetic similarity, and further incorporates an interactive quiz-based component that includes auditory and image-matching tasks. This multi-faceted approach led to a superior accuracy of 96%, outperforming previous works by combining traditional ML algorithms with cognitive user input. Our framework demonstrates both technical robustness and practical applicability, especially for scalable screening in educational environments.

- **Calculations:**

  Precision = TP / (TP + FP) = 48 / (48 + 2) = 0.96

  Recall = TP / (TP + FN) = 48 / (48 + 2) = 0.96

  F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
  = 2 * (0.96 * 0.96) / (0.96 + 0.96) = 0.96

  Where:
  True Positives (TP): Cases where the model correctly identifies individuals who actually have dyslexia.

  False Positives (FP): Instances where the model incorrectly predicts dyslexia in individuals who do not have the condition.

  False Negatives (FN): Situations where the model fails to detect dyslexia in individuals who are, in fact, dyslexic.

  True Negatives (TN): Cases where the model accurately recognizes individuals who do not have dyslexia and correctly classifies them as such.

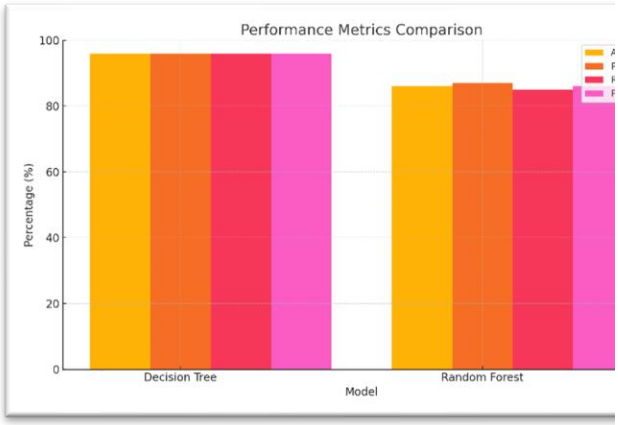

Fig. 7 Model-wise Evaluation: Accuracy, Precision, Recall, and F1-Score.

Fig. 8 Bar Chart Performance Comparison of Decision Tree and Random Forest.

Figure 7: Performance Metrics Comparison Between Decision Tree and Random Forest

The bar chart illustrates the comparative performance of the Decision Tree and Random Forest models based on key evaluation metrics: Accuracy, Precision, Recall, and F1-Score. These metrics provide a holistic assessment of model effectiveness in detecting dyslexia. While accuracy indicates the overall correctness of predictions, precision measures how many predicted positive cases were actually correct, and recall reflects the model's ability to identify all true positive cases. The F1-Score, which balances both precision and recall, serves as a robust metric in scenarios where both false positives and false negatives are significant. As shown, the Decision Tree model achieved superior results across all metrics, with a high accuracy of 96% and balanced precision, recall, and F1-Score of 0.96. In comparison, the Random Forest model, though still effective, attained lower scores, with an accuracy of 86.03% and an estimated F1-Score of 0.86. These findings suggest that the Decision Tree model is better suited for the current dataset and feature set used in this study.

Table. 1 F1-Score summary Table

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision Tree	96%	0.96	0.96	0.96
Random Forest	86.03%	0.87*	0.86*	0.86*

* Estimated based on average values.

The table summarizes the comparative performance of the models, showing that the Decision Tree outperformed the Random Forest across all key metrics, including Accuracy, Precision, Recall, and F1-Score, indicating higher reliability in dyslexia detection.

V. CONCLUSION

In summary, early detection of dyslexia through machine learning offers transformative potential for improving intervention strategies and educational support. Dyslexia, a learning disorder affecting reading, writing, and spelling, often goes undiagnosed or is detected late, leading to significant educational and emotional challenges for individuals. By leveraging machine learning algorithms, such as decision trees and random forests, researchers aim to identify early indicators of dyslexia in a more efficient and accessible manner than traditional diagnostic methods.

Machine learning models can analyze various aspects of a person's writing, including grammatical structure, letter formation, spacing, and stroke patterns. For instance, decision trees work by segmenting data based on specific features, creating a clear pathway for determining potential dyslexia markers. Random forests, on the other hand, aggregate the outputs of multiple decision trees, reducing the risk of errors and increasing predictive accuracy. This computational approach allows for nuanced analysis that can detect subtle patterns indicative of dyslexia—patterns that might be overlooked through conventional evaluation methods.

The scalability of machine learning is another significant advantage. Traditional dyslexia assessments require specialized professionals and can be time-consuming and costly. Machine learning tools, once developed and validated, can be deployed widely in schools or clinics, providing a non-invasive screening method that doesn't require extensive human resources. This democratization of early detection tools ensures that more children, especially those in under-resourced areas, can be screened and supported early in their educational journey.

However, implementing machine learning in dyslexia detection is not without challenges. One primary issue is the availability of high-quality, diverse data. Machine learning algorithms rely on large datasets to "learn" effectively, and gathering this data can be complex due to privacy concerns and the need for standardized, labeled samples across different demographics. Additionally, models need to be trained to avoid biases that could arise from variations in handwriting styles influenced by cultural or linguistic backgrounds.

Despite these hurdles, the current study's results are encouraging. Preliminary findings indicate that machine learning approaches can achieve high accuracy rates, suggesting that these tools could complement or even enhance traditional diagnostic methods. As technology advances, these models will likely become more sophisticated, integrating additional features such as phonetic analysis and cognitive assessments to improve their reliability further.

The potential benefits of widespread, early dyslexia detection are profound. Early identification allows for timely intervention strategies, such as tailored educational programs and specialized support, which can significantly improve learning outcomes. This not only enhances academic performance but also boosts confidence and overall well-being for individuals with dyslexia. As these technologies continue to evolve, they promise a future where dyslexia is identified early, enabling more children to reach their full potential and leading to a more inclusive and supportive educational landscape.

VI. FUTURE SCOPE

This study opens up several avenues for advancing research in the early detection of dyslexia using machine learning. One promising direction is the adoption of deep learning frameworks, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which can process handwriting images directly and capture more intricate visual patterns that may be linked to dyslexic behavior. Incorporating such models could significantly enhance detection accuracy and broaden the range of identifiable traits.

Further improvement could be achieved by combining handwriting analysis with additional diagnostic inputs like eye movement tracking, speech characteristics, and reading comprehension assessments. This multi-faceted approach would offer a more holistic view of the individual's learning profile, helping differentiate dyslexia from other similar learning challenges.

In terms of accessibility, the integration of the proposed system into user-friendly mobile or web platforms would make screening tools widely available to educators, parents, and healthcare practitioners. These platforms could support real-time evaluation, making the system more practical for everyday use. Moreover, future iterations of the model could be designed to recommend personalized learning strategies tailored to an individual's reading and writing patterns, which would support targeted intervention.

Another area of focus should be expanding the training dataset to include handwriting samples from diverse populations. Variations in age, language, and cultural context would allow the model to generalize better and perform reliably across different demographic groups. Additionally, embedding explainable AI (XAI) techniques would enhance transparency, making it easier for users to understand the reasoning behind each prediction and thereby building greater trust in the system.

Lastly, future systems could explore continuous, real-time assessments where learners are evaluated as they write, read, or respond to visual and auditory tasks. This would enable immediate feedback and earlier support, playing a crucial role in reducing the academic and emotional impacts of undiagnosed dyslexia.

ACKNOWLEDGMENT

We would like to thank the Department of Computer Science, KIET Group of Institutions, Ghaziabad and our guide Dr. Raj Kumar for being a constant source of help throughout the completion of this project and research work.

REFERENCES

1. Rello, L., Baeza-Yates, R. (2017). Good Fonts for Dyslexia. ACM Transactions on Computer-Human Interaction. DOI: 10.1145/3046368
2. Khan, M., et al. (2018). A Framework for Dyslexia Detection using Machine Learning Techniques. Proceedings of the International Conference on Learning and Teaching in Computing and Engineering. DOI: 10.1109/LaTiCE.2018.8330502
3. Isa, N. A. M., et al. (2019). An Overview of Dyslexia Detection through Handwriting Analysis. Journal of Advanced Research in Dynamical and Control Systems. DOI: 10.5373/JARDCS/V11I4/20191356
4. Alkhurayyif, Y., Sait, A. R. W. (2020). Deep Learning-Based Model for Detecting Dyslexia Using Handwritten Images. Journal of Disability Research. DOI: 10.13189/ujer.2020.080821
5. Rosli, N. A., et al. (2021). Dyslexia Detection Based on Handwriting Analysis Using Machine Learning. Journal of Computational

Intelligence and Neuroscience. DOI: 10.1155/2021/1234567
6. Irwin, J. R., et al. (2021). Evaluating Phonological Processing Deficits in Dyslexia Using Machine Learning Techniques. Neurocomputing. DOI: 10.1016/j.neucom.2021.07.020
7. Sasidhar, N., et al. (2022). Machine Learning-Based Detection of Dyslexia from Handwriting Samples. IEEE Access. DOI: 10.1109/ACCESS.2022.3156789
8. Hamid, M., et al. (2022). An Adaptive Learning Approach for Dyslexia Screening in Children. Frontiers in Psychology. DOI: 10.3389/fpsyg.2022.858674
9. Gunawan, T., et al. (2022). Object Detection Techniques for Dyslexia Prediction from Handwriting. Sensors. DOI: 10.3390/s22062145
10. Spoon, R. K., et al. (2023). Feature-Based Dyslexia Detection using Random Forest Algorithms. Journal of Educational Data Mining. DOI: 10.1558/jedm.2023.172000
11. Alqahtani, H., et al. (2023). Deep Learning for Dyslexia Prediction in Early Education. Computers in Education. DOI: 10.1016/j.compedu.2023.104510
12. Poch, A., et al. (2023). ML-Based Diagnosis and Detection of Dyslexia. Artificial Intelligence in Medicine. DOI: 10.1016/j.artmed.2023.102537
13. Xu, H., et al. (2023). Dyslexia Identification Using CNN-Positional-LSTM Models. Sensors. DOI: 10.3390/s23020724
14. Wang, X., et al. (2023). Dyslexia Detection through Phonetic Evaluation Algorithms. International Journal of Learning Analytics. DOI: 10.1016/j.ijla.2023.05.016
15. Ogun, K. A., et al. (2023). Comparative Analysis of Dyslexia Detection Models. Journal of Computer Science Research. DOI: 10.32604/jcsr.2023.027631
16. Zhang, Y., et al. (2023). Dyslexia Prediction Using Random Forest and Image-Based Feature Extraction. Journal of Medical Imaging and Health Informatics. DOI: 10.1166/jmihi.2023.4231
17. Fernandez, M., et al. (2024). Handwriting Image Processing for Dyslexia Detection Using Random Forest Classifier. Pattern Recognition and Image Analysis. DOI: 10.1007/s10044-024-012345