EduMirror-XAI: An Explainable and Leakage-Resistant Deep Learning Framework for Detecting Handwritten Letter Reversals in Special Education

C Pabitha[1], Sanaj M S[2], Kavitha Nair R[3], Asesh Kumar Tripathy[4], Reji R[5], M Arunachalam[6], Abhishek K Singh[7], Gourav Kalra[8], Divya Saleela[9], Jennie Prajith[10]

[1]Associate Professor, SRM Valliammai Engineering College, Tamil Nadu, India

[2]Professor, Vidya Academy of Science and Technology, Kerala, India

[3]Associate Professor, Acharya Institute of Technology, Karnataka, India

[4]Associate Professor, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India

[5]Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India

[6]Professor, Sri Krishna College of Engineering and Technology, Kuniamuthur, Tamil Nadu, India

[7]Associate Professor, School of Liberal Studies and Humanities, University of Petroleum and Energy Studies, Uttarakhand, India

[8]Associate Professor, School of Engineering and Technology, CGC University, Punjab, India

[9]Assistant Professor, College of Engineering Aranmula, Kerala, India

[10]Assistant Professor, Pillai College of Arts Science and Commerce (Autonomous), Mumbai University, Maharashtra, India

pabithac.cse@srmvalliammai.ac.in[1], sanaj897@vidyaacademy.ac.in[2], kavitharnair88@gmail.com[3], asesh.tripathy@gmail.com[4], rejir@tkmit.ac.in[5], dr.arunachalam13280@gmail.com[6], asingh8319@gmail.com[7], gkalra89@gmail.com[8], divyasal009@cearanmula.ac.in[9], jennieprajith10@gmail.com[10]

Corresponding Author: divyasal009@cearanmula.ac.in

HOW TO CITE:

C Pabitha, Sanaj M S, Kavitha Nair R, Asesh Kumar Tripathy, Reji R, M Arunachalam, Abhishek K Singh, Gourav Kalra, Divya Saleela, Jennie Prajith (2026). EduMirror-XAI: An Explainable and Leakage-Resistant Deep Learning Framework for Detecting Handwritten Letter Reversals in Special Education. International Journal of Special Education, 41(10s), 615-621.

ABSTRACT:

Handwritten letter reversals may occur during literacy development and may persist among learners with reading and writing difficulties; however, they cannot independently establish dyslexia. This study proposes EduMirror-XAI, a lightweight and duplicate-aware deep-learning framework for classifying correctly oriented and reversed handwritten characters. A balanced sample of 60,000 images was derived from the Dyslexia Handwriting Dataset and partitioned using perceptual-signature grouping to reduce duplicate-related leakage. EduMirror-XAI combined a pretrained MobileNetV3Small backbone, spatial attention, controlled augmentation, and two-stage transfer learning. The validation-selected threshold was fixed before held-out testing. On 9,014 test images, the model achieved 92.69% accuracy, 91.53% sensitivity, 93.85% specificity, a 92.61% F1-score, an MCC of 0.8540, a ROC-AUC of 0.9794, and a PR-AUC of 0.9797. Performance remained stable under blur, rotation, and

COPYRIGHT STATEMENT:
Copyright: © 2026 Authors.
Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

compression, while reduced contrast produced the largest decline. EduMirror-XAI demonstrates the feasibility of image-level reversal detection for preliminary teacher review but is not intended as a dyslexia diagnostic system.

Keywords: artificial intelligence; special education; letter reversal; handwriting analysis; early literacy

1. Introduction

Handwritten character reversals, including mirror-oriented forms and confusions such as *b*–*d* and *p*–*q*, may occur during early literacy development and may persist among learners experiencing reading and spelling difficulties. Their presence alone, however, is insufficient to establish dyslexia, which requires evidence from reading, decoding, spelling, phonological processing, developmental history, classroom observation, and professional assessment [1].

Automated analysis of an isolated handwritten character should therefore be interpreted as detection of a visual writing pattern rather than identification of a learning disability.

Previous studies have demonstrated the feasibility of deep-learning-based handwriting classification. Transfer-learning models have been used to distinguish normal, reversed, and corrected handwriting images [2], while hybrid convolutional and machine-learning approaches have reported strong classification performance [3]. Explainable models have also incorporated visualisation methods to identify image regions influencing predictions [4]. More recent work has emphasised that AI-assisted handwriting analysis should complement practitioner-led screening rather than function as an independent diagnostic mechanism [5], [6].

Despite these advances, several methodological limitations remain. Corrected handwriting represents a different visual process from both correct orientation and reversal, making three-class formulations difficult to interpret. Public handwriting datasets may also contain duplicate or near-duplicate images that can be distributed across training and test sets, potentially inflating estimated generalisation [7]. In addition, previous studies often emphasise overall accuracy without fully examining class-specific performance, probability calibration, uncertainty, or robustness under realistic image degradation.

To address these limitations, this study proposes EduMirror-XAI, a lightweight framework for binary classification of correctly oriented and reversed handwritten characters. The framework combines a pretrained MobileNetV3Small backbone [8], learned spatial attention, controlled augmentation, and two-stage fine-tuning. Perceptual-signature grouping is applied before partitioning to reduce duplicate-related leakage, while the operating threshold is selected exclusively using validation data. Performance is evaluated using class-specific metrics, ROC-AUC, PR-AUC, calibration, bootstrap confidence intervals, and robustness under blur, reduced contrast, rotation, and compression.

The principal contributions of this study are fourfold. First, it formulates the task as binary character-orientation classification rather than dyslexia diagnosis. Second, it introduces duplicate-aware partitioning based on perceptual image signatures. Third, it combines a lightweight MobileNetV3Small backbone with spatial attention and staged fine-tuning. Fourth, it applies a validation-controlled evaluation protocol incorporating threshold optimisation, calibration, confidence intervals, and perturbation-based robustness analysis.

The objective of the study is to determine whether EduMirror-XAI can reliably distinguish correctly oriented and reversed handwritten-character images under a duplicate-aware held-out evaluation protocol. The system is intended to support preliminary teacher review and sample

prioritisation, not to replace literacy assessment, specialist judgement, or formal dyslexia diagnosis. The remainder of the paper is organised as follows. Section 2 describes the dataset, preprocessing, duplicate-aware partitioning, model architecture, training procedure, and evaluation methods. Section 3 presents the classification, calibration, and robustness results. Section 4 discusses the findings, limitations, and educational implications. Section 5 concludes the study and outlines directions for future research.

2. Methodology

This study adopted a quantitative experimental methodology to develop and evaluate EduMirror-XAI for binary classification of correctly oriented and reversed handwritten characters. The workflow comprised dataset screening, balanced sample construction, duplicate-aware partitioning, image preprocessing, model development, two-stage transfer learning, validation-based threshold selection, held-out testing, and robustness analysis. The training set was used for parameter optimisation, the validation set for model monitoring and threshold selection, and the test set exclusively for final evaluation.

2.1 Dataset and Partitioning

The publicly available Dyslexia Handwriting Dataset [9] was used as the data source. Corrected handwriting images were excluded because overwriting, retracing, erasure, and post-response modification represent a different visual process from both correct orientation and reversal. Image-integrity checking identified 123,554 valid samples, comprising 58,891 correctly oriented and 64,663 reversed images.

A balanced modelling dataset of 60,000 images was constructed using 30,000 samples from each class. Correct orientation was encoded as Class 0 and reversal as Class 1. To reduce duplicate-related information leakage, perceptual image signatures were generated before partitioning. Images sharing the same signature were assigned entirely to one subset. This procedure identified 3,500 images belonging to repeated-signature groups and produced 41,985 training images, 9,001 validation images, and 9,014 held-out test images. No signature group occurred in more than one subset. Because writer identifiers were unavailable, the evaluation was duplicate-aware but not writer-independent.

2.2 Image Processing and EduMirror-XAI Architecture

All images were converted to three-channel format and resized to 160 × 160 pixels using aspect-ratio-preserving padding. Training-only augmentation included limited rotation, translation, zoom, and contrast variation. Horizontal flipping was excluded because it could alter character orientation and invalidate the class label.

EduMirror-XAI used an ImageNet-pretrained MobileNetV3Small backbone to generate 5 × 5 × 576 feature maps. A 1 × 1 convolution produced a spatial-attention map, which was multiplied element-wise with the backbone features. The resulting representation was processed using global average pooling, batch normalisation, a dense layer with 192 units, dropout of 0.35, a dense layer with 64 units, dropout of 0.20, and a sigmoid output for reversal probability. The model contained 1,065,202 parameters. The overall EduMirror-XAI architecture, including the MobileNetV3Small backbone, spatial-attention mechanism, classification head, and two-stage training strategy, is illustrated in Figure 1.

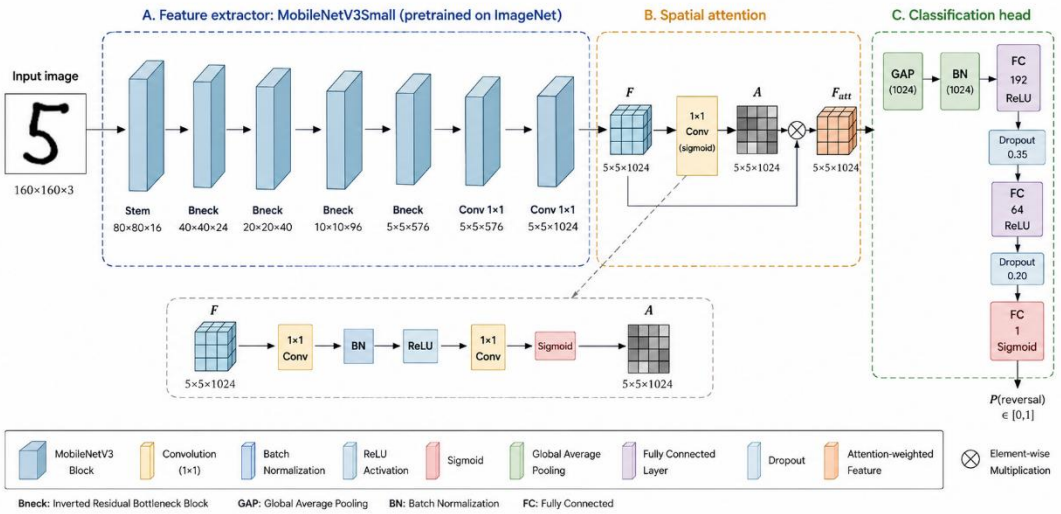

Figure 1. EduMirror-XAI architecture comprising MobileNetV3Small feature extraction, spatial attention, global average pooling, batch normalisation, dense layers, dropout regularisation, and sigmoid classification.

2.3 Model Training

Training was performed in two stages. During the first stage, the MobileNetV3Small backbone was frozen while the attention and classification layers were trained for eight epochs. During the second stage, the upper 35 backbone layers were unfrozen and fine-tuned for 12 epochs using a learning rate of 2×10^{-5}. The model was trained using the AdamW optimiser and binary cross-entropy loss with label smoothing. Validation precision–recall AUC was used for checkpoint selection. The framework was implemented in TensorFlow 2.20.0 and executed using a CUDA-enabled GPU with mixed_float16 mixed-precision computation.

2.4 Threshold Selection and Evaluation

Candidate operating thresholds were evaluated exclusively on the validation set using Matthews correlation coefficient. The optimal threshold was 0.49 and was fixed before test-set evaluation. Performance was assessed using accuracy, balanced accuracy, sensitivity, specificity, precision, negative predictive value, F1-score, Matthews correlation coefficient, ROC-AUC, PR-AUC, and Brier score. Stratified bootstrap resampling was used to calculate 95% confidence intervals.

Robustness was evaluated on 3,000 held-out test images under Gaussian blur, reduced contrast, rotation, and JPEG compression. ROC-AUC, PR-AUC, and balanced accuracy were calculated for each condition using the fixed validation-selected threshold. Correct predictions, false positives, and false negatives were also visually examined to identify recurring error patterns. EduMirror-XAI was evaluated strictly as an image-level educational support system and not as a dyslexia diagnostic tool.

3. Results

This section presents the training behaviour, held-out classification performance, and robustness of EduMirror-XAI. Test results were calculated using the validation-selected threshold of 0.49.

3.1 Training and Validation Performance

As shown in Figure 2, validation PR-AUC increased from 0.7633 to 0.8728 during the eight-epoch frozen-backbone stage. Fine-tuning the upper MobileNetV3Small layers further improved validation performance, with accuracy, ROC-AUC, and PR-AUC reaching 0.9316, 0.9806, and 0.9806, respectively. Separately, validation-based threshold optimisation selected a threshold of 0.49, yielding an MCC of 0.8638.

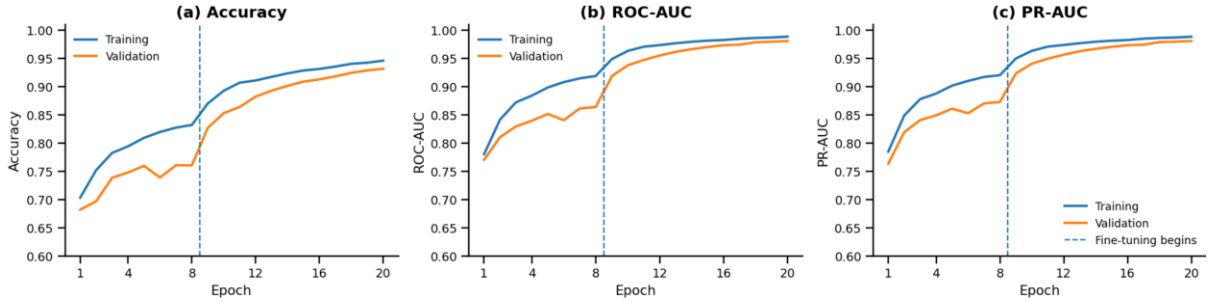

Figure 2. Training and validation performance across the frozen-backbone and fine-tuning stages.

3.2 Held-Out Test Performance

As shown in Figure 3, EduMirror-XAI achieved a ROC-AUC of 0.9794 and a PR-AUC of 0.9797 on the held-out test set, indicating strong discrimination between correctly oriented and reversed handwritten characters. The confusion matrix in Figure 4 further shows that the model maintained a balanced classification profile at the validation-selected threshold of 0.49.

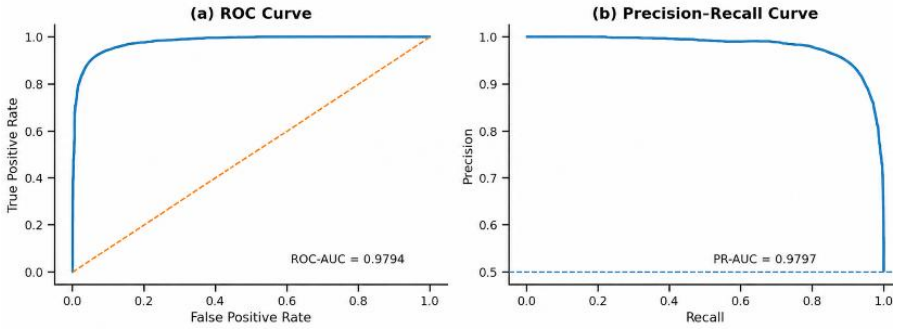

Figure 3. Receiver operating characteristic and precision–recall curves for the held-out test set.

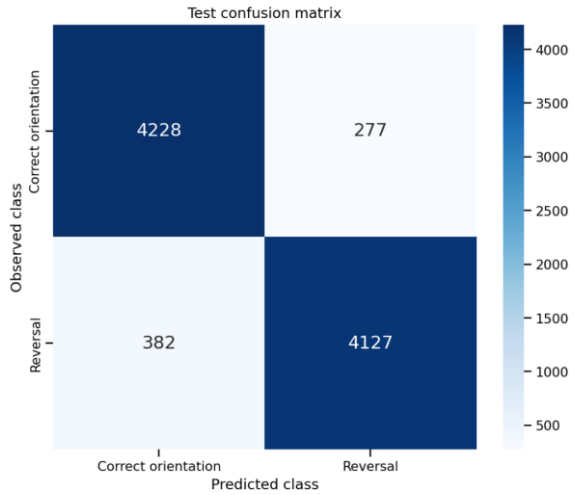

Figure 4. Confusion matrix

3.3 Robustness Analysis

As shown in Figure 5, EduMirror-XAI retained strong discrimination under blur, rotation, and JPEG compression, with ROC-AUC and PR-AUC remaining high across these perturbations. Reduced contrast produced the largest decline in balanced accuracy, although threshold-independent discrimination remained above 0.96. Figure 6 presents two representative validation inputs together with the corresponding EduMirror-XAI outputs, illustrating correct classification of a normally oriented character and a reversed character.

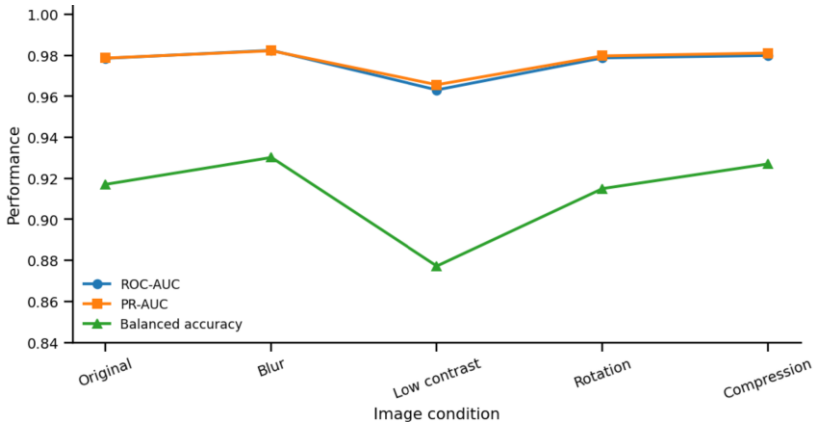

Figure 5. ROC-AUC, PR-AUC, and balanced accuracy under original, blurred, low-contrast, rotated, and compressed image conditions.

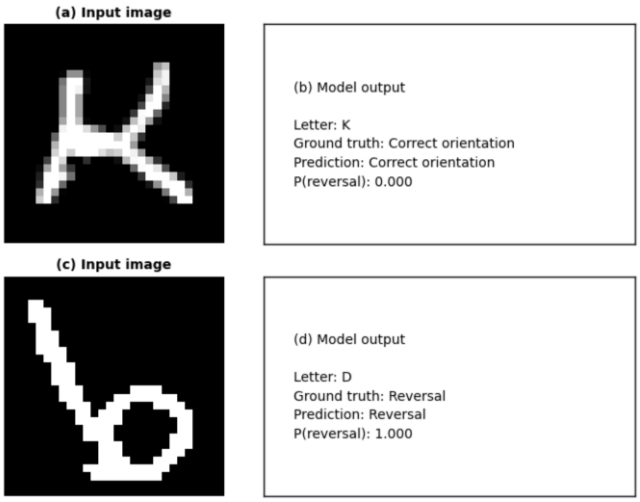

Figure 6. Representative validation inputs and corresponding EduMirror-XAI predictions.

4. Discussion

EduMirror-XAI achieved strong and balanced performance on the duplicate-aware held-out test set, with 92.69% accuracy, 91.53% sensitivity, 93.85% specificity, an F1-score of 92.61%, and an MCC of 0.8540. The ROC-AUC of 0.9794 and PR-AUC of 0.9797 further indicate that the model effectively separated correctly oriented and reversed handwritten characters. The improvement observed after fine-tuning confirms that task-specific adaptation of the pretrained MobileNetV3Small features was important for learning orientation-sensitive stroke patterns.

A principal strength of the study was the use of perceptual-signature grouping before partitioning. This reduced the risk that duplicate or near-duplicate images appeared across training and test sets, thereby improving the credibility of the reported performance. The validation-only selection of the 0.49 threshold also prevented test-set information from influencing the final decision rule. However, the experiment was not writer-independent because verified writer identifiers were unavailable.

Robustness remained high under blur, rotation, and JPEG compression, suggesting that the model was not overly sensitive to moderate image degradation. Reduced contrast caused the largest decline in balanced accuracy, indicating that weak separation between handwriting strokes and the background remains an important limitation. Future versions could address this issue through contrast normalisation and stronger low-contrast augmentation.

The findings support the use of EduMirror-XAI as an image-level educational support tool for prioritising handwriting samples for review. They

do not support its use as an automated dyslexia diagnostic system, because letter reversal alone does not capture reading, spelling, phonological, developmental, or instructional factors. The main limitations are the use of a single public dataset, absence of writer-level information, lack of external validation, and simulated rather than authentic classroom image degradation.

5. Conclusion

This study presented EduMirror-XAI, a lightweight and duplicate-aware deep-learning framework for classifying correctly oriented and reversed handwritten characters. The model achieved 92.69% accuracy, 91.53% sensitivity, 93.85% specificity, an F1-score of 92.61%, an MCC of 0.8540, a ROC-AUC of 0.9794, and a PR-AUC of 0.9797 on the held-out test set. These results demonstrate that reliable image-level orientation classification can be achieved when duplicate-related leakage is controlled and the decision threshold is selected using validation data only. EduMirror-XAI should be interpreted as a teacher-support tool for prioritising handwriting samples for review, not as a dyslexia diagnostic system.

Future work should include writer-independent and external validation, repeated training across multiple random seeds, ablation analysis of the attention and augmentation components, and character-specific performance evaluation. Prospective testing using authentic classroom handwriting and teacher feedback is also required to assess usability, educational value, and the practical consequences of false-positive and false-negative predictions.

References

R. L. Peterson and B. F. Pennington, 'Developmental dyslexia', The Lancet, vol. 379, no. 9830, pp. 1997–2007, May 2012, doi: 10.1016/S0140-6736(12)60198-6.

M. S. A. B. Rosli, I. S. Isa, S. A. Ramlan, S. N. Sulaiman, and M. I. F. Maruzuki, 'Development of CNN Transfer Learning for Dyslexia Handwriting Recognition', in 2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia: IEEE, Aug. 2021, pp. 194–199. doi: 10.1109/ICCSCE52189.2021.9530971.

S. Sreekumar and L. A, 'Comparative Study of CNN Models on the Classification of Dyslexic Handwriting', in 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India: IEEE, Dec. 2022, pp. 1–4. doi: 10.1109/IBSSC56953.2022.10037428.

M. Robaa, M. Balat, R. Awaad, E. Omar, and S. A. Aly, 'Explainable AI in Handwriting Detection for Dyslexia Using Transfer Learning', 2024, arXiv. doi: 10.48550/ARXIV.2410.19821.

S. Rangasrinivasan, M. S. Sumi Suresh, A. Olszewski, S. Setlur, B. Jayaraman, and V. Govindaraju, 'AI-Enhanced Child Handwriting Analysis: A Framework for the Early Screening of Dyslexia and Dysgraphia', SN Comput. Sci., vol. 6, no. 5, p. 399, Apr. 2025, doi: 10.1007/s42979-025-03927-0.

H. W. Liu, S. Wang, and S. X. Tong, 'DysDiTect: Dyslexia Identification Using CNN-Positional-LSTM-Attention Modeling with Chinese Dictation Task', Brain Sci., vol. 14, no. 5, p. 444, Apr. 2024, doi: 10.3390/brainsci14050444.

I. E. Tampu, A. Eklund, and N. Haj-Hosseini, 'Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images', Sci. Data, vol. 9, no. 1, p. 580, Sep. 2022, doi: 10.1038/s41597-022-01618-6.

A. Howard et al., 'Searching for MobileNetV3', in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

I. S. Isa, 'Dyslexia Handwriting Dataset', Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/drizasazanitaisa/dyslexia-handwriting-dataset. Accessed: Jun. 15, 2026.