

LDDetect: A Machine Learning Based System for Detection and Classification of Specific Learning Disabilities

Dr. Leena H. Patil¹, Aditya Singh², Aman Mishra³, Harish Pagarwar⁴, Ayush Pawar⁵, Vranda Rahangdale⁶
^{1,2,3,4,5,6}*Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur*
(An Autonomous Institution, Affiliated to R.T.M. Nagpur University)

Abstract—Learning disabilities such as Dyslexia and Dysgraphia affect 15–20% of students globally, yet remain largely undiagnosed due to costly clinical evaluation and a lack of automated screening tools. Early detection is essential because timely intervention directly improves academic outcomes and long-term learning support. This study reviews existing machine learning–based approaches for learning disability detection and proposes LDDetect, a web-based system for screening Dyslexia and Dysgraphia. The system collects handwriting and reading-assessment data through a web interface, extracts behavioral features, and classifies risk using Random Forest and Support Vector Machine (SVM) classifiers, comparing their performance to select the most accurate model for each disorder, followed by personalized exercise recommendations. The proposed approach is expected to offer an affordable and accessible screening alternative suited to schools and educational environments with limited access to specialists, diagnostic infrastructure or clinical resources.

Index Terms—Learning Disabilities, Dyslexia, Dysgraphia, Machine Learning, Random Forest, SVM, Early Screening, Web-based Assessment

I. INTRODUCTION

Specific Learning Disabilities (SLDs) are neurodevelopmental conditions that impair a child's ability to read, write, spell, or perform mathematical calculations despite otherwise normal intelligence [1]. Dyslexia, dysgraphia, and dyscalculia are commonly identified as the most frequent SLD subtypes encountered in school-age populations [6]. Reported estimates indicate that approximately 10-15% of children worldwide are affected by some form of SLD [5]. Early diagnosis is important because delayed

identification is linked to poor academic performance, low self-esteem, and long-term psychosocial difficulties [1]. Conventional diagnosis is based on standardized intelligence, achievement, visual-motor, and language assessments administered by trained professionals [6]. Such manual evaluation is time-consuming, depends on scarce specialist resources, and is difficult to scale across large student populations [2]. As a result, Machine Learning (ML) has been increasingly explored as a faster and more cost-effective alternative to conventional clinical screening [1]. ML models can be trained on behavioral signals, such as handwriting strokes, eye-tracking measurements, EEG recordings, and reading/spelling responses, to automatically flag children who may be at risk of an SLD [3]. Classical algorithms such as Support Vector Machines (SVM), random forests (RF), Decision Trees, and ensemble methods, including AdaBoost and Gradient Boosting, have been used for this purpose, with reported accuracies ranging from approximately 80% to above 99% [4], [5], [7], [10]. Handwriting-based screening is attractive because data can be captured using an ordinary digitizing tablet rather than specialized clinical hardware [5]. Eye-tracking-based screening exploits the finding that dyslexic readers exhibit longer fixations and more frequent regressive saccades than typical readers do [4]. Multimodal platforms that combine more than one assessment, such as reading, handwriting, and listening tasks, have also been proposed to widen diagnostic coverage within a single system [2]. Despite this progress, most existing ML-based systems remain limited by small and narrow datasets, a single disorder focus, and little to no integration with personalized remedial support [1].

Therefore, this review (i) surveys recent classical ML-based approaches to SLD detection organized by data modality, (ii) consolidates the resulting research gaps, and (iii) uses comparative accuracy evidence from the literature to justify Random Forest as the primary candidate algorithm for the proposed screening system.

II. LITERATURE REVIEW

Table I summarizes the 12 ML-based studies reviewed in this paper, indicating the data modality, algorithm(s) used, dataset size, and the best reported accuracy in each case.

Table I. Comparative summary of reviewed studies

Study	Modality	Best Algorithm	Accuracy
Zaibi & Bezine [5]	Handwriting	Gradient Boosting / RF	99%
Ahire et al. [7]	EEG	SVM	97%
Modak et al. [11]	E-learning behaviour	Random Forest	96.4%
Vakadkar et al. [12]	Screening checklist	Random Forest	95%
Joshi et al. [2]	Multimodal (eye, hand, speech)	Trained classifier	93.6%
Raatikainen et al. [4]	Eye-tracking	RF + SVM	89.7%
Kaisar (survey) [3]	Eye-tracking / EEG	SVM (most common)	80-97%
Khan et al. [8]	Reading test	K-NN	n/a*
Dutta & Sait [9]	Speech	Gradient Boosting + SVM	High**
Drotár & Dobeš [10]	Handwriting	AdaBoost	~80%
Arundhati et al. [6]	Quiz-based assessment	Rule-based scoring	n/a*
Patel et al. (review) [1]	Multiple (56 studies)	Hybrid / Ensemble ML	80-99%

Random Forest was selected as the primary classifier for the proposed system for the following reasons, based purely on the comparative literature evidence in Tables I-II rather than on our own experimental

results, since no algorithm has yet been trained or tested on our own collected data:

* Reported as classification/flag rate rather than accuracy.

** Reported as qualitatively high; exact figure not specified in the source.

A. Handwriting-Based Detection

- Zaibi and Bezine [5] segmented digitized strokes using the beta-elliptic model and compared Decision Tree, SVM, Random Forest, AdaBoost and Gradient Boosting; the ensemble models reached 99% accuracy on 120 schoolchildren.
- Drotár and Dobeš [10] used AdaBoost for dysgraphia detection from handwriting, reporting an accuracy of approximately 80 %, which is lower than the ensemble results above.

B. Eye-Tracking and EEG-Based Detection

- Raatikainen et al. [4] used Random Forest for feature selection, followed by SVM classification on eye-movement data from 161 students, achieving 89.7% accuracy and 84.8% recall for below percentile reading fluency.
- Kaisar's survey [3] found SVM to be the most common classifier across eye-tracking and EEG-based dyslexia studies, with accuracies typically between 80% and 97%.
- Ahire et al. [7], as reported in [1], compared SVM, Random Forest, Logistic Regression, K-NN, Decision Tree and XGBoost on EEG data for dyslexia and ADHD, with SVM achieving the highest accuracy (97% and 83%, respectively).

C. Multimodal and Assessment-Based Systems

- Joshi et al. [2] combined eyeball tracking, handwriting analysis, a listening/logical test, and a color-blindness check in one web/mobile platform, reporting 93.6% accuracy for the tracking-based module.
- Arundhati et al. [6] used age-grouped reading, spelling, comprehension, and arithmetic quizzes with a 12-case rule-based decision table to classify dyslexia, dysgraphia, dyscalculia, and dyspraxia, and additionally recommended remedial exercises.
- Modak et al. [11] and Vakadkar et al. [12], as reported in [1], applied Random Forest ensembles to e-learning behavioral data and screening

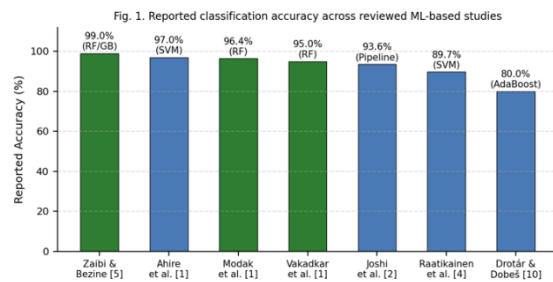
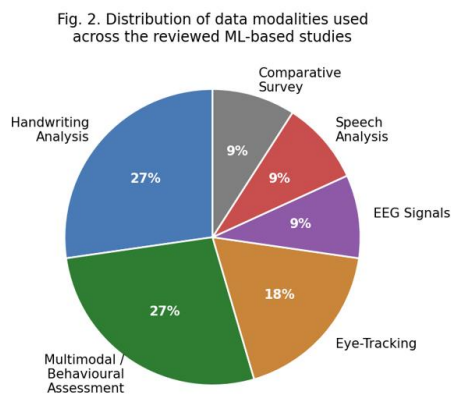
checklists, respectively, achieving 96.4% and 95% accuracy.

- Dutta and Sait [9] used Gradient Boosting for feature extraction from voice spectrograms, followed by SVM classification for speech disorder detection, reporting high accuracy relative to manual diagnosis.

D. Comparative Review

Patel et al. [1] conducted a PRISMA-guided systematic review of 56 studies spanning handwriting, EEG, speech and behavioural-checklist based ML models. Their key finding was that hybrid and ensemble ML combinations consistently outperformed single classifiers, and that preprocessing (handling missing values, dimensionality reduction) meaningfully improved accuracy.

III. COMPARATIVE ANALYSIS



As shown in Fig. 1, ensemble and random forest-based methods ([5], [11], [12]) occupy the top accuracy positions among the reviewed studies, whereas single classical classifiers such as AdaBoost trail behind [10]. Fig. 2 shows that handwriting analysis and multimodal/behavioral assessment are the two most

common modalities among the reviewed studies, together accounting for over half of the surveyed literature, followed by eye-tracking, EEG, and speech-based approaches.

IV. RESEARCH GAPS AND CHALLENGES

- Dataset size and diversity: Most studies use fewer than 300 participants drawn from a single institution, region, or language group, limiting generalizability [1].
- Hardware dependence: EEG headsets and infrared eye-trackers used in several studies are costly and impractical for routine school deployment [3], [4].
- Single-disorder focus: Most systems target one disorder in isolation, whereas SLDs frequently co-occur in the same child [1].
- Manual rule-based thresholds: Some practical systems rely on hand-coded scoring rules rather than trained, validated ML models, limiting their robustness [6].
- No integrated remediation: Most reviewed systems stop at detection without linking results to personalized learning material, with [6] being a notable exception.

V. PROPOSED SYSTEM AND CHOICE OF ALGORITHM

Fig. 3 shows the generic ML-based pipeline followed by the reviewed studies, from data collection through preprocessing, feature extraction/selection, classification, and risk output. Fig. 4 shows how this generic pipeline is adapted in our proposed system, LDDetect, which pairs a software-only handwriting canvas and a reading/spelling assessment with two independent classifiers, avoiding the specialized EEG or infrared eye-tracking hardware used in several reviewed studies [3], [4].

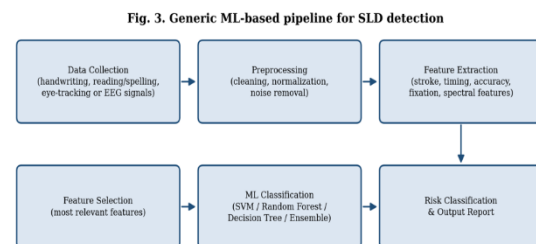


Fig. 4. Proposed architecture of the LDDetect screening system

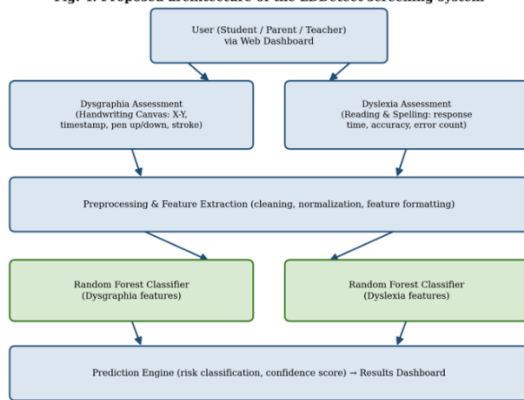


Table II. Algorithm-wise accuracy across reviewed studies

Algorithm	Studies reporting it as best	Accuracy range
Random Forest	[11], [12], and joint-best in [5]	95% - 99%
Gradient Boosting / Ensemble	[5], [9]	~99% (High)
SVM	[3], [4], [7]	80% - 97%
AdaBoost	[10]	~80%
K-NN	[8]	n/a*

* See note under Table I.

- It achieves the single highest or joint-highest reported accuracy among all reviewed studies (95%-99%) [5], [11], [12].
- As an ensemble of decision trees, it is comparatively robust to the noise typically present in handwriting and behavioral feature data [1].
- It provides feature-importance scores, giving interpretable insight into which handwriting or reading features drive a prediction, which is useful for generating recommendations [1].
- Its ensemble structure reduces the overfitting risk on comparatively small datasets typical of this domain [1].

This choice will be validated empirically once our dataset has been collected and multiple classifiers (including SVM and Decision Tree, for comparison) have been trained and tested.

VI. CONCLUSION

This review compared 12 studies that applied classical Machine Learning to the detection of Specific Learning Disabilities across handwriting, eye-

tracking, EEG, speech, and multimodal assessment modalities. Reported accuracies ranged from approximately 80% to above 99%, with random forest and ensemble-based methods consistently at the top of this range. Five recurring gaps were identified: small and narrow datasets, dependence on specialized hardware, single-disorder focus, non-validated rule-based logic, and the general absence of integrated remedial recommendations. Based on this comparative evidence, Random Forest is proposed as the primary classification algorithm for our system, LDDetect, although this choice remains to be empirically validated once our own assessment data have been collected. Future work will focus on training and comparing multiple classifiers on this data and integrating detection with personalized, actionable learning recommendations.

REFERENCES

- [1]. H. Patel, V. Sheth, and M. Shah, "An early detection of learning disabilities using machine learning: A comparative study of models, datasets, and diagnostic strategies," *Intelligent Hospital*, vol. 1, Art. no. 100022, 2025.
- [2]. A. Joshi, R. Bagate, Y. Hambir, A. Sapkal, N. P. Sable, and M. Lonare, "System for detection of specific learning disabilities based on assessment," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 9s, pp. 362–368, 2024.
- [3]. S. Kaisar, "Developmental dyslexia detection using machine learning techniques: A survey," *ICT Express*, vol. 6, pp. 181–184, 2020.
- [4]. P. Raatikainen, J. Hautala, O. Loberg, T. Kärkkäinen, P. Leppänen, and P. Nieminen, "Detection of developmental dyslexia with machine learning using eye movement data," *Array*, vol. 12, Art. no. 100087, 2021.
- [5]. T. Zaibi and H. Bezine, "Early detection of learning disabilities through handwriting analysis and machine learning," *Procedia Computer Science*, vol. 246, pp. 3702–3712, 2024.
- [6]. A. R. Arundhati, G. Sinha, D. Sinha, and P. T. Siva Sankar, "Learning disability detector and classification system with solution," *IEEE, Jain University, Bengaluru, India*.

- [7]. N. Ahire, R. N. Awale, and A. Wagh, "Development of EEG-based identification of learning disability using machine learning algorithms," *Journal of Neurological Disorders*, vol. 10, no. 10, Art. no. 521, 2022.
- [8]. R. U. Khan, J. L. A. Cheng, and O. Y. Bee, "Machine learning and dyslexia: Diagnostic and classification system (DCS) for kids with learning disabilities," *International Journal of Engineering & Technology*, vol. 7, no. 3.18, pp. 97–100, 2018.
- [9]. A. K. Dutta and A. R. Wahab Sait, "A speech disorder detection model using ensemble learning approach," *Journal of Disability Research*, vol. 3, no. 3, 2024.
- [10]. P. Drotár and M. Dobeš, "Dysgraphia detection through machine learning," *Scientific Reports*, vol. 10, Art. no. 21541, 2020.
- [11]. M. Modak, P. Gharpure, and M. Sasikumar, "Learning traits and capture mode of learning disability with classification in e-learning for detecting learning disability using machine learning," *Health Sciences*, pp. 9050–9077, 2022.
- [12]. K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of autism spectrum disorder in children using machine learning techniques," *SN Computer Science*, vol. 2, no. 5, 2021.