IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Evaluating the Diagnostic Value of Unimodal Datasets for Dyslexia Using Machine Learning

Tanvi Srinivasan

Oakridge International School, Varthur Road, Bangalore, Karnataka, 562125, India; tanvi.srinivasan22@gmail.com
Mentor: Dr. Anu Prabha

ABSTRACT: In this study, we evaluated the performance of three machine learning models—Random Forest, Gradient Boosting, and Multilayer Perceptron—across three different dyslexia-related datasets: handwriting images, visual and phonological assessments, and response-based behavioral performance. These datasets not only vary in their cognitive and biological relevance, but also in their collection methods, their accessibility, and their noise content. By focusing on each data type, we aimed to evaluate the contributions of each dataset for dyslexia classification and prediction. All three types of data contain dyslexia signals, but their predictive power varies greatly. Handwriting data provided the highest accuracy, 99.5% from Random Forest, suggesting that motor and visual aspects are strong indicators. Visual and phonological assessment data also performed well, especially with neural network models, reflecting known issues in language processing. In comparison, response-based behavioral data had lower accuracy and more variability, likely due to differences in individual context affecting performance. Overall, these findings support the use of inexpensive and non-invasive data, especially handwriting and phonological measures, for early dyslexia prediction. Despite limitations such as the absence of unified multimodal clinical datasets, this study demonstrates the potential of accessible machine learning approaches to improve diagnostic efficiency and outcomes for individuals with dyslexia.

KEYWORDS: Biology, Neuroscience, Dyslexia, Machine Learning, Unimodal Data.

Introduction

Dyslexia is a neurodevelopmental disorder, also known as 'word blindness', characterized by learning difficulties of reading and writing. It is a dysfunction in learning-related brain systems caused by the interaction of genetic and environmental factors. Problems may include difficulties in reading, storing memory, phonological or language processing, and more, despite conventional instruction and normal intelligence.[1,2] Its etiology is rooted in the left hemisphere of the brain, responsible for the language network, specifically Broca's area (for articulation and speech production), the parieto-temporal region (phonological decoding and letter-sound mapping), and the occipito-temporal region (for rapid word recognition).[3,4]

The global prevalence of the disorder is estimated to be almost 5-10% of the world's population, while the school population contributes to almost 15-20%, making dyslexia the most common learning disability.[5,6] The Dyslexia Association of India reports that 10-15% of children may suffer some form of dyslexia, with over 250 million children, adolescents, and adults suffering some form of learning disability.[7] Across countries, the prevalence ranges from 3.9% to 27% with lifelong implications that may impact education, psychological well-being, and overall quality of life.[8]

Traditional methods of diagnosis in dyslexia are heavily reliant on the reactive approach, thereby missing out on the critical window for the most effective early intervention.[9] The behavioral assessments, though effective, can only be identified post-academic failure of the child, accounting for a huge mental burden both on the child and their family. The diagnosis is often subjective, heavily reliant on the clinical expertise available and behavioral observations, and is frequently delayed or misdiagnosed, leading to an extended and painful diagnostic odyssey.[10] According to the Mayo Clinic, the current paradigm of diagnoses includes questionnaires, comprehension, vision, phonological testing, and academic skills/reading fluency. These methods of identification of cognitive anomalies for early intervention would include standardized tests of phonological processing, Rapid Automatized Naming (RAN), and reading fluency and comprehension.[11-13]

The burden on children and families suffering from neurodevelopmental disorders leads to significant impairments to their quality of life as they struggle with processing language and reading skills needed in their day-to-day life.[8] The need for early detection and the most effective early intervention is an absolute must to alleviate the dyslexia burden from suffering families.[14]

The advent of neuroimaging techniques such as electroencephalography (EEG), magnetic resonance imaging (MRI), and positron emission tomography (PET) has proven to be quite useful in illuminating the neural correlates of neurodevelopmental disorders, enabling objective and pre-symptomatic risk assessment. This has paved the way for the identification of aberrant neural patterns and familial risk for dyslexia early, much before behavioral patterns emerge or formal school education commences.[15,16] The problem now shifts to reliability in understanding the high complexity and dimensionality of this data, subject to clinical expertise and human interpretation.

Machine Learning (ML) and Artificial Intelligence (AI) can be very effectively used to combat the challenges that individual interpretation and clinical expertise can bring in accurate

DOI: 10.36838/IJHSR815.95

diagnosis using neuroimaging techniques for neurodevelopmental disorders. ML models can effectively analyze vast, complex data to identify subtle patterns that are indicative of neurological disorders. These tools have become increasingly useful in analysing and interpreting information, allowing for accurate predictions.[17] The significance of machine learning techniques in early dyslexia prediction through analysis of multimodal data with an accuracy of 98.6% has been shown in the study conducted by Alkhurayyif *et al.*[18] Models such as these can leverage different data types such as electrophysiological signals (EEG/ERPs), structural and functional brain images (fMRI) and behavioral assessments (RAN, phonological tests, questionnaires), each offering a unique lens on dyslexia's underlying mechanisms facilitating unique, directed and accurate early clinical interventions.[19]

One of the glaring limitations is the choice of dataset used to train these models for dyslexia and other neurodevelopmental disorder diagnoses. Unimodal data, as used by Da Silva *et al.* in their study, where only fMRI data is used to train models for dyslexia prediction, clearly demonstrates the limited ability of the model to identify significant dyslexia biomarkers, severely hampering the prediction accuracy. The unimodal approach fails to capture the multifaceted nature and complex mechanisms of neurodevelopmental disorders, which are defined by a distributed network of cognitive and neurological deficits. Every dataset type has its own markers that indicate dyslexia, and the usage of the right ones when training and testing ML models is imperative to ensure high accuracy when used in practice.[20,21]

A study conducted by R. Hernández-Vásquez *et al.*[19] showed that electrophysiological markers (ERPs) and functional/structural MRI both highlight abnormalities in crucial yet different areas in language-related networks (the left hemisphere). Models trained on a unimodal dataset may often lack robustness and generalizability as they are only able to capture a fraction of the variance of the disorder and are inherently flawed for accurate diagnosis.[18] The critical question, therefore, is not only whether an AI can predict dyslexia, but rather it is the understanding of which data modalities can effectively capture the variance of the disorder and provide the most accurate and reliable prediction.

The primary objective of this paper is to systematically evaluate the relative importance of different data modalities in machine learning models for dyslexia detection. Specifically, this study compares the predictive performance of three unimodal data sources: handwriting images, visual and phonological assessment data, and response-based behavioural measures. We hypothesize that the handwriting data would provide stronger predictive signals for dyslexia when analyzed using machine learning techniques.

To investigate this, this study will evaluate multiple machine learning classifiers trained on each dataset to assess the predictivity of each modality. By comparing model performance across these unimodal datasets, this study aims to identify which data types provide the most reliable signals for dyslexia prediction. These findings could contribute to the development of accessible tools that can support earlier identification of dyslexia in educational and clinical settings.

Methods

This study uses machine learning (ML) to evaluate the potential for early dyslexia prediction. Three models were utilized: Random Forest, Gradient Boosting, and Multilayer Perceptron (MLP). These models are the most prevalent across neurological studies, especially those regarding neurodevelopmental disorders, due to their high complexity. These models were used across all datasets tested to conduct a comprehensive analysis to understand the overall accuracy of each dataset type. This would therefore make it possible to gain invaluable insights, allowing for conclusions to be drawn about the dataset type that allows for the most successful early prediction of dyslexia. The methodology follows an organized and logical approach from data collection, preprocessing, feature engineering, model development and selection, training, testing, validation, and evaluation.

Data Collection:

The datasets were collected from Kaggle. Visual and audio assessment data, performance-based behavioral data, and handwriting image data were collected.

Visual and audio assessment data from Kaggle measured 6 parameters based on which a label of 0, 1, or 2 was designated. 0 corresponded to low risk of being dyslexic, 1 to moderate risk of being dyslexic, and 2 to high risk of being dyslexic. The features that were measured in this dataset included vocabulary, memory, speed, visual discrimination, audio discrimination, and survey responses. These parameters all correspond to visual and audio assessment, and therefore provide insight into how accurately this dataset type can predict dyslexia with the use of the different models.[22]

Performance-based data from Kaggle measured and calculated 5 parameters based on the sample's performance on a gamified test. The sample comprised 392 dyslexic individuals (45.2% female, 54.8% male) and 3,252 non-dyslexic people (49.7% female, 50.3% male). Every participant was aged 7 to 17 years. The test included 32 questions that targeted executive functions of activation and attention, and sustained attention. For every question response, the following was calculated: the number of clicks, number of correct answers (hits), sum of hits per set of exercises (score), number of hits divided by the number of clicks (accuracy), the number of incorrect answers (misses), and the number of misses divided by the number of clicks (miss rate). This information provided an understanding of the attention levels of dyslexic and non-dyslexic individuals, allowing for a relation to be drawn by the models.[23]

The handwriting dataset from Kaggle included images for each of the 26 letters of the alphabet under each folder (train, test, and validation). Each folder included 8-100+ images for each letter, likely from different dyslexic or non-dyslexic individuals. This dataset included individual images for each letter; however, there were no clinical labels assigned to classify the handwriting as dyslexic or non-dyslexic. Therefore, K-means clustering was applied to identify latent handwriting pattern

groups, which were used as pseudo-labels for supervised classification.[24]

Data Pre-processing:

Cleaning the Data.

The rows of data in all datasets were checked for missing values, and if found, these rows were dropped to ensure that the model would be able to process the data and accurately predict dyslexia when given the features. This was done using df.isnull() and dropna. Missing values were not found in any of the data sets; they were all complete with every parameter filled for every participant.

Label/Class Distribution:

For all datasets, the imbalance of the class sizes was checked, and the same proportions were kept for both the training and test sets. Imbalance was checked by understanding and plotting on a bar graph the label distribution, as shown in Figure 1. The proportions of the classes and labels were kept constant for the train and test sets to match the original dataset by using stratify=y.

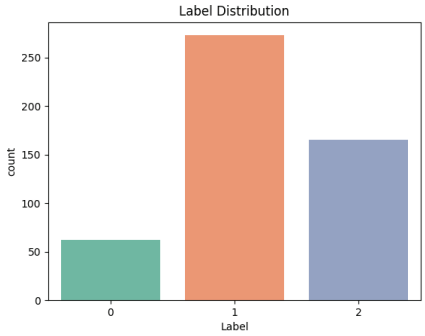


Figure 1: Bar plot of label distribution for visual and audio assessment data. This bar chart displays the distribution of data for the visual and phonological dataset used. The graph provides insights into the variation and constitution of the dataset, which was required to maintain proportions in the train and test sets.

Dimensionality Reduction:

Dimensionality reduction was applied to the handwriting dataset using Principal Component Analysis (PCA) to reduce the high-dimensional features generated by the ResNet50 convolutional neural network (CNN), enabling more efficient and accurate model analysis. Through this, noise was reduced while the important, characterizing features of the images were retained, 95% of the original variance was kept. This was a crucial step for the handwriting dataset to ensure efficiency and accuracy.

Clustering:

Clustering was used for the handwriting dataset, which lacked labels identifying images of handwriting that belonged to a dyslexic or non-dyslexic individual. K-Means was therefore used to observe patterns such as smoothness, shape, and pressure to split the dataset into 2 clusters: 0 (likely non-dyslexic) and 1 (likely dyslexic). After this, the silhouette score was calculated to understand whether the clusters were well-defined and made sense.

Model Development:

Three different models were developed and applied to all the datasets. Traditional machine learning models used include Random Forest and Gradient Boosting. Along with this, a neural network was used: Multilayer Perceptron (MLP). The traditional ML models (n_estimators=100, random_state=42) allowed for the reduction of overfitting and were suitable for the complexity of the data, therefore ensuring that they could capture the complexity of the relationships between parameters. The neural network (hidden_layer_sizes=(100,), max_iter=500, random_state=42) was compared against the traditional models, which was done to assess and analyze the accuracy and performance of prediction. Feature importance plots were generated for each model.

Training, Testing, and Validation:

Model training was performed on 80% of the data, with the remaining 20% used for testing. The proportions of the labels were kept the same for the training and test datasets through stratification. The 80-20 train-test split utilized hold-out validation, which was maintained for all three models used. This approach ensured consistent representation of labels while also ensuring that the model had sufficient data to be trained with, leaving some to be tested. This allowed for a verification of whether the model overfitted on the training data. For all datasets, an 80/20 train-test split was used. Preprocessing steps were implemented using a scikit-learn pipeline to prevent data leakage, ensuring that transformations were fitted only on the training data and then subsequently applied to the held-out test data. Specifically for the handwriting dataset, feature scaling and dimensionality reduction (PCA) were fitted exclusively on the training data, and the learned parameters were then applied to transform the test set.

Software and Tools:

Table 1: Software and tools utilized. To develop the programs and generate graphs and images, multiple tools and libraries were utilized, as shown in this table. These were essential to ensure efficiency and success in data analysis.

Tool/Library	Purpose	Version
Python	Programming language used for data preprocessing, feature engineering, model development, and analysis.	3.13.7
scikit-learn	Importing machine learning algorithms (Random Forest, Gradient Boosting, MLP), train-test splitting, evaluation metrics, feature selection, and preprocessing.	1.7.2
pandas	Used for data manipulation and preprocessing.	2.3.3
NumPy	Used to process the features and for numerical calculations.	2.3.3
Matplotlib	Generation of plots: confusion matrices, ROC curves, and other visualizations.	3.10.6
Seaborn	Visualization of statistical data through bar plots, heatmaps, and distribution plots.	0.13.2
torch / torchvision	To capture patterns in the image handwriting dataset.	2.9.0 / 0.24.0

The code pipeline used for data preprocessing, feature extraction, model training, and evaluation, along with the environment specifications, can be made available upon request by contacting the authors.

DOI: 10.36838/IJHSR815.95

Results

Three dyslexia datasets were evaluated through three different models: 2 traditional ML models (GB and RF) and a neural network (MLP). The accuracy of the models, as well as other evaluation metrics (precision, recall, F1), was compared to understand the variance between datasets in dyslexia prediction and to understand which features are most informative for distinguishing between groups of data. The handwriting dataset utilized weak clustering due to the lack of clinical labels, and therefore, the parameters were calculated based on how well the models could discriminate between the two groups of data that were created.

Handwriting:

Table 2: Classification performance metrics for the handwriting dataset. This table showcases the performance of the handwriting dataset across the three models tested. The parameters measured include accuracy, precision, recall, and weighted F1-score.

Model	Accuracy	Precision	Recall	F1-score (weighted)
MLP	0.970	0.970	0.970	0.970
Gradient Boosting	0.985	0.985	0.985	0.985
RF	0.995	0.995	0.995	0.995

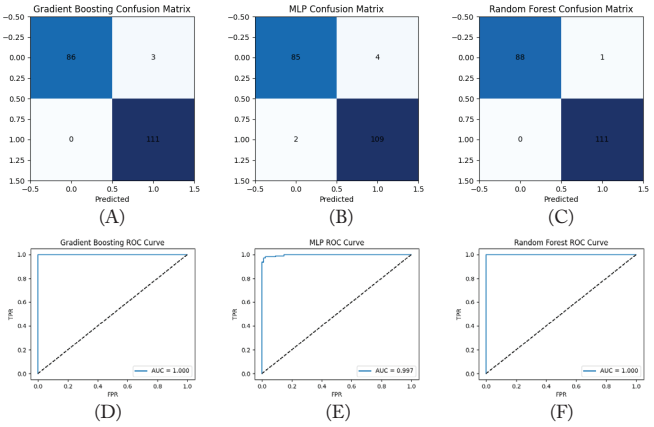


Figure 2: Model Performance (A) Gradient Boosting - Confusion Matrix (B) MLP - Confusion Matrix (C) Random Forest - Confusion Matrix (D) Gradient Boosting - ROC Curve (E) MLP - ROC Curve (F) Random Forest - ROC Curve. This figure displays the confusion matrices and ROC curves for each model. These graphs were created to understand exactly how well the data allows for classification.

The handwriting dataset consistently achieved high accuracies across all models, which is visible in Table 2, with the lowest accuracy being 97% for MLP and the highest being 99.5% for RF. Gradient Boosting performed slightly better than MLP, reaching 98.5% across all evaluation metrics, suggesting that its ensemble structure captured the dataset's patterns more effectively. The Random Forest model performed the best overall, obtaining 99.5% for accuracy, precision, recall, and F1-score. This suggests that RF had no issues with overfitting. Overall, the consistent numbers observed in the parameters suggest that the data were very easy to separate based on the images' characteristics. The significant number of images of handwriting proved sufficient for the models to identify patterns in the images to be able to distinguish between the groups created through clustering. The confusion matrices and ROC curves displayed show that RF had the best true negatives, no missed dyslexic cases, and only one false positive.

Visual and Phonological:

Table 3: Classification performance metrics for the visual and phonological datasets. This table displays the performance of the visual and phonological dataset across the three different models. The parameters measured include accuracy, precision, recall, and weighted F1-score.

Model	Accuracy	Precision	Recall	F1-score (weighted)
MLP	0.980	0.981	0.980	0.980
Gradient Boosting	0.930	0.930	0.930	0.930
RF	0.960	0.960	0.960	0.960

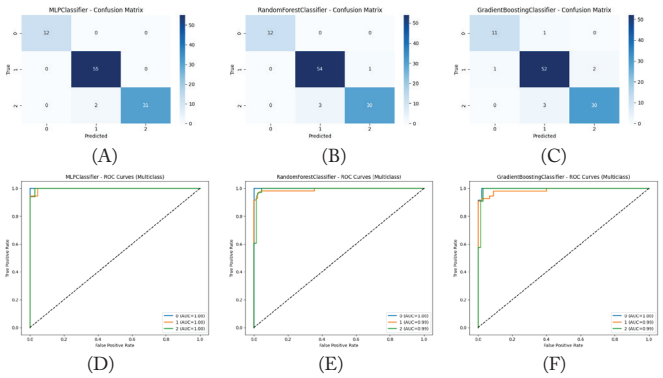


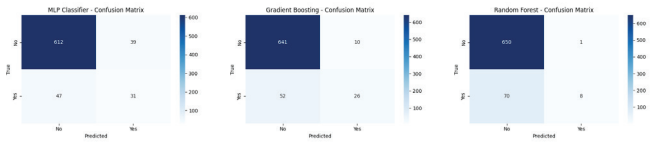
Figure 3: Model Performance (A) MLP Classifier - Confusion Matrix (B) Random Forest Classifier - Confusion Matrix (C) Gradient Boosting Classifier - Confusion Matrix (D) MLP Classifier - ROC Curves (Multiclass) (E) Random Forest Classifier - ROC Curves (Multiclass) (F) Gradient Boosting Classifier - ROC Curves (Multiclass). This figure displays the confusion matrices and ROC curves for each model. These graphs were created to understand exactly the number of errors and correct predictions for the visual and phonological data.

In the visual and phonological dataset, MLP achieved the strongest performance, with 98% accuracy and correspondingly high precision, recall, and F1-scores. This is evident in Table 3. The confusion matrix and ROC curve in Figure 3 show perfect classification for two of the three classes, with only two samples from the third class being misclassified, indicating strong separation between feature groups. RF performed almost equally as well, with a 96% accuracy, and had 2 more misclassified samples compared to MLP, as seen in the confusion matrix. Gradient Boosting performed lower than both, with an accuracy of 93%. There were misclassified samples under all three classes/labels.

Response Based:

Table 4: Classification Performance Metrics for Different ML Models for Response-based Dataset. This table displays the performance of the response-based dataset across the three different models. The parameters measured include accuracy, precision, recall, and weighted F1-score.

Model	Accuracy	Precision	Recall	F1-score (weighted)
MLP	0.882	0.877	0.882	0.880
Gradient Boosting	0.915	0.903	0.915	0.901
RF	0.903	0.902	0.903	0.866



(A) (B) (C)

Figure 4: Model Performance (A) MLP Classifier - Confusion Matrix (B) Gradient Boosting - Confusion Matrix (C) Random Forest - Confusion Matrix. This figure displays the confusion matrices for each of the three models. These graphs were created to understand exactly the number of errors and correct predictions for the response-based data.

In the response-based dataset, as we can see in Table 4, the values for the three models were varied and slightly low. MLP achieved the weakest performance, with an accuracy of 88.2% and similarly low precision, recall, and F1-score values. The confusion matrix in Figure 4 shows that the MLP frequently misclassified both classes. It produced 39 false positives and 47 false negatives; this model struggled to separate data into the classes. The Random Forest model achieved 90.3% accuracy, but it showed the lowest weighted F1-score (0.866) due to an imbalance in how it classified the two groups. RF produced 70 false negatives and only 1 false positive, showing that the model struggled to detect samples part of the "Yes" label correctly. Gradient Boosting performed the strongest overall on this dataset, achieving 91.5% for accuracy, precision, and recall, with a slightly lower weighted F1-score of 0.901. Its confusion matrix shows far fewer mistakes than MLP and RF, misclassifying only 10 non-dyslexic cases and 26 dyslexic cases, suggesting that GB was better at capturing the nonlinear structure of the response patterns.

Discussion

Comparative Predictive Power:

Our findings clearly indicate that the three different data modalities, though all carrying dyslexia specific signals, had very different predictive power. The Random Forest algorithm applied to handwriting samples yielded the highest prediction accuracy of 99.5%, surpassing the accuracy obtained from neuroimaging-based models (PMC11232624).[6] This aligns with more recent research that illustrates that there are specific and marked discrepancies, such as inconsistent letter size, spacing differences, and reversals within dyslexic handwriting that lead to near-perfect classification. The accuracy on the visual and phonological variables was also high and reflective of strong prediction skills (~98% accuracy on MLP), representative of fundamental language network dysfunction, as opposed to metrics tied to response variables (clicks, accuracy, misses), which were suboptimal (~88-91%) and had a higher variability/ noise due to task performance variability. In summary, variables that are associated more closely with motor/ language systems (handwriting and phonology) yielded more robust models when compared to pure performance or attention-based variables.

Classifier Performance and Feature Importance:

While comparing classifiers, we observed that ensemble tree-based classifiers (Random Forest, Gradient Boosting) generally performed the best. Random Forest achieved the highest accuracy of 99.5% due to its robustness against overfitting.[26] The deep neural network (MLP) performed exceptionally well on the dense visual and phonological data with 98% accuracy, but did perform poorly on sparse response data, suggesting its sensitivity to sample size and feature complexity. These observations mirror findings in related fields. Raatikainen *et al.* achieved an accuracy of 89.7% using eye-tracking data with a Random Forest-SVM hybrid, and Alkhurayyif *et al.* used a deep learning multimodal approach, reporting a 98.6% accuracy.[18,25] The handwriting model achieved an accuracy of 99.5%, but this should be interpreted keeping in mind that our handwriting labels were derived from unsupervised clustering rather than clinical diagnosis.

We have extracted feature importances within these models to emphasize key predictors such as phonological speed and accuracy metrics, which emerged as top features in the visual and phonological dataset, consistent with known strong markers of dyslexia (Letter-sound mapping deficits).[27]

The ROC AUC values were near perfect, which indicates strong and effective dyslexia classification performance for the visual and phonological dataset as well as the handwriting dataset. The ROC curves and AUC values were computed using predictions on the held-out test set. The misclassifications visible in the confusion matrices can be justified as the final predictions depend on a fixed probability threshold, whereas ROC AUC evaluates performance across all possible thresholds. The relatively small dataset size amplifies the discrepancy, as it could have inflated the AUC values.

Dataset Characteristics:

Our datasets differed in their method of collection and the quality of the data itself. The handwriting images were easily acquired, where children simply wrote letters requiring minimal to no instrumentation, but in this case, lacked clinical labels. The handwriting dataset used in this study does not provide metadata identifying individual writers. As a result, it is not possible to control writer-level separation between training and test sets. This limitation does introduce the possibility that stylistic similarities from the same writer could appear in both sets. The silhouette score after applying K-means clustering was 0.1614, which indicates weak but present cluster structure. This indicates weak but present cluster structure, suggesting that handwriting features contain detectable variation within the dataset, although the separation between groups is not strongly defined. Nevertheless, the models trained on this data achieved high predictive performance, suggesting that ML algorithms were able to capture subtle patterns. The visual/ phonological dataset consisted of questionnaires and test scores of vocabulary, memory, sensory discrimination, etc., similar to standard dyslexic clinical screening.[8] These features were carefully normalized (as the case may be) so that speed, accuracy, and memory scores were on comparable scales across individuals, a necessary step given their different units. The response-based data came from an online attentional task with thousands of participants, providing statistical power but also introducing noise - for instance, motivation and fatigue. In this study's analysis of the imbal-

DOI: 10.36838/IJHSR815.95

anced response-based dataset, standard evaluation metrics (Accuracy, Precision, Recall, F1-score, and ROC AUC) were used to maintain consistency across the datasets examined in this study, which allowed for efficient comparison. Nevertheless, the use of these metrics may not fully account for class imbalance. Future work could incorporate imbalance evaluation metrics such as balanced accuracy, AUROC, AUPRC, and MCC, as well as class-weighting or resampling strategies. We stratified the train/ test split in order to maintain class balance. In summary, each modality entails trade-offs: handwriting and visual tasks are fast and non-invasive, while response-based measures are inexpensive yet indirect.

Clinical and Practical Implications of Early Screening and Technology:

Our results emphasize how non-invasive assessments can successfully identify a risk for dyslexic issues in a child. For patients and families, these methods can be used as a first-layer of intervention, causing little to no distress to parents or children. Take an example of a simple exercise, such as a brief handwriting exercise or a short computerised phonology game, which individually might take a couple of minutes but prove to be strong indicators in their diagnosis using a simple forecasting software predicting the likelihood of dyslexia.[25] This is less dependent on clinical assessment, giving families early indicators of poor learning. Early identification means early intervention, which is known to significantly impact the quality of life as well as the self-esteem of suffering children and families.[8] In practice, parents who can identify these early signs of dyslexia/ learning difficulties could use computerized screening tools (such as our model) at home, making this a less emotionally taxing process (given the model has clinical labels and is trained on the same).

These findings also indicate that the combination of machine learning tools with standard clinical practice could dramatically improve the diagnostic accuracy and efficiency. Clinicians may focus efforts on obtaining the most informative data. Based on our analysis, handwriting samples and standard phonological measures gathered alone can direct the clinician to quickly assess the risk for dyslexia.[25] Traditional steps of clinical evaluation, such as sight word tests or RAN tasks, can then be reserved for those cases requiring confirmation in suspect cases. Machine learning can also give objective rankings to features, providing the clinician with evidence for specific deficits such as phonemic memory vs. visual discrimination, thereby guiding directed and personalized clinical interventions or clinical confirmatory tests. AI tools can therefore act as a 'second opinion' or pre-screen that reduces the number of people who may actually need expensive imaging or comprehensive testing, thus shortening what is often a many-year-long 'diagnostic odyssey' that suffering families face.

This research, therefore, is a demonstration of how technology can assist in the diagnostic journey and contribute to a better quality of life for suffering children or families with dyslexia. It is also an exemplification of how 'everyday' data provides enough signal for a reliable prediction. By automating early screening/pre-screening, we can reduce the stigma and stress that surround the diagnosis of dyslexia. Our feature analysis also provides a biological explanation, such as the prevalence of strong signals in handwriting tasks, highlighting that the left-hemisphere language network attributes more to someone with dyslexia. This convergence of computational results with the earlier studies in neuroscience strengthens confidence in the validity and importance of such studies in dyslexia and other neuro-developmental disorders.

Limitations and Future Directions:

Our study certainly had limitations. This study utilizes data that is primarily open-sourced, with no access to different facets of data for the same subjects in the study. We did not have brain imaging/EPS data for our subjects, so we cannot check our hypothesis concerning the preeminence of neuroimaging. Other researchers have demonstrated that multimodal models, fMRI/MRI/EEG, can achieve a vastly high degree of accuracy (about 98% to 99%).[27] Incorporating such data would have provided a much richer biological grounding covering all aspects of causation and measurement that are imperative to dyslexia identification. Unfortunately, no single existing dataset collects handwriting, cognitive scores, and imaging from the same individuals, ideal for such analysis and unified modelling experiments. The scarcity of large, multimodal dyslexia datasets is an obstacle well noted by others working in the domain.[27] For future work, we plan to use such emerging datasets with verified/available clinical labels that link behavioral and neural markers, enabling evaluation of cross-modal synergies. We also used a hold-out validation approach, applying cross-validation or external replication that would strengthen claims of generalizability and rule out overfitting.[13]

Letter categories were not explicitly stratified across splits. This means that some letters may be slightly over- or under-represented in each split. Therefore, the model may partially learn patterns related to specific letter shapes rather than only dyslexia-related handwriting characteristics. Future studies should control the distribution of letters across splits to reduce this potential bias.

Model evaluation in this study was performed using a single 80/20 train-test split. Since the datasets used in this study contain limited sample sizes, this constrained the use of cross-validation. The lack of publicly available datasets for the data types being used here prevented external validation. Future work should incorporate larger datasets and utilize cross-validation and external validation to provide more robust estimates of model generalizability.

Simpler models, such as logistic regression or linear SVM, were not explored in this study, and systematic tuning procedures were not implemented. This is because the primary aim of this work was to evaluate the comparative predictive capacity of different data modalities rather than to optimize model performance. Future work could incorporate simpler baseline classifiers and more complex hyperparameter tuning to further validate the results.

Lastly, although our models are robust, the application of our models to real-world scenarios is still necessary. This is because the labeling of the handwriting dataset in this research is

created from clusters, but this is only a surrogate for real dyslexic patients. A real, clinically labelled handwriting dataset (or a mobile application to obtain writing samples from patients with known diagnoses) can be used for supervised learning to validate this hypothesis. The reported high classification performance reflects the separability of handwriting pattern clusters rather than validated diagnostic detection of dyslexia. Due to the lack of clinically labelled handwriting datasets, pseudo-labels were used; therefore, the results should be interpreted as exploratory. Moreover, real-world application for fair treatment (independent of language and demographics) is also a requirement because patients with different types of writing systems may have different types of dyslexic symptoms. Despite this, research such as ours is urgently required because, by far, a diagnosis of a patient with a learning disorder such as dyslexia has the longest path compared to other learning disorders. Our research indicates that when cheap, fast data is used, diagnoses with high accuracy are possible, which in turn leads to shortening of the diagnostic odyssey.

■ Conclusion

In conclusion, our research supports a targeted diagnosis approach, where patients are initially assessed with high-yield, low-cost tests (handwriting, phonological skills), with further comprehensive evaluation (imaging, etc.) offered only when necessary, thereby enabling a more directive and personalized diagnostic journey for affected individuals.

■ Acknowledgments

I would like to thank my professor, Dr. Anu Prabha, for her constant guidance and support throughout this process. I am also grateful to my parents and my school, without whose encouragement and help this work would not have been possible.

■ References

1. Hulme, C.; Snowling, M. J. Reading Disorders and Dyslexia. *Curr. Opin. Pediatr*. 2016, *28* (6), 731–735. https://doi.org/10.1097/MOP.0000000000000411.
2. Lyon, G. R.; Shaywitz, S. E.; Shaywitz, B. A. A Definition of Dyslexia. *Ann. Dyslexia* 2003, *53* (1), 1–14. https://doi.org/10.1007/s11881-003-0001-9.
3. Shaywitz, S. E.; Shaywitz, B. A.; Fulbright, R. K.; Skudlarski, P.; Mencl, W. E.; Constable, R. T.; Pugh, K. R.; Holahan, J. M.; Marchione, K. E.; Fletcher, J. M.; Lyon, G. R.; Gore, J. C. Neural Systems for Compensation and Persistence: Young Adult Outcome of Childhood Reading Disability. *Biol. Psychiatry* 2006, *57* (11), 1251–1262. https://doi.org/10.1016/j.biopsych.2005.10.038.
4. Peterson, R. L.; Pennington, B. F. Developmental Dyslexia. *Lancet* 2012, *379* (9830), 1997–2007. https://doi.org/10.1016/S0140-6736(12)60198-6.
5. Shaywitz SE, Shaywitz JE, Shaywitz BA. Dyslexia in the 21st century. Curr Opin Psychiatry. 2021 Mar 1;34(2):80-86. doi: 10.1097/YCO.0000000000000670. PMID: 33278155.
6. Wagner RK, Zirps FA, Edwards AA, Wood SG, Joyner RE, Becker BJ, Liu G, Beal B. The Prevalence of Dyslexia: A New Approach to Its Estimation. J Learn Disabil. 2020 Sep/Oct;53(5):354-365. doi: 10.1177/0022219420920377. Epub 2020 May 26. PMID: 32452713; PMCID: PMC8183124.
7. Robaa, M.; Balat, M.; Awaad, R.; Omar, E.; Aly, S. A. Explainable AI in Handwriting Detection for Dyslexia Using Transfer Learning. *arXiv* 2024, 2410.19821v1. https://arxiv.org/abs/2410.19821 (accessed Dec 2025).
8. Huang, Y.; Liu, D.; Wang, Y.; Li, Y.; Zhang, X.; Yang, X.; Wu, H.; Chen, X. Personality, Behavior Characteristics, and Life Quality of Primary Students with Dyslexia. *Int. J. Environ. Res. Public Health* 2020, *17* (5), 1559. https://doi.org/10.3390/ijerph17051559.
9. Gabrieli, J. D. E. Dyslexia: A New Synergy between Education and Cognitive Neuroscience. *Science* 2009, *325* (5938), 280–283. https://doi.org/10.1126/science.1174705.
10. Wu, Y.; Cheng, Y.; Yang, X.; Yu, W.; Wan, Y. Dyslexia: A Bibliometric and Visualization Analysis. *Frontiers in Public Health* 2022, *10*, 915053. https://doi.org/10.3389/fpubh.2022.915053
11. Pennington, B. F.; McGrath, L. M.; Peterson, R. L. *Diagnosing Learning Disorders: A Neuropsychological Framework*, 3rd ed.; Guilford Press: New York, 2019.
12. Mather, N.; Wendling, B. J. *Essentials of Dyslexia Assessment and Intervention*, 2nd ed.; Wiley: Hoboken, NJ, 2024.
13. Wilmot, A.; Pizzey, H.; Leitão, S.; Hasking, P.; Boyes, M. Growing Up with Dyslexia: Child and Parent Perspectives on School Struggles, Self-Esteem, and Mental Health. Dyslexia 2023. https://doi.org/10.1002/dys.1729.
14. Shofiah, N.; Putera, Z. F. Importance of Early Literacy Intervention for Children with Dyslexia. In Proc. First Conf. Psychol. Flourishing Humanity (PFH 2022); Atlantis Press, 2023; pp 42–56.
15. Ridzwan, E. H. M.; Husni, H.; Na'in, N. M. Designing for Dyslexia: Challenges and Insights into Interaction Design in Augmented Reality. International Journal of Information and Education Technology 2025, 15 (6), 1211. https://doi.org/10.18178/ijiet.2025.15.6.2324
16. Vanderauwera J, Altarelli I, Vandermosten M, De Vos A, Wouters J, Ghesquière P. Atypical Structural Asymmetry of the Planum Temporale is Related to Family History of Dyslexia. Cereb Cortex. 2018 Jan 1;28(1):63-72. doi: 10.1093/cercor/bhw348. PMID: 29253247.
17. Alkhurayyif, Y.; Sait, A. R. W. A Review of Artificial Intelligence-Based Dyslexia Detection Techniques. Diagnostics 2024, 14, 2362. https://doi.org/10.3390/diagnostics14212362
18. Alkhurayyif, Y.; Sait, A. R. W. Deep learning-driven dyslexia detection model using multi-modality data. PeerJ Computer Science 2024, 10, e2077. https://doi.org/10.7717/peerj-cs.2077
19. Hernández-Vásquez R, Córdova García U, Barreto AMB, Rojas MLR, Ponce-Meza J, Saavedra-López M. An Overview on Electrophysiological and Neuroimaging Findings in Dyslexia. Iran J Psychiatry. 2023 Oct;18(4):503-509. doi: 10.18502/ijps.v18i4.13638. PMID: 37881421; PMCID: PMC10593994.
20. Cara C, Zantonello G, Ghio M, Tettamanti M. Multimodal investigation of the neurocognitive deficits underlying dyslexia in adulthood. Cereb Cortex. 2025 Oct 2;35(10):bhaf193. doi: 10.1093/cercor/bhaf193. PMID: 41052271; PMCID: PMC12499769.
21. Velmurugan, S. Predicting Dyslexia with Machine Learning: A Comprehensive Review of Feature Selection, Algorithms, and Evaluation Metrics. Journal of Behavioral Data Science 2023, 3 (1), 70–83. https://doi.org/10.35566/jbds/v3n1/s
22. thenikhilnj45. Dyslexia Project—labelled_dysx.csv. Kaggle Dataset. https://www.kaggle.com/datasets/thenikhilnj45/dyslexiaproject (accessed Dec 2025).
23. Rello, L. Predicting Risk of Dyslexia (dataset). Kaggle. https://www.kaggle.com/datasets/luzrello/dyslexia (accessed Dec 2025)
24. Sumitaich. Dyslexia Datasets (dataset). Kaggle. https://www.kaggle.com/datasets/sumitaich/dyslexia-datasets (accessed Dec 2025).

DOI: 10.36838/IJHSR815.95

25. Raatikainen, P.; Hautala, J.; Loberg, O.; Kärkkäinen, T.; Leppanen, P.; Nieminen, P. Detection of Developmental Dyslexia with Machine Learning Using Eye Movement Data. Array 2021, 12, 100087. https://doi.org/10.1016/j.array.2021.100087
26. Snowling MJ, Hulme C, Nation K. Defining and understanding dyslexia: past, present and future. Oxf Rev Educ. 2020 Aug 13;46(4):501-513. doi: 10.1080/03054985.2020.1765756. PMID: 32939103; PMCID: PMC7455053.
27. Dyslexia—Diagnosis and Treatment. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/dyslexia/diagnosis-treatment/drc-20353557 (accessed Dec 2025).

Author

Tanvi Srinivasan is a Grade 11 IB student at Oakridge International School, Bangalore. She studies Biology and Chemistry at Higher Level and is deeply interested in neuroscience, particularly neurodegenerative diseases. She has completed related courses and taught students globally, aiming to explore the intersection of neuroscience and technology.