Article

Not peer-reviewed version

Explainable Deep Learning Approaches for Dyslexia Detection in English and Arabic Handwriting Using Convolutional Neural Networks and Transfer Learning

Marwa Abu Najm * and Hamid Mukhtar

Posted Date: 28 February 2026

doi: 10.20944/preprints202602.1939.v1

Keywords: dyslexia screening; convolutional neural networks transfer learning; explainable AI

 doi:10.20944/preprints202602.1939.v1

Article

Explainable Deep Learning Approaches for Dyslexia Detection in English and Arabic Handwriting Using Convolutional Neural Networks and Transfer Learning

Marwa Abu Najm * and Hamid Mukhtar

School of Computer Science, College of Engineering and Physical Sciences, University of Birmingham Dubai

* Correspondence: mma346@alumni.bham.ac.uk

Abstract

Dyslexia impacts 5–15% of school-aged children globally, but automated screening mechanisms to detect it are rare, and such tools are relatively scarce in non-Latin scripts. The work introduces a bilingual deep learning model for dyslexia preliminary diagnosis through digitalized handwriting samples in both English and Arabic. Two computational methods were employed and compared systematically: the page-oriented classification strategy and the character-oriented classification method. For Arabic, an EnhancedCNN architecture is proposed to classify whole-page scans end-to-end by coping with cursive script and contextual letter forms. Both a baseline SimpleCNN model and a MobileNetV3-Small transfer learning model were trained on segmented letter crops from 123,554 labeled English samples. Preprocessing steps included the removal of instructor annotations, the Otsu adaptive thresholding method binarization and morphological processing noise removal and stroke refinement. Grad-CAM visualizations were included for model transparency and education decision aids, showing discriminative regions in page-level as well as character-level predictions. Experimental results proved that the proposed Arabic page-level model obtained 77% test accuracy, which constitutes preliminary proof of concept for AI-driven dyslexia screening in Arabic. English character-level approach using MobileNetV3 achieved 99% accuracy on the single letter detection task. This work also contributes to one of the earliest AI-assisted reading screening systems which is specifically designed for detecting dyslexia in Arabic script and brings systematic evidence on comparing hybrid page- and letter-level strategies for bilingual handwriting analysis.

Keywords: dyslexia screening; convolutional neural networks transfer learning; explainable AI

Introduction

Dyslexia is a kind of specific learning disability, which presents with difficulty in reading related to decoding (the translation of print into speech), spelling and language comprehension that impairs a child ability despite having the opportunity to be educated and has at least average intelligence [1]. These problems originate from disturbances in phonological processing and are frequently impaired on reading, writing, and language tasks. The prevalence has been reported between 5–15%, depending on assessment procedures, diagnostic criteria, and language orthography in school-age children worldwide [1].

The early diagnosis and intervention of dyslexia is important for the children to achieve normal outcome and support a normal life. Interventions are most effective when they are provided before the age of eight, with children identified after this age achieving significantly less academic growth and fewer gains in reading fluency, self-concept [2] The conventional diagnosis of dyslexia generally depends upon specialized expertise such as that offered by trained educational psychologists or specialist practitioners and involves a battery of standardized tests. However, even the standardized

 doi:10.20944/preprints202602.1939.v1

tests show consistent differences in test results if the motivation and attention are counted for during the test [29]. In addition, these processes tend to be expensive, time-consuming and typically take weeks, making the timely evaluation of affected subjects difficult (especially in poorly resource schools, rural habitats and underdeveloped countries). This difference is a barrier to the early recognition and, in turn, intervention.

A viable alternative is to examine handwriting with posing multiple-dimension perspective on a cognitive and motor writing model. Writing as a process involves motor control, accuracy (spelling), and fluency which can in turn all indicate challenges related to dyslexia [3]. A number of dyslexic children have characteristic handwriting features such as reversal of letters, variable shapes of letters, uneven short spaces between words, and excessive pauses [3]. These characteristics can be observed through basic pen-and-paper tasks; and with digital technology (e.g., smartphone cameras or scanners), these writings samples can be quickly digitized for analysis. Capitalizing on these developments, state-of-the-art-machine-vision and explainable AI have advanced to the point where automated tools can be created for screening handwriting samples taken in a classroom environment. These systems simultaneously produce a high-quality recognition of handwriting images and visual explanations such as attention maps and activation heatmaps. These advancements provide accessible solutions that can be scaled by educators to better enable early and more equitable identification and intervention of children with dyslexia [4,5].

Handwriting analysis for the detection of dyslexia is most often studied in English samples, yet there is a growing need to consider the specificities and challenges that other scripts, such as Arabic, impose on this method. Unlike English, Arabic script is written right-to-left and it has connected cursive letters that have different forms based on their position in a word [6]. These features complicate handwriting analysis, and it is difficult to extend commonly used English-based features to the case of Arabic. Currently, few automated tools that are able to assess Arabic handwriting, and most artificial intelligence systems do not show the evidence used for making a prediction, which mitigates educators from trusting their recommendations with sufficient actionable evidence upon reflecting on assistance of a referral [7]. To meet these challenges, this study intended to develop a bilingual handwriting assessment system using explainable AI technology for educational screening. In the U.A.E., school students are learning in English and Arabic and taking tests in writing in both. The main aim of the study was to assess dyslexia risk and generate understandable visual explanations that can help teachers, parents and specialists gain a deeper insight in this field and trust the obtained diagnosis. It's also worth noting that this research developed a screening tool, not a diagnostic device.

The primary research question motivating this research was: How well can a Convolutional Neural Network (CNN)-based AI system, enriched with Explainability aspects (as discussed above), identify dyslexia tendencies in English and Arabic Handwriting generated by children?

Five major contributions are made in this project:

- We developed individual page-level classifiers corresponding to the handwriting in English and Arabic.
- We developed and compared a letter-level classification pipeline using MobileNetV3-Small for transfer learning to the baseline SimpleCNN.
- Grad-CAM visualization methods were incorporated to enhance interpretability for educational users, including visual explanations of model predictions.
- An empirical analysis was performed, comparing the effect of our automatic segmentation to a manually optimized segmentation.
- One of the first explainable AI-based screening systems for Arabic dyslexia diagnosis was developed, closing a substantial void in multilingual educational contexts technologies.

We believe that, these contributions will advance the state-of-the-art of detection techniques for dyslexic handwriting by transferring validated writing pattern analysis methodologies to Arabic, and

 doi:10.20944/preprints202602.1939.v1

by prioritizing explainability as means to facilitate practical adoption in real-life educational contexts.

Related Work

This section surveys prior studies of automated dyslexia screening from handwriting, organized by methodological categories: page-level (document) classification, character/letter-level classification (including transfer learning), segmentation and object-detection methods, sequence/temporal modelling, and explainability for educational deployment. Each subsection summarizes representative works, highlights their strengths and weaknesses, and identifies gaps that motivated the bilingual page-and-letter comparative framework developed here.

Page-Level Classification and Document-Level Approaches

Page-level approaches treat the full scanned page as the network input and predict page- or writer-level labels directly from a global image representation. The main benefit of this approach is to be robust to segmentation failures that are quite common in naturalistic scans of the classroom or in scripts with cursive joining. Early research on handwriting analysis at the document and page levels focused on holistic features and layout descriptors for the identification of the writer and document classification, which serve as the basis for the application of page-level learning to dyslexia screening [21]. Asselborn et al. demonstrated that aggregated page statistics and global image features contain diagnostically relevant cues e.g., irregular spacing, line misalignment and overall stroke density which can serve as useful screening signals for dysgraphia/dyslexia [22].

More recently, research has shifted from engineered page statistics to end-to-end convolutional and hybrid architectures that extract hierarchical representations from full pages. Hybrid CNN–LSTM and CNN–attention networks capture both local stroke textures and longer-range layout patterns; they are typically more robust to local noise yet may be sensitive to global layout variability [23]. Page-level approaches have been found to be of particular use for scripts where segmentation itself is inherently ambiguous - such as Arabic, for instance, due to the use of contextual letter forms and ligatures combined with the fact that an extraction using isolated characters alone is often unreliable [24,25]. However, page models are also limited in some significant ways. They often need larger characterized page corpora to train, since patterns on pages are more diverse than individual character shapes; page predictions can also be less readily visible at the character scale, so it is more difficult to generate actionable, character-by-character evidence that can be examined by the teacher or by a clinician. In brief, page-level models trade reduced segmentation brittleness for loss of fine-grained interpretability and increased dependence on page-level training data.

Character/Letter-Level Classification and Transfer Learning

Character-level pipelines divide pages into candidate crops and make decisions on each crop separately, page or writer level decisions are then calculated by summing the per-crop probabilities. When large high-quality pre-cropped datasets are available, this pipeline provides reliable fine-grained evidence for character-level analysis. Transfer learning with compact backbones such as MobileNetV3 is common for single-letter recognition because pretrained ImageNet features adapt quickly, preserve useful shape/texture priors, and require modest computation [17]. Some of the best results have been obtained by Robaa et al who report near-perfect single letter performance with MobileNetV3 variants trained on a large pre-cropped corpus with aggressive data augmentation Robaa et al. * Experiments with curated letter datasets [4] Simpler bespoke CNNs (e.g. SimpleCNN32) are still useful baselines that provide computational efficiency and good reproducibility for serving as conservative reference points for ablation studies [9].

Nevertheless, the basic vulnerability of the letter pipeline is the quality of segmentation. In cases where there is fragmentation or merging of regions in the results of segmentation (often during cursive handwriting, with varying lightness, or excessive teacher markings), the per-letter

 doi:10.20944/preprints202602.1939.v1

classification accuracy decreases greatly, and aggregation principles generate poor page-level classifications. Two complementary solutions to this are pointed out in the literature: (1) train stronger segmentation (or detection) algorithms to generate cleaner letter crops and (2) devise strategies of aggregation that are robust to noisy or missing crops - such as weighted by classifier confidence, or by ignoring low-confidence crops [26,27]. Character pipelines are therefore extremely effective in situations where the quality of the crop is high and datasets are plentiful, but in situations where segmentation is intrinsically difficult or data for isolated characters is scarce.

Segmentation, Object Detection and Hybrid Designs

Segmentation quality is reported to be the single most important factor as a determinant of end-to-end letter-pipeline robustness. Classical methods of image processing based on connected component analysis, contour filtering, morphology cleaning and heuristic thresholds are still widely used as they are simple and do not require any annotation other than page-level labels [15]. These approaches work well for medium quality scans with good stroke contrast and encounter problems with fused ligatures, overlaps and complex noise. Alternatively, object-detection models (e.g., YOLO family) learn joint localization and classification of letters, mitigating fragmentation by predicting context-sensitive bounding boxes [11]. The synthetic-data strategy - synthesizing the image of words by composing pre-cropped examples of letters and training detectors on these synthetic layouts - provides fast detectors and exemplary performance in controlled settings [11]. Nevertheless, synthetic pretraining often creates gaps between domains that restrict their extrapolation to real classroom data, and the quality of detectors is also sensitive to the quality of annotations as well as the representativeness of synthetic augmentations.

The compromise of hybrid two-stage pipelines seems to work: a detector or contour stage suggests candidate regions, and a more powerful per-crop classifier completes and checks its suggestions so that a compromise can be achieved, as well as allowing the results to be readily interpreted. Hybrid pipelines also enable aggregation rules based on both confidence of detector and classifier, which is very beneficial when some of the crops are blurred due to ink smudging or erasures [31]. Annotation cost is the primary barrier to detector adoption; when annotation budgets are limited, careful contour-based heuristics and conservative filtering (extent, aspect ratio and area thresholds) remain pragmatic alternatives. The literature shows that empirical comparisons between detectors show that that fragmentation and false negatives are minimized by using the detection-based approach at the cost of annotation effort and may give brittle domain transfer; hybrid pipelines are appealing in cases where partial annotation or weak supervision is present.

Temporal and Sequence-Aware Modeling

Temporal dynamics: stroke order, pen velocity and pause patterns provide diagnostic information that cannot be obtained from static images. Studies based on the capture of tablet or pen-trace data indicate that the profiles of dyslexic writing are often characterized by altered timing, more frequent pauses, and different sequencing of strokes, that is used by sequence models [3,10]. Liu et al. combined CNN features with positional encodings with LSTM + attention modules to model Chinese dictation sequence, and achieved great improvements in sensitivity and AUC when temporal order information was available [10]. In scanned static documents there are no native temporal signals; however, some positional regularities can be approximated (e.g., via stroke-thinning and curvature estimation) and multi-crop aggregation can capture coarse ordering information.

Integrating of temporal modeling requires special capture hardware or special datasets annotated with information on strokes. When dynamic capture is available, sequence-aware methods expand the set of diagnostic features well beyond static spatial descriptors. In the case of only static scans being available, practitioners have to consider the marginal benefit to be gained from approximated temporal features in favor of their additional complexity and potential fragility.

 doi:10.20944/preprints202602.1939.v1

Explainability and Deployment in Educational Settings

Explainability is a precondition for their deployment in schools as it should be. Teachers and clinicians need the available evidence to inform their referral decisions to be interpretable, and to be able to reconcile this evidence with pedagogical observations. Grad-CAM and other saliency-based class activation methods have become the de facto standard to generate intuitive heatmaps that localize regions of the image that affect the decision of a classifier [13]. Recent studies of dyslexia-handwriting use a combination of high-performing classification and Grad CAM overlays to achieve per-crop and per-page visual explanations, which makes them more transparent and allows them to be analyzed for errors. [4,5] However, methods for explainability have been documented to have limitations: saliency maps identify correlations and not causal mechanisms, can be unstable when input is slightly perturbed, and are widely misinterpreted if they are provided without guidelines and/or confidence context [14]. Educational AI surveys emphasise that visual explanations should be paired with clear confidence estimates and practical guidance that distinguishes screening (risk flagging) from clinical diagnosis [12].

Real-world deployments should therefore combine visual explanations with interfaces that (a) display per-crop confidence, (b) document common failure modes (e.g., poor lighting, heavy teacher annotations), and (c) include explicit usage disclaimers. Without such safeguards, heatmaps can engender false confidence in automated decisions.

Critical Synthesis and Open Gaps

The literature shows that there are obvious trade-offs. Page-level models are based on high robustness on cursive scripts and on noisy images, but they need much page-level training data and offer lower interpretability. Letter-level transfer learning models exhibit remarkable accuracy on per-thought curated data sets and are prone to errors in segmentation. It is a weaker segmentation brittle Object detection also reduces brittle segmentation but requires annotation investment or advanced synthetic pretraining that may not generalize. Temporal models provide orthogonal signals which however demand dynamic capture. Explainability enhances acceptability but needs to be incorporated in a prudent manner with the calibration of confidence and user-centered UI design.

A notable gap is the absence of rigorous bilingual comparative studies that systematically evaluate page-level and letter-level strategies across scripts with different segmentation properties (e.g., Latin vs. Arabic). The existing literature is usually restricted to one particular script and assesses either page-level or letter-level approaches separately. The following issues, then, remain (1) how to divide resources of annotating and engineering between page and character-level pipelines when deploying to a multilingual system, (2) how the explainability output of page-centric and crop-centric models qualitatively changes, and (3) how systematically variability in segmentation reliability affects aggregation rules. The current work specifically addresses these limitations by proposing an Arabic page-centric model with reduced reliance on fragile segmentation, large reuse of English letter corpora using MobileNetV3 transfer learning for per-letter classification and experimental investigation of aggregation strategies and behavior of Grad-CAM for both the scripts. The purpose of the cross-script analysis is to give a practical recommendation to practitioners developing bilingual screening instruments that have to work within the constraints of real-world situations of annotation budget, heterogeneity of devices, and classroom imaging conditions.

Methodology

Bilingual Pipeline, Page- vs. Letter-Level Comparison and Explainability

This study implements a bilingual screening framework that runs two complementary strategies and compares them systematically. The first is a page-level classifier that consumes a full scanned handwriting page and outputs a page-level dyslexia-risk label. The second is a character-level pipeline that segments pages into candidate crops, classifies each crop with a letter model

 doi:10.20944/preprints202602.1939.v1

(SimpleCNN or MobileNetV3-Small), and aggregates per-crop probabilities into a page decision using predefined aggregation rules. The page-level pipeline is especially appropriate for Arabic because Arabic's cursive, context-sensitive glyphs and frequent ligatures substantially undermine reliable character segmentation; consequently, Arabic was modelled primarily with an end-to-end page network. By contrast, English was modelled chiefly with a letter-level transfer-learning pipeline trained on a large corpus of pre-cropped English letters, with character probabilities converted to page labels via aggregation. An overview diagram of the bilingual pipeline showing the page-level and character-level branches, the segmentation/aggregation flow, and the evaluation splits is provided in Figure 1.

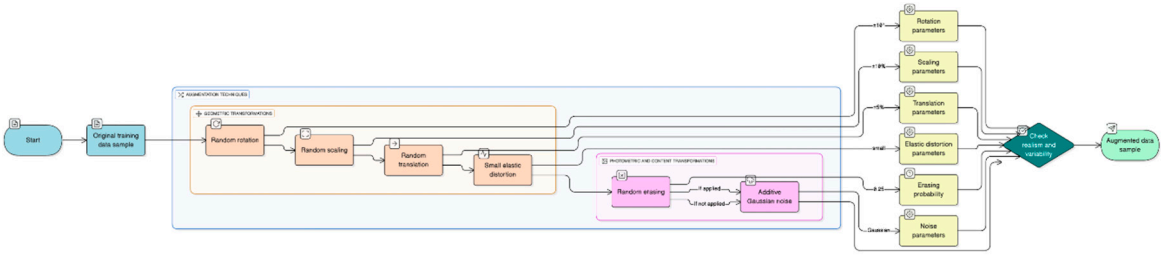

Figure 1. Workflow illustrating the sequence and parameter control of data augmentation techniques applied during training.

To reconcile the scripts' different demands, the evaluation adopted a script-aware, bifurcated strategy. English pages were processed primarily through the segmentation → letter → aggregation route because abundant, high-quality pre-cropped English letter data and MobileNetV3 transfer learning produce robust per-letter predictions. Arabic pages were processed primarily with an EnhancedSmallCNN trained end-to-end on whole-page images because cursive joins and position-dependent letter forms render robust character extraction unreliable. For completeness, cross-evaluations were performed: the English pipeline was also assessed with a page-level model, and the Arabic pipeline was tested using a pragmatic segmentation routine. Those cross-evaluations quantify the contribution of segmentation quality and dataset scale to final performance, and they inform practical migration strategies when isolated Arabic letter corpora are unavailable (see Figure 2(a) for the Enhanced page-level architecture).

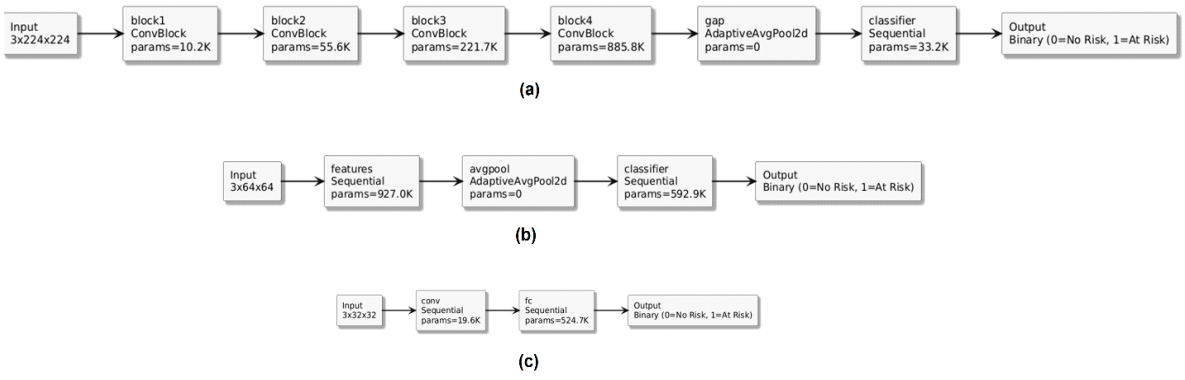

Figure 2. Architecture diagrams for (a) EnhancedSmallCNN (page-level), (b) MobileNetV3-Small (character-level), and (c) SimpleCNN32 (character-level baseline).

Dataset Acquisition and Composition

Three datasets were used in the current study. A dataset comprising 120 English handwriting pages and another one comprising 122 Arabic handwriting pages were acquired from local schools

 doi:10.20944/preprints202602.1939.v1

(after obtaining ethics approval from the institutions and with parents' consent) for page-level classification. Any identifying data was immediately deleted and replaced with a random identifier from the original to guarantee the study remained anonymous. For letter level classification, the publicly available Kaggle Dyslexia Handwriting Dataset [4] which consists of 123,554 labeled letter crops over two classes (dyslexic handwriting and non-dyslexic handwriting) was used. The stratified sampling is used to split the page-level datasets into training, validation and test sets that respect those of class in Table 1. For the English task, this led to 72 training images (37 without any risk of dyslexia and 35 at-risk samples for dyslexia), 18 validation images, and 30 test images. In this latter case the collection of Arabic dataset consists of 73 training images (51 control and 22 at risk for dyslexia, where there is a severe imbalance between the classes), 19 validation images and 30 test images. Dataset sizes and the train/validation/test splits used in all experiments are summarized in Table 1.

Table 1. Dataset Composition.

Train Distribution	Train/Val/Test	Total	Source	Dataset
Control: 37, Dyslexic: 35	72/18/30	120	Schools	English Pages
Control: 51, Dyslexic: 22	73/19/30	122	Schools	Arabic Pages
Stratified split	70%/15%/15%	123,554	Kaggle	Letter Crops

Page-Level Classification

The page-level classification strategy takes each handwriting sample at page level as a holistic input unit, for the purpose of dyslexia risk analysis. This is especially useful for scripts such as Arabic, where more complex joiners and context-dependent letter shapes may make segmentation of individual characters technically cumbersome and error-prone. Because they utilize features from the full-page image and do not require analysis of individual letters, page-based models are able to capture characteristics in the writing on a global level including overall spatial organization, text alignment, variability in letter size, rhythm and frequency of specific dyslexic confounders such as inconsistent spacing or reversal patterns., Page-level methods have been effective for tasks such as writer identification and general handwriting evaluation (e.g., [21]) in English and attempted also for dyslexia detection by applying the ground truth (summary) statistics of writing samples (Asselborn et al. [22]; Yılmaz et al., [23]). For Arabic, the benefit of the page-level model is more significant due to fragmentations and loss of context when trying to segment connected letters [24]. Utilizing holistic analysis, the page-level classification model can effectively deal with the nature of Arabic script, allow cross-language exploration and facilitate finding meaningful dyslexia-related indicators. Finally, this method can be a strong front-line screening tool especially if combined with explainable AI visualizations that let educators and clinicians understand why predictions were made.

Page-Level Preprocessing Pipeline

All page-level input images were processed with a standardized pre-processing pipeline, specifically developed to improve on text-readability while discarding artifacts. The pipeline was composed of five steps: (1) conversion into RGB and normalization; (2) removal of the teacher annotations by using an HSV-based masking for red ink with the masked pixels replaced by white background; (3) conversion into grayscale followed by Otsu's thresholding aimed at generating binary images emphasizing handwritten strokes, 4) morphological closing operations with 3×3 size kernels to smooth stroke edges and fill small gaps, 5) resizing images down to 224×224 pixels in a three channel RGB format, allowing us to use pre-trained models. Data augmentation Training involved significant data augmentation to make the model more robust: random rotations within ±10°, random scaling by ±10%, random translations of ±5%, random erasing with probability 0.25, Gaussian noise addition and small elastic distortions. The augmentation parameters were chosen to

 doi:10.20944/preprints202602.1939.v1

add the necessary variation that can be present in normal handwriting and not distort the characters unrealistically.

Page-Level Model Architecture

CNNs are the most frequently used models based on deep learning to learn discriminative features directly from raw images without reliance on manually designed segmentation rules. For page-level classification, a custom architecture called EnhancedSmallCNN (figure 2(a)) was developed, inspired by lightweight CNN designs commonly used in handwriting and document image classification tasks. This model was trained from scratch. Its custom architecture comprises four convolutional blocks with channel dimensions increasing from 32 to 64 to 128 to 256. Two 3×3 convolutions, batch normalization, ReLU activation, 2×2 max pooling, and dropout (p=0.4) are contained in each block. A 256-dimensional vector is produced by global average pooling, followed by a two-layer MLP classifier (256→128→2). The model contains approximately 1.1M parameters total. A block diagram of the EnhancedSmallCNN page-level architecture is illustrated in Figure 2(a).

Character-Level Classification Approach

In this line of research, one considers classification for individual characters on a handwriting page (or a portion thereof), and the decisions are an aggregated summary for page-/writer-level perceptions. This fine-grained analysis allows for identifying minor orthographic characteristics, such as letter reversals, substitutions and inconsistent letter shapes that are known to be sensitive markers of dyslexia in alphabetic writing systems.

This method is well-adapted to English in which letters require a narrow enough exposure window and are produced with a nearly constant size and shape. Current best practices use CNNs and transfer learning models (e.g., MobileNet, ResNet) to characterize characters and offer feature analysis automatically [22]. More recent work has gone beyond these models via multi-instance learning and aggregation strategies, which are used to boost robustness against segmentation errors as well as handwriting variations [27].

While it has been successful in English and similar scripts, character-level recognition has some drawbacks for cursive or context-dependent writing systems such as Arabic. The nature of handwriting in Arabic and the common practice of ligature formation significantly complicate segmenting the script, leading to erratic breaks, misrecognition of diacritics, and disrupted visual patterning. Thus, relatively few studies have yet been carried out on the use of segmentation features and training when developing a dyslexia screening system for Arabic (most still focusing purely on holistic page-level methods [25]). The character-level approach therefore still provides within the same framework interpretable fine-resolution markers of dyslexic handwriting, where feature level explainability is a valuable asset to assist teachers and parents in interpreting predictions [35], particularly in scripts where isolated-character boundaries aid interpretable visual inspection.

Letter Segmentation and Filtering

Letter segmentation for the English character-level pipeline employs a contour-based connected-component procedure designed for robustness in classroom-collected page scans. Steps are: (1) convert to grayscale and apply Otsu's thresholding for binarization, which automatically selects a global threshold that minimizes intra-class variance [33]; (2) invert the binary image so foreground strokes become bright, then apply morphological closing followed by opening with a 3×3 rectangular kernel to smooth stroke boundaries and remove small speckle noise; (3) extract external contours and compute axis-aligned bounding rectangles for each contour; (4) apply conservative filtering heuristics to remove spurious regions using minimum area, minimum height, aspect-ratio bounds (width/height) and an extent ratio (contour area / bbox area); and (5) pad each retained bbox by 6 pixels and clip to image boundaries to include stroke tails and anti-aliasing. Crops failing a simple contrast check (max − min intensity below a small threshold) are discarded. These steps produce

 doi:10.20944/preprints202602.1939.v1

candidate character images that are then passed to the letter-level classifiers. Representative segmentation outputs, including typical successful crops and common failure cases (merged ligatures and small noisy contours), are presented in Figure 3. The approach follows established offline handwriting segmentation practice and is tuned empirically on a validation split to balance keeping true letters versus removing noise [15,33].

The filter thresholds (e.g., minimum area, minimum height, aspect-ratio limits, extent threshold and padding) were chosen by combining prior domain practice and empirical validation. Connected-component extent and aspect-ratio filtering are common in document analysis to reject non-character ink [15]; we then performed grid-style tuning on the validation set to select operating points that optimize page-level aggregation performance (weighted F1) instead of per-crop accuracy alone. A sensitivity visualization that illustrates how segmentation thresholds affect crop quality and resultant aggregation is shown in Figure 3. Typical values used in our experiments were minimum area = 200 px, minimum height = 10 px, aspect ratio ∈ [0.25, 6.0], extent ≥ 0.03 and padding = 6 px; a sensitivity analysis is provided in Appendix A showing performance trends when these parameters vary by ±25%, demonstrating stability of the end-to-end screening metric across a reasonable parameter range.

Figure 3. Preprocessing Examples. Shows before/after for page and letter samples.

Character-Level Preprocessing

Character-level preprocessing converts each retained crop to a standardized input for the letter models: (1) automatic contrast enhancement, (2) fixed-threshold binarization (threshold=90) for MobileNet training experiments, (3) resizing to model-specific dimensions (32×32 for SimpleCNN, 64×64 for MobileNetV3-Small), and (4) channel replication to 3 channels plus ImageNet normalization when using transfer learning. This pipeline preserves stroke morphology while producing consistent inputs for both scratch-trained and pretrained models.

Character-Level Model Architectures

We started model building by developing a lightweight baseline architecture consisting of two convolutional blocks followed by flattening and a fully connected layer with dropout was implemented for comparison. Design for 32×32 input resulted in approximately 600K parameters. This model was trained from scratch and serves as a performance baseline to highlight the benefit of transfer learning. We call this model the SimpleCNN32 (letter-level baseline)

Subsequently, we built and tested more complex models. The experiments led us to choose MobileNetV3-Small pretrained on ImageNet [17] as the primary transfer learning model, with replacement of the classification head by a custom two-class layer. Two stages occurred in fine-tuning: first training only the new head with frozen backbone with a small learning rate and for only 2 epochs or passes, then joint fine-tuning where the backbone is also retrained but even a smaller learning rate (head: 1e-3, backbone: 1e-4). Among various architectural configurations, this model

 doi:10.20944/preprints202602.1939.v1

provided the highest character-level accuracy and was therefore selected as the main model for subsequent analysis. The MobileNetV3-Small transfer-learning backbone and the SimpleCNN32 baseline architecture are shown schematically in Figure 2(b).

Training Protocols

The training methods were designed to fit each model type and data through the use of specific hyper-parameters. In general, each choice of hyperparameter was based on the size and distribution of the dataset, characteristics of the model architecture and previous literature, in order to maintain a good trade-off between stability, efficiency and performance.

For example, to alleviate class imbalance, Focal Loss [18] with a focusing parameter γ=2.0 and the class-weighted α was used to suppress the relative loss of well-classified examples comparing to that from misclassified ones, so as to attract model results more on difficult or underrepresented classes. AdamW optimization [19] was selected due to its adaptive learning rates and decoupled weight decay, facilitating better convergence and generalization. Page-level models were trained with smaller batch sizes of 16–32 due to larger input sizes and lower memory consumption (learning rate=1e-3), which have been found to show stable convergence from preliminary experiments. The learning rate was adaptively changed with the scheduler according to the macor-F1 score during the validation phase in order to let the optimizer be able to fine-tune weights, when no significant improvement is achieved. Early stopping with a patience of 6–8 epochs avoided the overfitting and a maximum of 60 epochs was chosen to get enough training iterations.

Large batch sizes (64–128) were applicable for letter-level models because of the small input size. Training of SimpleCNN32 which was performed during 5–20 epochs with the learning rate fixed in 1e-3 (optimal experimentally to get stable performance). Pretrained feature extractor and classifier head of different learning rates lead MobileNetV3-Small to adapt slowly after fine-tuning, while the classifier learned fast. To prevent overfitting, early stopping was used and the progress on validation weighted F1 was monitored. To address the problem of class imbalance, weighted random sampling and class-weighted loss were adopted We applied clipping of the gradients to enhance training stability by preventing exploding gradients. Lastly, all experiments were reproducible with fixed random seeds.

Models Evaluation

To evaluate the performance of the developed models, we used the standard metrics of accuracy, precision, recall and F1-score on the test sets. There are two possibilities to aggregate predictions per letter into a decision at the page level. The all-letters approach takes the average of predicted probability for dyslexia over all sub-segmented letters on a page and assigns equal importance to each letter in the determination of page label. In comparison, the accepted-letters approach keeps only letters with high confidence from the model (probability greater than 0.7), which decreases the importance of uncertain or noisy predictions. This method can enhance robustness by only considering high-confidence predictions and discarding misleading letters

Results

Summary statistics for Arabic page-level and English letter-level experiments are presented in Table 2 and Table 3, respectively. The system was evaluated using standard metrics such as accuracy, F1 score, precision and recall, confusion matrices, as well as average probabilities followed by thresholding. As for the performance, the EnhancedCNN model obtained a validation and test accuracy of 74% and 77%, respectively. In the character-level classification problem, we compared with a SimpleCNN baseline that achieved 96% test accuracy and investigated MobileNetV3-Small which was fine-tuned using transfer learning by pre-trained weights that reached 99% test accuracy. Here we used a strategy that made the most of pretrained feature representations while fine-tuning the classifier on our dataset, resulting in better performance and faster convergence.

 doi:10.20944/preprints202602.1939.v1

For enhancing the trust of user, we have integrated with Grad-CAM visualizations to visualize diagnostic areas in handwriting samples. The results show that page-level models are more efficient when dealing with cursive scripts (e.g. Arabic) and letter-level classifier performs better on high quality segmented crops. Arabic page-level validation and test metrics are reported in Table 2.

Table 2. Arabic EnhancedCNN Results.

Comments	Recall	Precision	Weighted F1	Accuracy	Dataset
Performance is affected by class imbalance and cursive complexity.	0.74	0.72	0.73	0.74	Validation
Lower generalization than English; suggests need for more Arabic samples.	0.77	0.76	0.76	0.74	Test

Table 3. English Letter Models Results.

Comments	Recall	Precision	Weighted F1	Accuracy	Dataset
Strong, balanced performance; stable training.	0.96	0.96	0.96	0.96	Simple CNN Validation
Slight drop; good generalization; minor style mismatch.	0.96	0.96	0.96	0.96	Simple CNN Test
Excellent, consistent results; well-learned features.	0.98	0.98	0.98	0.98	MobilnetV3-small Validation
Very strong generalization; best overall model.	0.99	0.99	0.99	0.99	MobilnetV3-small Test

Explainability Visualization

To increase the interpretability of the model, visual explanations for letter-level predictions were generated using Grad-CAM. Gradient-weighted Class Activation Mapping (Grad-CAM) [13] is a model explainability method for deep CNNs designed to localize image regions that most influence the predicted class. Representative Grad-CAM overlays for true positives, false positives and ambiguous crops (used to illustrate typical attention patterns) are presented in Figure 4. Utilizing gradient information of the final convolutional layer, Grad-CAM produces heatmaps that visually direct a user towards diagnostically relevant regions. This not only helps to interpret the model's decision process but also enables domain experts to validate that the system attends at relevant morphological and structural handwriting pattern. In the present investigation, we were the first to have utilized Grad-CAM for letter level dyslexia screening in English and Arabic scripts filling in between black-box AI decisions on one hand and human interpretability on the other, by directly supporting referral decisions by teachers when combined with confidence scores and visual explanations" as shown in the Grad-CAM paragraph above.

Examination of the Grad-CAM overlays on all crops demonstrated consistent trends in which visual information was given emphasis by the model. For true positive crops. the heatmaps were succinctly concentrated around morphological relevant regions such as stroke intersections, ascenders, descenders and character terminals. This is the same structural information handwriting analysts use to differentiate between character shapes. In Crop 1, the model's attention had also adhered to the curve of the ascender, while in Crop 4, the heatmap stressed lower loop and terminal stroke, suggesting confident and interpretable reasoning.

In comparison, mis-classified or ambiguous predictions like in Crop 6 and Crop 9 presented diffuse activation patterning or completely incorrect locations for the attention was placed on background regions or non-informative stokes. This separation indicates that the model could not reach a consensus decision, which is in accordance with its low confidence values and misclassifications. This trend indicated the uncertainty or even indistinctive discriminative characteristic, demonstrating that the model design and data preprocessing were essential. The addition of such explainability tools not only engenders trust between educators and clinicians by rendering the AI decisions more interpretable but also meets an important need in school settings

 doi:10.20944/preprints202602.1939.v1

where clear actionable evidence is needed to inform referral and screening decision-making for dyslexia. Visualizing how and where the model activates during its decision making, Grad-CAM promotes user's trust for responsible application of AI based screening tools in actual practice.

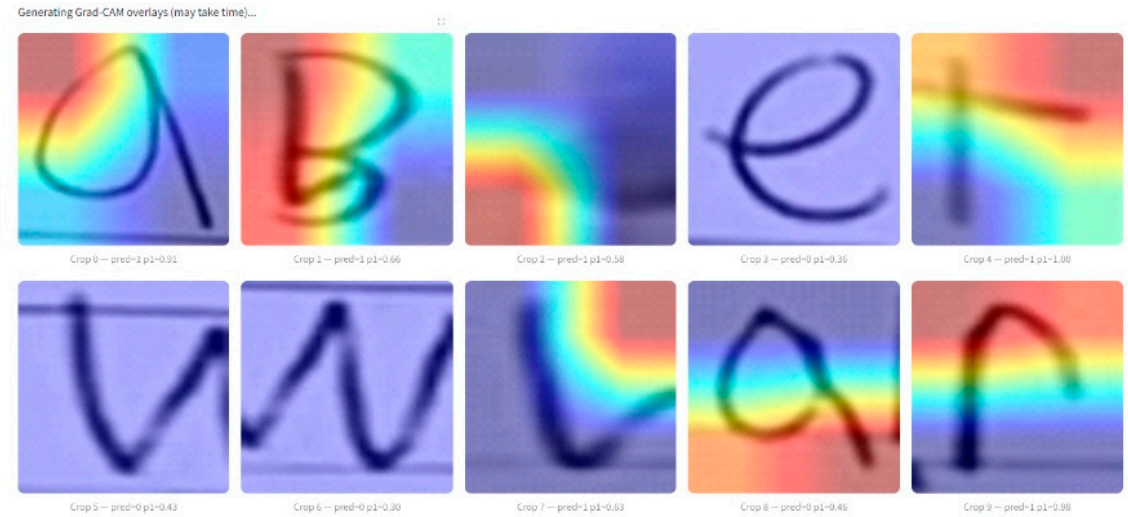

Figure 4. Grad-CAM visualization examples showing original letter crops with overlaid heatmaps highlighting discriminative regions for correctly and incorrectly classified samples.

Discussion

English character-level models were trained using a large pre-cropped letter dataset and transfer learning (MobileNetV3-Small), which produced high per-letter accuracies that could be reliably aggregated to page-level decisions. Arabic poses specific challenges due to cursive connections and context-dependent letter shapes; therefore, rather than applying the English character pipeline directly, we trained an Arabic page-level model (EnhancedSmallCNN) on full-page images. This decision was motivated by the lack of a large, labeled corpus of isolated Arabic character crops and by the fundamentally different visual structure of Arabic script, which makes character isolation and transfer from English less effective.

Interpretation of Findings

We observe significant differences between page level and letter-level modeling. The performance of page-level models remained strong even when segmentation fails, for instance, the test accuracy on Arabic pages is 0.77, which is suitable for preliminary screening. Exploiting transfer learning significantly boosted English letter-level model, which achieved test accuracy of 0.99, indicating the importance of pre-trained networks. The performance of Arabic models was constrained by a lack of large, high-quality datasets indicating the need for more diverse data. Ultimately, Grad-CAM offered explainability by imaging which input areas affected predictions and making models more transparent.

Comparison with Prior Work

The English letter-level scores obtained in this paper (i.e. 0.99 accuracy with MobileNetV3) are similar to those of Robaa et al. [4], achieved between 99% and 99.6% accuracy on the Kaggle Dyslexia Handwriting Dataset (KdHWD). They used MobileNetV3, along with a similar transfer learning approaches as ours. Indeed, their data were all clean characters that had been pre-cropped to minimize variation and avoid the need for segmentation. This probably accounts for their slight performance advantage, as their models were trained on cleaner and more uniform inputs. Nevertheless, the value of the current study lies not only in letter classification. In contrast to the one by Robaa et al., in this work, a letter-to-page aggregation strategy is proposed to transform the

 doi:10.20944/preprints202602.1939.v1

results at character level into a decision on page-level dyslexia detection, which could be crucial for real-world deployment at educational and clinical sites. In addition, this work directly studies the impact of segmentation quality on end-to-end screening robustness and shows that extraction accuracy narrows down diagnostic confidence. Thus, the proposed framework brings an all-in-one interpretable and deployable screening pipeline while Robaa et al. focus exclusively on character-level classification.

The Simple CNN baseline obtained 96% of test accuracy in this work. This yield is comparable to the 96.4% value obtained by Aldehim et al. [9] under the NIST SD19 handwritten letters dataset using a four-layer CNN. Their higher accuracy is obtained partly due to a more structured and cleaner letter samples in NIST SD19, where they have less variation in writing style as well as image noise.. However, the comparable performance shows that lightweight architecture can still be competitive and enjoy faster training speed and smaller computational cost. While this study and the aforementioned works [28] focus on analyzing static features in **handwriting**, another line of research centers on computationally modeling the **reading process** itself. For instance, Hautala et al. (2024) propose a neural network model design that simulates continuous reading by coordinating word recognition and eye movements. Such models are theoretically significant for understanding dyslexia; they suggest that a principal deficit may lie in "early visuo-orthographic processing" and difficulties in "decoding efficiency". This provides an important theoretical grounding for the present study: the features our models detect in handwriting (such as letter reversals or inconsistent shapes) can be interpreted as the physical and motor manifestations of the same underlying cognitive deficits in visuo-orthographic processing and decoding that continuous reading models aim to simulate. Arabic page-level performance (77% accuracy) fills a gap in the current research.

Although there exists literature on Arabic dyslexia based on reading-based tests and isolated handwritten character classification, to the best of our knowledge, no previous work has introduced an end-to-end handwriting-based page level screening with segmentation, character modeling and aggregated dyslexia decision. Accordingly, the scope of this paper presents a complete page-level handwriting-based screening pipeline for Arabic dyslexia for the first time. Although this moderate accuracy is in line with difficulties known in the context of Arabic handwriting cursive connectivity, contextual shapes of letters and large writer variance it still constitutes a basis for future work. Better performance is also expected when training on larger and diverse handwriting datasets, or using architectures adapted to the patterns of Arabic script. The addition of Grad-CAM for explainability accords with increasing focus on interpretable AI in educational contexts [12,14]; however, our results also demonstrate that localization heatmaps capture correlation rather than causation and must be interpreted thoughtfully by domain experts.

Limitations

This study has several limitations. The dataset used is small (120 pages in English and 122 in Arabic) which constrains the statistical significance of the results. The minority class has more serious damage caused by Arabic class imbalance in learning. The Kaggle dataset, which is not classroom handwriting we seek to infer insightful letter-level pipeline from. We did not model temporal (or sequence) information, although writing dynamics can offer more diagnostic signals [10]. Similarly, cross-linguistic transfer learning was not investigated – thus it is unclear whether models trained on English handwriting could be utilized for Arabic screening purposes (or in reverse).

Ethical Considerations

A screening tool, not a diagnostic device, is what this system is important for ethical deployment in schools. GDPR principles including consent, data minimization, and anonymization were followed by the project. Explicit consent and clear disclosure accompanied collection of all samples. Immediate anonymization of images occurred by removing names and assigning random IDs.

 doi:10.20944/preprints202602.1939.v1

Conclusions and Future Work

This work presented a bilingual handwriting-based screening system for dyslexia, where page-level vs. character-level classification strategies are compared for English and Arabic scripts. Three main findings emerged. First, Arabic page-level classification reached 0.77 test accuracy as an initial sign for the possibility of AI-based screening for Arabic dyslexia despite cursive connectivity and variety in writer. Second, the English letter-level classification can already achieve1 accuracy with MobileNetV3-Small), which verifies the effectiveness of transfer learning for dyslexia detection based on handwriting. Finally, examination of Grad-CAM visualizations in each case revealed stroke level patterns (e.g., intersections, ascenders, descenders) that corresponded with elements utilized for expert methodology and served to make the model interpretable to educational stakeholders.

Our main contributions are as follows: A thorough comparison of page-level and character-level based screening strategies across two different writing systems; A data driven analysis revealing the impact of segmentation quality on end-to-end performance; The proposal, adoption and illustration of visual explainability throughout the classification pipeline, to provide insights into how predictions are made; and The development of one of the first deep learning-based screening system specifically designed for Arabic dyslexia.

There are a number of natural avenues for future work suggested by these results. First, enriching the enriching the dataset with additional, diverse handwriting samples might help to enhance generalization for the current system (for which learning curves do not seem saturated and where performance is still influenced by sample variability). Second, the quality of the segmentation having been proved to have an impact on the page-level performance, learning letter localization methods (e.g., object detection models trained on annotated pages) could help improve robustness and consistency in segmentation.

Other research opportunities also go beyond what directly tested here. Further lines of research could model temporal writing dynamics as well as explore the contribution of cross-linguistic transfer in both acquisition and diagnosis between English and Arabic handwriting, to test with teachers' pilot studies addressing usability and decision threshold selection in real classroom environments. Collectively, such work would provide support for ongoing refinement of valuable, interpretable, and linguistically sound screening instruments that lead toward earlier and more equitable identification of dyslexia.

References

1. Van Heuverswyn, E., Gosse, C., and Van Reybroeck, M., 2024. Handwriting difficulties in children with dyslexia: Poorer legibility in dictation and alphabet tasks. Dyslexia, 30(2), pp.e1767.
2. Snowling, M.J. and Hulme, C., 2012. Annual research review: The nature and classification of reading disorders. Journal of Child Psychology and Psychiatry, 53(5), pp.593-607.
3. Berninger, V.W., Nielsen, K.H., Abbott, R.D., Wijsman, E. and Raskind, W., 2008. Writing problems in developmental dyslexia. Journal of School Psychology, 46(1), pp.1-21.
4. Robaa, M., Balat, M., Awaad, R., Omar, E. and Aly, S.A., 2024. Explainable AI in Handwriting Detection for Dyslexia Using Transfer Learning. arXiv preprint arXiv:2410.19821.
5. Rangasrinivasan, S., Suresh, S., Olszewski, A., Setlur, S., Jayaraman, B. and Govindaraju, V., 2025. AI-Enhanced Child Handwriting Analysis: A Framework for the Early Screening of Dyslexia and Dysgraphia. SN Computer Science, 6(5), pp.1-26.
6. Abandah, G.A., Younis, K.S. and Khedher, M.Z., 2014. Handwritten Arabic character recognition using multiple classifiers. In Proceedings of the 5th International Conference on Signal Processing and Communication Systems, IEEE, pp.1-8.
7. Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence. IEEE Access, 6, pp.52138-52160.
8. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature, 521(7553), pp.436-444.

 doi:10.20944/preprints202602.1939.v1

9. Aldehim, G., Rashid, M., Alluhaidan, A.S., Sakri, S.B. and Basheer, S., 2024. Deep learning for dyslexia detection: a comprehensive CNN approach. Journal of Disability Research, 3(2), p.20240010.
10. Liu, H.W., Wang, S. and Tong, S.X., 2024. DysDiTect: Dyslexia Identification Using CNN-LSTM-Attention. Brain Sciences, 14(5), p.444.
11. Fink, N., 2025. Explainable YOLO-Based Dyslexia Detection in Synthetic Handwriting Data. arXiv preprint arXiv:2501.15263.
12. Holstein, K., McLaren, B.M. and Aleven, V., 2019. Co-designing a real-time classroom orchestration tool. Journal of Learning Analytics, 6(2), pp.27-52.
13. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-CAM: Visual explanations from deep networks. In Proceedings of the IEEE International Conference on Computer Vision, pp.618-626.
14. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions. Nature Machine Intelligence, 1(5), pp.206-215.
15. Lorigo, L.M. and Govindaraju, V., 2006. Offline Arabic handwriting recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5), pp.712-724.
16. Boufenar, C., Kerboua, A. and Batouche, M., 2018. Investigation on deep learning for off-line handwritten Arabic character recognition. Cognitive Systems Research, 50, pp.180-195.
17. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V. and Le, Q.V., 2019. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.1314-1324.
18. Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, pp.2980-2988.
19. Loshchilov, I. and Hutter, F., 2019. Decoupled weight decay regularization. In International Conference on Learning Representations (ICLR).
20. Zaibi, T. and Bezine, H., 2024. Early detection of learning disabilities through handwriting analysis. Procedia Computer Science, 246, pp.3702-3712.
21. Bulacu, M., & Schomaker, L. (2007). Text-Independent Writer Identification and Verification on Offline Handwriting: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4), 701–717.
22. Asselborn, T., Chapatte, M., & Huber, P. (2018). Automatic Dysgraphia Screening Using Handwriting Analysis. Scientific Reports, 8, 16256.
23. Yılmaz, M., Kılıç, F., & Gürbüz, E. (2023). Deep Learning-Based Detection of Dyslexia in Handwritten Texts. Computers in Biology and Medicine, 155, 106570.
24. Abd-Almageed, W., Elhoseiny, M., & Mahmoud, T. (2018). Arabic Handwritten Script Segmentation Based on Deep Learning. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR).
25. Alhaddad, A., Al-Maadeed, S., & Ali, S. (2021). End-to-End Arabic Handwriting Recognition Using Deep Neural Networks. Pattern Recognition Letters, 142, 35–42.
26. Zgarbová, H., Doležel, M., & Haluška, J. (2023). Handwriting-Based Dyslexia Detection Using Deep Neural Networks: Letter-Level Approaches versus Page-Level Features. Pattern Recognition Letters, 171, 86–92.
27. Frid, A., Breznitz, Z., & Breznitz, D. (2019). A Two-Stage Deep Learning Model for Handwriting Dyslexia Detection. In Proceedings of the 2019 IEEE International Conference on Data Mining Workshops (ICDMW), 120–127.
28. Hautala, J., Saarela, M., Loberg, O., & Kärkkäinen, T. (2024). A design for neural network model of continuous reading. Cognitive Systems Research, 88, 101.
29. Koegel, L.K., Koegel, R.L. and Smith, A., 1997. Variables related to differences in standardized test outcomes for children with autism. *Journal of Autism and Developmental Disorders*, *27*(3), pp.233-243.
30. Spoon, Katie, David Crandall, and Katie Siek. "Towards detecting dyslexia in children's handwriting using neural networks." In *Proceedings of the international conference on machine learning AI for social good workshop, Long Beach, CA, USA*, pp. 1-5. 2019.

 doi:10.20944/preprints202602.1939.v1

31. Patil, S.P., Apare, R.S., Borhade, R.H. and Mahalle, P.N., 2024. Automated Dyslexia Screening Using Children's Handwriting in English Language with Convolutional Neural Network and Bidirectional Long Short-Term Memory Model. *Engineered Science*, *32*, p.1345.
32. Alkhurayyif, Y. and Sait, A.R.W., 2024. A Review of Artificial Intelligence-Based Dyslexia Detection Techniques. *Diagnostics*, *14*(21), p.2362.