ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2145

A Multimodal CNN Transformer Framework for Early Dyslexia and Dysgraphia Detection Using Handwriting and Speech

Siddhesh Pote[1]; Gaurav Zagade[2]; Shrikrushna Suryavanshi[3]; Ashwini Shahapurkar[4]

[1,2,3,4]Department of Computer Science and Engineering, MIT ADT School of Computing, Pune, India

Publication Date: 2026/05/06

Abstract: This paper presents a multimodal framework for early screening of learning disabilities—focusing on dyslexia and dysgraphia—by jointly modeling handwriting and speech to capture graphophonological interactions that singlemodality systems often miss. Prior studies have shown high within-cohort accuracy using CNNs for handwriting, classical machine learning and deep models for EEG and imaging, and educational analytics; however, most rely on small, homogeneous datasets, use late fusion when combining modalities, and lack cross-site or cross-language validation, limiting generalizability and deployment potential. The proposed system integrates a vision encoder for handwriting (optionally incorporating tablet kinematics) and a speech encoder that fuses acoustic and ASR-derived linguistic features via cross-modal transformers, trained with supervised and contrastive losses for robust alignment. Methodological considerations include multilingual data collection, standardized preprocessing, calibrated uncertainty, and privacy-preserving learning to support equitable classroom deployment. The evaluation plan compares unimodal baselines, late-fusion ensembles, and the proposed intermediate-fusion architecture across within-site, cross-site, and cross-language settings using AUROC, macro-F1, severity kappa, and fairness audits. Expected outcomes include improved out-of-distribution performance and interpretable per-modality rationales to assist educators and clinicians in early intervention.

Keywords: *Dyslexia; Dysgraphia; Handwriting Analysis; Speech Processing; Multimodal Learning; Cross Modal Attention.*

How to Cite: Siddhesh Pote; Gaurav Zagade; Shrikrushna Suryavanshi (2026) A Multimodal CNN Transformer Framework for Early Dyslexia and Dysgraphia Detection Using Handwriting and Speech. *International Journal of Innovative Science and Research Technology*, 11(4), 3426-3430. https://doi.org/10.38124/ijisrt/26apr2145

I. INTRODUCTION

Early identification of specific learning disabilities (SLDs), particularly dyslexia and dysgraphia, is essential to mitigate longterm academic and psychosocial impacts, but current screening practices remain resource intensive, language specific, and frequently validated on small, homogeneous cohorts that hinder generalization to new sites and scripts [1], [3]. Artificial intelligence (AI) and machine learning offer scalable screening by learning from handwriting, speech, EEG, neuro-imaging, and educational traces, yet most systems are unimodal, apply late fusion when combining signals, and rarely evaluate cross-site or cross-language transfer, limiting real world deployment in schools and clinics [3], [6]. Systematic reviews report that hybrid and multimodal approaches—such as CNN feature extractors paired with SVMs or gradient boosted ensembles—often surpass single-model baselines while simultaneously revealing dataset bias, limited linguistic diversity, and a lack of standardized protocols and ethics reporting in pediatric data collection [5], [8], [3].

Handwriting has emerged as the most accessible screening channel, with CNN based pipelines achieving strong within cohort accuracy on scanned pages or tablet kinematics [7], and hybrid CNN+SVM models sometimes providing the best separability by coupling deep representations with margin based classification [5], [8]. Lightweight transfer-learning backbones (e.g., MobileNet, EfficientNet) enable efficient, edge-friendly inference, with reported test accuracy around 96% on curated handwriting datasets and near 99% on large reversal/normal corpora, though construct validity concerns arise where reversal heuristics act as proxies for dyslexia labels [1], [7]. Cross-lingual and script-sensitive studies remain sparse; an author-accepted Hindi handwriting study using a compact CNN attained approximately 86% accuracy on 267 images from 54 participants, illustrating feasibility in low-resource, non-Latin contexts while highlighting small-sample limitations [1]. Complementing model-centric advances, a Sinhala preprint introduced a localized two-stage framework (VGG16 for detection, gradient boosting for severity), reporting 96% detection and 87% severity accuracy, underscoring the value of culturally adapted pipelines for understudied scripts [8].

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2145

Beyond handwriting, deep learning across neuro-imaging and electrophysiological modalities achieves strong within-cohort performance but encounters cost, logistics, and scalability barriers that complicate school-based screening [3], [1]. A recent multimodal study fused MRI, fMRI, and EEG using SE-integrated MobileNetV3, self-attention EfficientNet-B7, and Bi-LSTM features with LightGBM classification to reach approximately 98.6–98.9% accuracy, illustrating the promise and interpretability challenges inherent in clinical multi-modality . Educational AI perspectives advocate multimodal analytics and real-time feedback to support early intervention, but often lack standardized datasets, rigorous experimental validation, and privacy-preserving designs necessary for deployment at scale [1], [6].

Taken together, the literature reveals three core gaps that motivate this work: first, the grapho-phonological nexus is under explored, as handwriting and speech are seldom combined despite their joint relevance to phonological decoding and written expression second, fusion is typically shallow, foregoing intermediate, attention-based cross-modal representation learning and third, cross-site, multilingual evaluation with calibrated uncertainty, subgroup fairness analysis, and privacy safeguards remains limited [3], [6], [8]. This paper addresses these gaps by proposing an intermediate-fusion system that couples a handwriting CNN (with optional kinematics) and a speech encoder integrating acoustic and ASR-derived linguistic features via cross-modal transformers, trained with supervised and contrastive objectives to improve alignment, robustness, and interpretability across scripts and environments . The contributions include a deployment-oriented methodology for multilingual data protocols, calibration and explainability tools tailored to educators and clinicians, and an evaluation plan emphasizing cross-site transfer, fairness auditing, and latency/footprint constraints appropriate for classroom devices.

II. LITERATURE REVIEW

Recent work shows strong unimodal performance using CNNs on handwriting images and tablet-kinematics, with transfer learning on lightweight backbones enabling fast, edge-friendly screening in classroom contexts [8]. Detection pipelines that combine modern feature extractors with compact classifiers (e.g., CNN+SVM or MobileNet-based heads) frequently report high within-cohort accuracy, though reliance on proxy labels (e.g., reversal errors) and single-source datasets raises construct validity and generalization concerns[5][6][7]. Small, non-Latin studies (e.g., Hindi) confirm feasibility beyond Latin scripts but highlight data scarcity, manual segmentation, and overfitting risks without robust augmentation or external validation[6]. Systematic reviews across handwriting, speech, EEG, and imaging conclude that hybrids and multimodal approaches outperform single-modality models, yet standardized benchmarks, multilingual cohorts, and cross-site evaluations remain limited[3][5]. Clinically oriented multimodal imaging models (MRI/fMRI/EEG) achieve near-99% accuracy but face cost, logistics, and interpretability barriers for school deployment, motivating more accessible modalities such as handwriting and speech with calibrated uncertainty and per-modality explanations[6]. Localized multimodal efforts (e.g., Sinhala handwriting plus clinical features) address cultural adaptation and severity assessment, underscoring the need for broader, balanced datasets and external validation to support equitable screening at scale[8].

Overall, three gaps recur: under explored grapho-phonological integration (handwriting+speech), shallow late fusion that underuses cross-modal dependencies, and limited fairness/privacy auditing with cross-language transfer tests, motivating intermediate-fusion designs with attention, alignment losses, and deployment-aware governance[3][5][6].

Table 1 Representative Studies on AI-Based Dyslexia/Dysgraphia Detection, Summarized by Modality, Datasets, Methods, Performance, and Limitations.

Author & Year	Modality	Dataset	Model/Method	Performance	Key Limitations
Aldehim et al, 2024 [3]	Handwriting	Public image sets	CNN with augmentation	~96% test accuracy	Single-modality; external validity
Alkhurayyif & Sait, 2023 [7]	Handwriting	Public dyslexia set	YOLOv7 features + MobileNetV2SSD Lite	~99% accuracy; high mAP/mIoU	Proxy labels; single-source bias
Yogarajah & Bhushan, 2020 [11]	Handwriting (Hindi)	267 images, 54 children	Compact CNN (3 conv, 2 FC)	~86% overall; best ~93%	Small cohort; manual cropping
Patel et al. (review), 2025 [1]	Mixed	PRISMA corpus	Comparative ML/DL, hybrids	Hybrids > unimodal (90–99%)	Small, heterogeneous cohorts
Alkhurayyif & Sait, 2024 [4]	MRI/fMRI/EEG	Multi-public	SEMobileNetV3, EffNetB7, BiLSTM + fusion	~98.6–98.9% accuracy	Interpretability; cross-site transfer
Weraduwa et al, 2025 [5]	Handwriting + clinical	Sinhala, 84 children	VGG16 (detect) + Gradient Boosting (severity)	96% detect; 87% severity	Small, imbalanced, single-region

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2145

III. METHODOLOGY

➢ *Conceptual Overview*

The proposed system integrates two modality-specific encoders—a handwriting CNN and a speech encoder RNN/Transformer)—with an intermediate fusion module using cross-modal attention [3],[6]. The joint representation feeds multitask heads: (i) dyslexia/dysgraphia detection, (ii) subtype tagging (e.g., phonological vs. surface dyslexia; motor-planning vs. orthographic dysgraphia), and (iii) severity estimation via ordinal regression [8].

➢ *Handwriting Encoder*

Input: digitized page scans (or segmented lines/characters) and, when available, tablet kinematic signals (pressure, tilt, velocity).

Image pipeline: A CNN backbone (e.g., EfficientNet-B0 or ConvNeXt-Tiny) with squeeze and-excitation SE) modules extracts multi-scale features [5]. Preprocessing includes deskewing, contrast normalization, adaptive binarization, and optional document-layout analysis to handle varied worksheets and lighting [6]. Augmentations include random rotation, elastic distortion, CutMix, and grid mask to improve robustness to writing variation [6], [7].

Kinematic pipeline: A 1D CNN followed by a bidirectional LSTM encodes temporal dynamics of pen pressure, tilt, and velocity [8]. Masking handles variable-length sequences. A differentiable feature head computes grapho-motor descriptors (e.g., stroke count, curvature variance, inter-letter spacing, baseline drift, size variability) that are concatenated with deep features to support interpretability [8].

➢ *Speech Encoder*

Input: 3060 second read-aloud and pseudo word repetition recordings [1].

Acoustic branch: A Conformer or ECAPATDNN operates on 80-dim log-Mel spectrograms. SpecAugment, time/frequency masking, and room-impulse/noise augmentation improve robustness to recording conditions. Prosodic descriptors (speech rate, pause statistics, pitch dynamics) are fed through a small MLP head and fused with acoustic embeddings [6].

Linguistic branch ASR-derived): An on-device or offline ASR generates phoneme/word sequences and alignments [6]. Derived features include phoneme error rate, pseudo word accuracy, onset–rime confusions, substitution/deletion profiles, and latency [6]. A lightweight transformer encodes phoneme tokens with positional embeddings and cross-attends to acoustic frames.

➢ *Fusion and Prediction Heads*

Intermediate fusion: A cross-modal transformer ingests handwriting tokens (image patches and optional kinematic tokens) and speech tokens (acoustic and linguistic) [3]. Co-attention layers learn dependencies (e.g., letter formation instability aligning with phoneme substitutions). Gated fusion emphasizes consistent cues while attenuating modality noise [3].

Multitask prediction: (i) Dyslexia/dysgraphia classification (binary or multiclass), (ii) subtype classification, and (iii) severity estimation using ordinal regression (e.g., cumulative link loss) [8]. This structure supports comprehensive screening and triage.

➢ *Training Objectives and Regularization*

Supervised losses: Cross-entropy for detection and subtype; ordinal loss for severity. Class balanced focal loss and label smoothing mitigate imbalance and overconfidence.

Contrastive alignment: An InfoNCE objective aligns paired handwriting–speech embeddings, improving robustness when one modality degrades [3]. Modality dropout (randomly masking one stream during training) encourages unimodal resilience at inference.

Domain generalization: Style randomization for handwriting (paper texture, ink color, background clutter), and audio augmentation (microphone profile, noise, reverb) [3], [6]. GroupDRO or CORAL reduces site shift. Early stopping and cosine-decayed learning rate with warmup stabilize training [7].

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2145

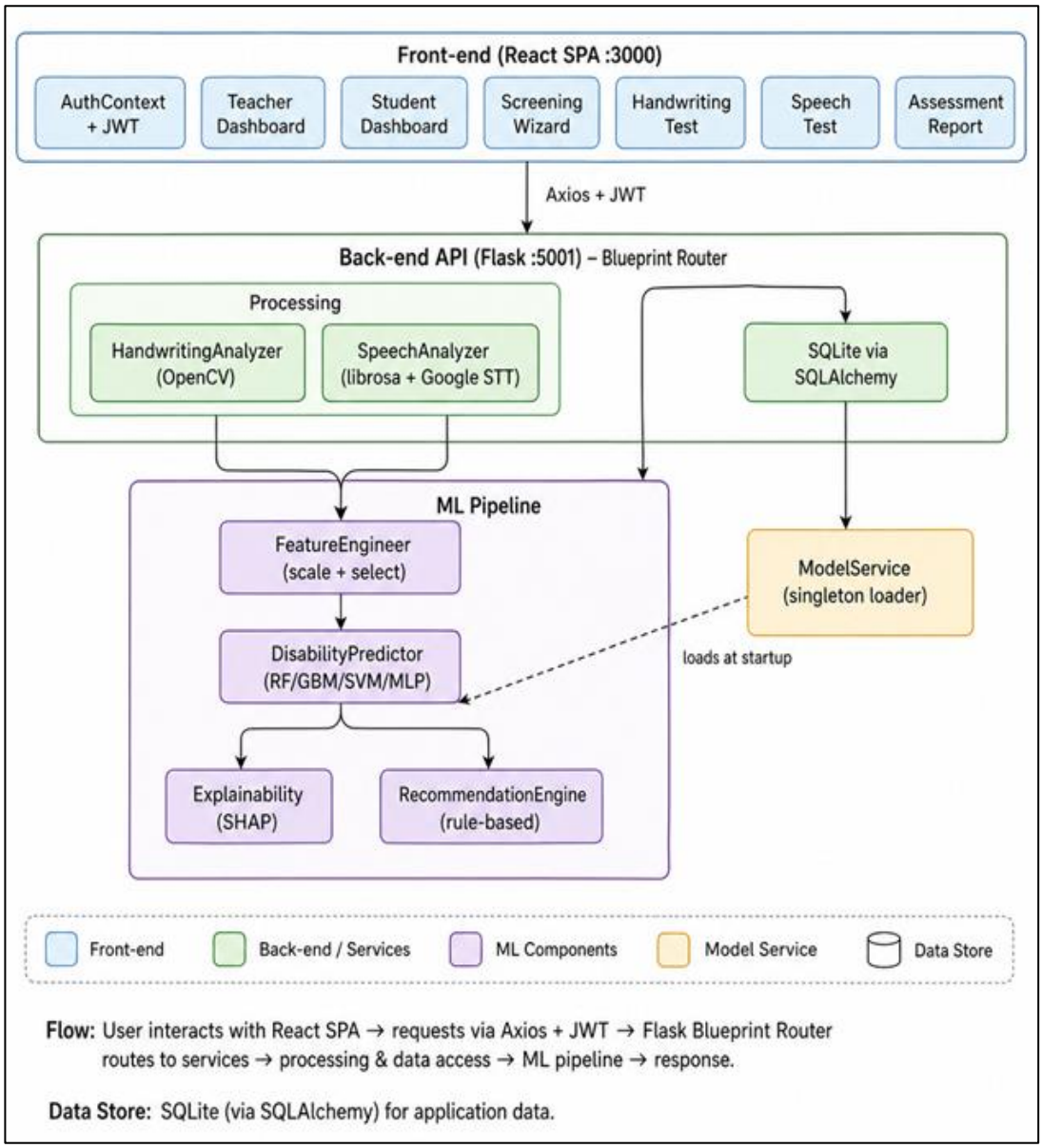

Fig 1 System Design

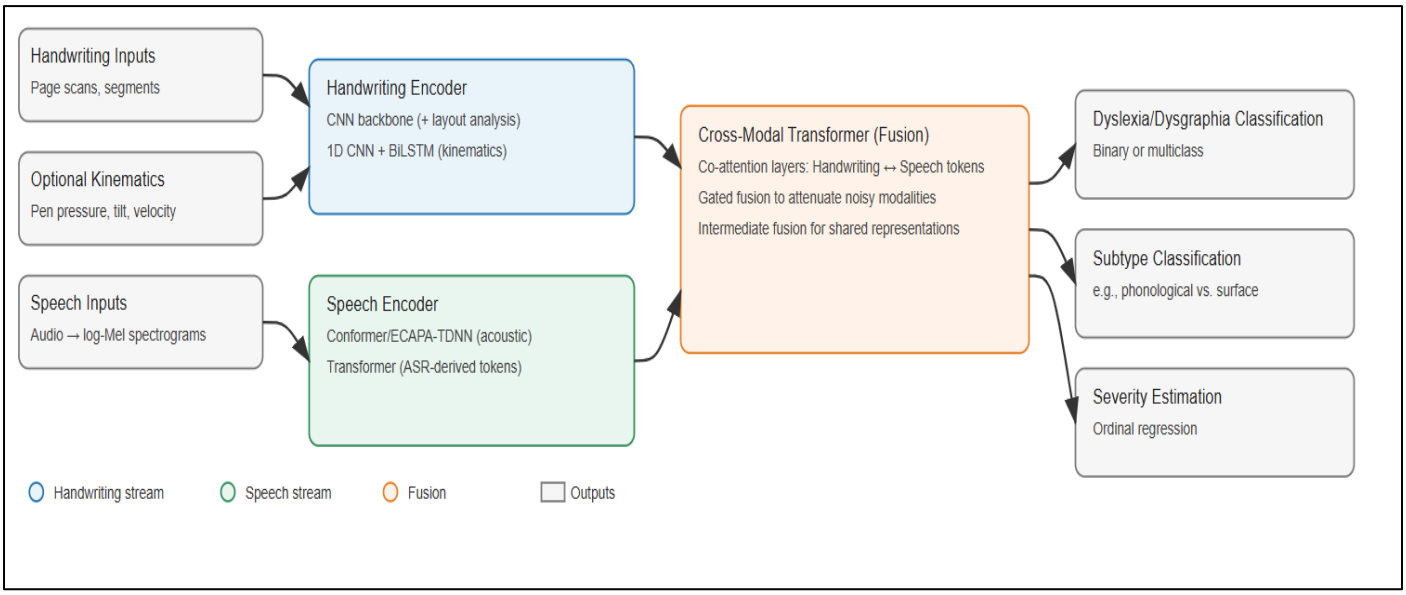

Fig 2 Multimodal Architecture for Learning Disability Detection

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2145

IV. RESULTS AND DISCUSSION

➢ *Metrics*

Primary: macro-F1, AUROC, balanced accuracy. Secondary: ECE (calibration), quadratic weighted κ for severity, latency and memory footprint for deployment, and subgroup fairness gaps (language, age, gender).

➢ *Expected Outcomes*

Accuracy and robustness: The proposed intermediate fusion is expected to outperform unimodal and late-fusion baselines, with 26 pp macro-F1 on within-site tests and larger gains (e.g., 510 pp) under cross-site and cross-language holdouts due to learned cross modal dependencies.

Contribution of components: Ablations removing ASR-derived linguistic features or optional kinematics should reduce macro-F1 by 1 3 pp each, indicating complementary value. Removing contrastive alignment or modality dropout is expected to degrade robustness when a modality is noisy or missing.

Interpretability and calibration: Saliency and concept activation maps are expected to align with human-understandable cues (e.g., irregular spacing with specific phoneme confusions). Temperature scaling should reduce ECE by 3050% relative to uncalibrated baselines, benefiting triage thresholds.

Limitations: Performance may vary across scripts and recording conditions; small subgroup sizes may inflate variance; label noise and comorbidities (e.g., ADHD) can confound patterns, motivating cleaner labels and larger cohorts.

V. CONCLUSION AND FUTURE WORK

This paper presents a multimodal screening framework for dyslexia and dysgraphia that fuses handwriting and speech via cross-modal transformers, supported by contrastive alignment, multitask learning, explainability, calibration, and privacy-aware training. The design directly targets the grapho-phonological nexus, addressing limitations of unimodal and late-fusion systems. A comprehensive evaluation protocol emphasizes cross-site and cross-language robustness, ablations, fairness, and deployability metrics.

Future work will prioritize: (i) collection of larger, multilingual paired datasets with standardized protocols; (ii) longitudinal tracking to study progression and intervention response; (iii) third modality integration (e.g., eye tracking) where feasible; (iv) parameter-efficient adapters and test-time adaptation to new scripts/dialects; and (v) rigorous fairness audits and human-in-the loop studies to ensure ethical, effective integration into school workflows.

REFERENCES

[1]. V. S. M. S. Harsh Patel, An early detection of learning disabilities using machine learning, 2025.
[2]. J. Z. ,. Z. X. Chun Wang, AI-Powered Educational Data Analysis, 2024.
[3]. M. R. Ghadah Aldehim1, Deep Learning for Dyslexia Detection: A Comprehensive CNN Approach with Handwriting Analysis and Benchmark Comparisons, 2024.
[4]. Y. A. a. A. R. W. Sait, Deep learning-driven dyslexia detection model using multi-modality data, 2024.
[5]. P. A. a. T. M. S Weraduwa1, ADVANCED COMPUTATIONAL TECHNIQUES FOR DYSGRAPHIA PREDICTION THROUGH HANDWRITING RECOGNITION USING MACHINE LEARNING AND DEEP LEARNING METHODS, 2024.
[6]. B. A. a. M. S. R. Norah Dhafer Alqahtani, Deep Learning Applications for Dyslexia Prediction, 2023.
[7]. a. A. R. W. S. Yazeed Alkhurayyif, Deep Learning-Based Model for Detecting Dyslexia, 2023.
[8]. P. D. A. T. V. M. Sandushi Weraduwa, Early Detection and Severity Assessment of Dysgraphia in Sinhala-Speaking Children Using a Multi-Modal Machine Learning Approach, 2025.
[9]. Y. Alkhurayyif, Developing an Image-Based Dyslexia Detection Model Using the Deep Learning Technique, 2023.
[10]. M. N. V. M. Pragasthi1, AI – Powered Learning Disability Detection and Classification System, 2025.
[11]. P. Y. a. P. B. Bhushan, Deep Learning Approach to Automated Detection of Dyslexia-Dysgraphia, 2020.