Convolutional neural network DenseNet in classifying dyslexic handwriting images

Chelsea Zaomi Pondayu[1,2], Widodo[1,2], Murien Nugraheni[2,3]
[1]Department of Informatics Education, Universitas Negeri Jakarta, Jakarta, Indonesia
[2]Research Group on Machine Learning and Natural Language Processing, Universitas Negeri Jakarta, Jakarta, Indonesia
[3]Department of Information Systems and Technology, Universitas Negeri Jakarta, Jakarta, Indonesia

Article Info

Article history:

Received Sep 27, 2025
Revised Nov 28, 2025
Accepted Dec 13, 2025

Keywords:

Convolutional neural network
Deep learning
DenseNet
Dyslexia
K-fold cross validation

ABSTRACT

Dyslexia is a specific learning disability (SLD) associated with word-level reading difficulties and often manifests in childhood handwriting through irregular spacing and inconsistent letter sizing, due to shared phonological and orthographic processing. Early identification is critical; however, traditional diagnostic procedures are time-consuming and unsuitable for large-scale screening. This study aimed to develop a handwriting analysis at the paragraph-level using a DenseNet121 convolutional neural network (CNN) model as a low-cost dyslexia screening tool for resource-constrained educational settings. One hundred English handwriting images were preprocessed and standardized into two hundred samples, with 70% of the dataset evaluated using 4-fold cross-validation and the remaining 30% used for testing. The model achieved 90% test accuracy and 92.86% training accuracy, significantly outperforming a random forest baseline that reached 83.57% train accuracy and 63.33% test accuracy, with statistical significance confirmed by McNemar’s test. The main contribution of this study is the demonstration that a lightweight, single-architecture DenseNet121 using paragraph-level analysis can achieve competitive performance compared to prior studies that relied on more complex hybrid models and character-level analysis, while requiring substantially lower computational resources and simplified pipeline. These findings indicate that DenseNet121 provides a robust and low-cost solution for preliminary dyslexia screening in resource-limited educational environments.

Corresponding Author:

Chelsea Zaomi Pondayu
Department of Informatics and Computer Engineering Education, Universitas Negeri Jakarta
11 Rawa Mangun Muka Raya Road, Pulo Gadung, East Jakarta 13220, Jakarta, Indonesia.
Email: chelseazaomi9@gmail.com

1. INTRODUCTION

Dyslexia as a specific learning disability (SLD) is primarily characterized by difficulties with accurate or fluent word recognition and decoding of unfamiliar words [1]-[3], stemming from deficits in phonological and orthographic processing, which are essential for literacy acquisition [4]. These difficulties arise independently of cognitive abilities and often persist despite effective classroom instruction [2], [3]. Developmental dyslexia affect approximately 7-10% of primary school children worldwide [5], [6], with significant numbers remaining undiagnosed [6]. Beyond literacy difficulties, children with dyslexia tend to exhibit negative self-perception and poor academic performance [7]; however, early identification and tailored educational interventions can improve both self-concept and academic outcomes [2], [4], [7]. Traditional dyslexia identification involves comprehensive psychoeducational evaluations with varying protocols due to long-standing debates over dyslexia markers [3], making these assessments costly and inefficient for mass screening. Recent consensus emphasizes core characteristics—difficulty with accurate and fluent word-level reading [3] and advocates for screening models that prioritize early assessment based on reading and spelling performance [2].

To support accessible early identification, recent studies have explored machine learning (ML) and deep learning (DL) approaches for automated dyslexia detection across diverse modalities [8], such as electroencephalogram (EEG), eye-tracking, and magnetic resonance imaging (MRI). Seshadri *et al*. [9] achieved a mean accuracy of 96.7% using K-nearest neighbor (KNN) on EEG data, although their study focused on cognitive and attentional measures rather than reading or literacy-related skills. Eye-tracking studies include Svaricek *et al.* [10] achieved 86.65% accuracy using fixation-image visualizations with ResNet18 on a limited dataset, and Vaitheeshwari *et al*. [11], who reported 98% accuracy using virtual reality-based eye-tracking features with a fusion model combining convolutional neural networks (CNN), deep neural networks (DNN), and bidirectional encoder representations from transformers (BERT). Multi-modality studies have also been conducted. Alkhurayyif and Sait [12] applied multiple DL models (MobileNetV3, EfficientNetB7 and Bi-directional long short-term memory/Bi-LSTM) to MRI, functional MRI (fMRI), and EEG data, achieving 98.6%, 98.9%, and 98.8% accuracy, respectively. Although highly effective, these modalities require specialized equipment and expertise, making them unsuitable for large-scale, low-cost educational screening [8], [12]. Furthermore, recent study suggests that despite neurobiological differences in individuals with dyslexia, it is more effective and reliable to identified dyslexia by behavioral indicators of literacy difficulties instead of neuroimaging [3].

Given that reading and writing are strongly correlated [13], and research confirms children with dyslexia exhibit poorer handwriting legibility and slower writing speed even in simple tasks [14], both skills reflect shared deficits in phonological and orthographic processing [4]. Therefore, handwriting analysis has emerged as a practical tool for behavioral screening. Handwriting samples exhibiting spatial features—such as inconsistent spacing, irregular letter formation, and overall disorganization [15], [16] can be easily collected in educational settings without high-cost equipment [17]. Several DL studies have explored this modality. Patil *et al.* [16] achieved 95.6% accuracy using a CNN-Bi-LSTM hybrid architecture for holistic handwritten analysis on the IAM dataset and a primary school dataset. Alqahtani *et al.* [18] achieved 99.33% accuracy classifying three categories of handwriting characters (normal, reversed, and corrected) using a CNN-SVM hybrid model. DysDiTect [19] employed a CNN-LSTM hybrid model, achieving an accuracy of 83.2% on a Chinese character dataset. Zaibi and Bezine [20] achieved 99% accuracy with both gradient boosting (GB) and random forest (RF) models on the "Handyg23" Arabic paragraph/text dataset. Additionally, a cross-modality study by Sait and Alkhurayyif [21] combined handwriting characters data with MRI and EEG data, achieving 99.1%-99.2% accuracy using hybrid transformer-based models.

Despite these advances, significant research gaps remain. First, very few handwriting-based studies perform paragraph-level analysis; most state-of-the-art approaches rely on character-level [18], [19], [21] or sequential analysis [16], which require complex hybrid architectures and limit the development of models capable of learning holistic graphomotor patterns instead of isolated character shapes. Second, paragraph-level dyslexic handwriting datasets in alphabetic languages are scarce. Studies using non-alphabetic scripts such as Chinese [19] and Arabic [20] may not generalize to alphabetic systems due to different graphomotor demands. Third, many existing datasets are small and proprietary, making reproducibility challenging and necessitating models that can maintain stable performance under small-data conditions. Finally, although hybrid architectures achieve high accuracy, their complexity hinders real-world deployment in typical schools with limited hardware.

To address these gaps, the objective of this study is to develop a computationally efficient, single-architecture dyslexia screening model using a CNN-based DenseNet121 that performs classification through paragraph-level handwriting analysis, enabling low-cost, scalable preliminary screening in resource-constrained educational environments. DenseNet121 was selected for its robustness against overfitting, particularly on small datasets [22]. This study uses an open-source dataset of English paragraph-level handwriting samples to ensure reproducibility and practical relevance for alphabetic writing systems. The main contributions of this study are:

i) Introducing a single-architecture DenseNet121 model with a preprocessing pipeline for dyslexia classification, demonstrating that a lightweight CNN can achieve accuracy competitive with prior studies that used hybrid or ensemble models, while requiring significantly lower computational resources.
ii) Implementing paragraph-level analysis on alphabetic handwriting to capture holistic spatial features; unlike most prior studies that focus on character-level or word-level segmentation, this approach addresses the dataset scarcity and removes the need for time-intensive segmentation pipelines.
iii) Demonstrating DenseNet121's robustness on a small alphabetic handwriting dataset, addressing a critical challenge in medical and educational DL applications where large datasets are often unavailable.

For a clear and coherent flow, the remainder of this paper is organized as follows: section 2 presents the materials and main methods. Section 3 presents the results and discussion. Section 4 concludes the study and provides directions for future work.

2. METHOD

Figure 1 illustrates the overall research workflow adopted in this study, starting from data acquisition to final performance evaluation. The workflow begins with dataset acquisition (section 2.1) and a structured preprocessing pipeline designed to produce clean and standardized scan-like handwriting images (section 2.2). The DenseNet121 model architecture and configuration are detailed in section 2.3., while section 2.4 covers the complete training and validation pipeline, including dataset splitting, defining 4-fold cross-validation (CV), model compilation, and the training and validation process. Model testing, performance metric evaluation, as well as statistical and comparative analysis are explained in section 2.5.

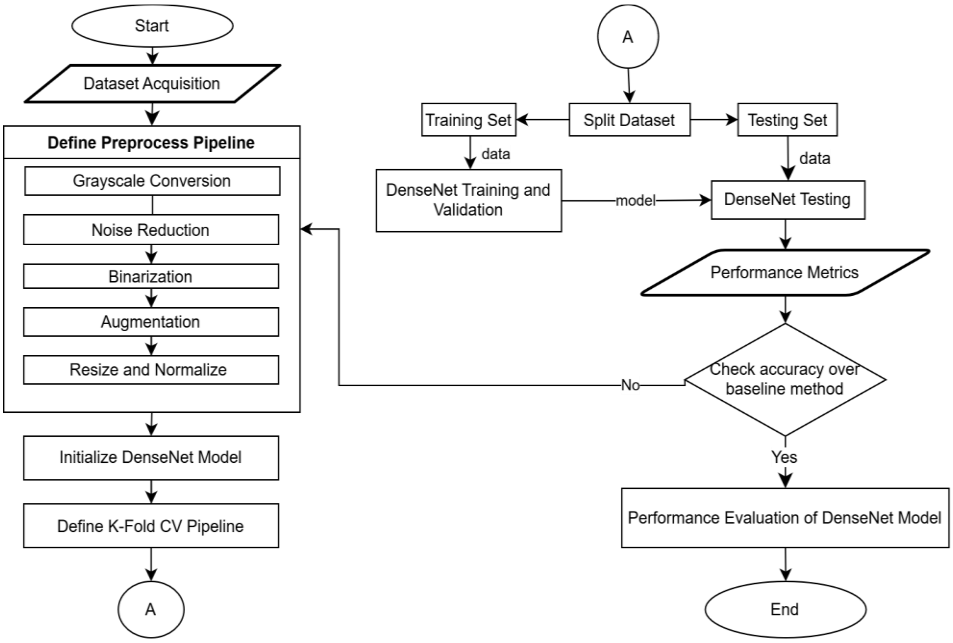


Figure 1. Research workflow

2.1. Dataset acquisition

The dataset was acquired from the open-source GitHub repository "*Dyslexia_Detection"* by user *dlsathvik04*. It contains one hundred English handwritten images, equally divided between dyslexia and non-dyslexia classes. Figure 2 shows samples of non-dyslexic (left) and dyslexic (right) handwriting, which consist of one to two paragraphs. Although this dataset lacks controlled participant metadata (*e.g.*, age, grade level), it was selected for its balanced classes, public availability, and suitability for holistic handwriting binary classification tasks.

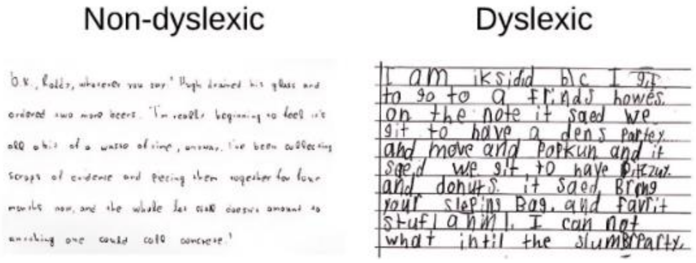

Non-dyslexic
Dyslexic
O.K., Rollo, whatever you say.' Hugh drained his glass and
ordered two more beers. 'I'm really beginning to feel it's
all a bit of a waste of time, anyway. I've been collecting
scraps of evidence and piecing them together for four
months now, and the whole lot still doesn't amount to
anything one could call concrete.'
I am iksidid b\c I git
to go to a frindJ howes.
on the note it saed we
git to have a dens Partey.
and move and Popkun and it
saed we git to have Pitzut.
and donuts. it saed Brmg
your sleping Bag. and favrit
stufl anml. I can not
what intil the slumBrParty.

Figure 2. Dataset samples

2.2. Dataset preprocessing

To transform raw data into clean and standardized inputs [23], this study employed preprocessing techniques suitable for image data, including grayscale conversion, noise reduction, binarization, augmentation, and standardization (resize and normalization). This study applied a specific sequence of grayscale conversion, noise reduction, and binarization, to produce cleaner, scan-like images. Grayscale conversion reduces color complexity and simplifies computations [24]. Noise reduction with a median blur removes noise that can be introduced or emphasized during grayscale conversion [25], [26]. Binarization with adaptive thresholding converts a grayscale image into a clean black–white image, based on the local intensity distribution [27], [28]. After that sequence, augmentation methods such as random rotation and brightness adjustment were then applied, adding another 100 images to the dataset (total images are 200), for enhancing training data and improving generalization [29], [30]. Finally, images were resized to 200x200 pixels with 3 channels and then normalized to ensure pixel values within a reasonable range [23].

2.3. Model architecture definition

CNNs, as a DL algorithm, utilize multiple hidden layers to detect complex patterns in data through convolution and pooling operations [23], [31]. The convolution operation extracts spatial features by replacing traditional matrix multiplications. In most DL frameworks, the convolution is implemented as cross-correlation, without flipping the kernel [23]. It is defined in (1).

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \tag{1}$$

Where *S* is the feature map, *I* is the input, *K* is the kernel, (m,n) is the kernel position, and (i,j) is the output pixel position [23]. Meanwhile, pooling operations reduce spatial dimensions by summarizing local features. Max pooling returns the maximum value in each patch, while average pooling returns the mean of elements in a patch [23]. This study utilized average pooling, defined as:

$$f_{ave}(X) = \frac{1}{N} \sum_{i=1}^{N} xi \tag{2}$$

where *X* is a feature patch, *xi* is the *i*-th element in the patch, and *N* is total number of elements [32].

DenseNet, as a CNN model, employs densely connected layers within dense block, where each layer receives all preceding feature maps and passes its outputs to all subsequent layers. This structure improves information flow, encourages feature reuse, and mitigates the vanishing gradient problem [22]. The dense connectivity is defined as:

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]) \tag{3}$$

with x_l is a feature map of l-th layer, $H_l(.)$ is a composite function, and $[x_0, x_1, \cdots, x_{l-1}]$ is the concatenation of all preceding feature maps [22].

Each composite function H_l consists of batch normalization (BN), rectified linear unit (ReLU) activation, and convolutions arranged as BN → ReLU → 1×1 Conv (bottleneck) → ReLU → 3×3 Conv [22]. BN stabilizes training by normalizing activations [23], [33], and ReLU introduces non-linearity to learn complex pattern [23], [34], the bottleneck layer reduces parameters, while the final convolution extracts spatial features [22]. The BN and ReLU core equations are defined consecutively in equations (4) and (5).

$$H' = \frac{H - \mu}{\sigma} \tag{4}$$

$$g(z) = max\{0, z\} \tag{5}$$

H is a mini-batch of activations, μ is the mean activation per unit, and σ is the standard deviation per unit [23], [33]. While $g(z)$ represent ReLU as identity function and *z* is the input [23], [34].

Transition layers between dense blocks include BN, 1×1 convolution, and 2×2 average pooling to reduce spatial dimensions and control model complexity [22], [35]. While DenseNet generally employs global average pooling followed by a softmax layer, this study used a sigmoid activation with a 0.5 threshold for binary classification of dyslexic and non-dyslexic handwriting. The sigmoid function maps any real-value (*z*) to the interval [0,1], as defined in (6) [23].

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{6}$$

Given these advantages, the pretrained DenseNet121 backbone was selected for its efficiency and relatively small number of parameters, and implemented in TensorFlow/Keras within Google Colab. The model was adapted for binary classification under limited data and computational constraints, as summarized in Pseudocode 1.

2.4. Training and validation pipeline

The dataset of 200 handwriting images was divided into a 70:30 split, with 70% (140 images) for training and validation, and 30% (60 images) for testing. This holdout ratio was selected to mitigate optimistic bias from an overly small test set and to ensure sufficient unseen samples. The 70% portion was evaluated using 4-fold CV with shuffled splits and a fixed random state (42) for reproducibility [36]. Each fold used approximately 105 training and 35 validation samples.

The DenseNet121 model was compiled with the AdamW optimizer (learning rate=1×10^{-4}, weight decay=1×10^{-5}) and binary cross-entropy loss. Training was performed for 15 epochs per fold with a batch size of 16, and model checkpoints were used to save the best-performing weights based on validation accuracy. This 4-fold CV setup mitigates overfitting and bias, ensuring better generalization for small datasets [23]. The complete training process is outlined in Pseudocode 1.

Pseudocode 1. Training algorithm for DenseNet121 with 4-fold CV

```
Input: Preprocessed train dataset D (140 images, 200×200×3), class labels y, number of
folds K=4, epochs E=15, batch size B=16, l earning rate ρ=1e-4, weight decay λ=1e-5,
dropout rates (0.2, 0.4), L2 regularization parameter α, random seed s=42.
Output: Best models {$M_1$,...,$M_4$}, validation metrics {Acc, Prec, Rec, F1} for each fold,
training histories H.

1: Load DenseNet121 backbone with ImageNet pretrained weights and exclude top layers)
2: Freeze all base layers in DenseNet121
3: Construct classification head:
    • Add GlobalAveragePooling2D layer
    • Add Dropout layer (0.2)
    • Add Dense layer (16, ReLU, L2 = α)
    • Add Dropout layer (0.4)
    • Add Batch Normalization layer
    • Add Dense layer (1 unit, Sigmoid)
4: Initialize 4-Fold CV (shuffle=True, random_state=s)
5: Split D into 4 folds {$F_1$,…., $F_4$}

For fold i=1 to 4 do:
   6: Split $F_i$ into training set D_train (~105 images) and validation set D_val (~35
images)
   7: Initialize model $M_i$ with DenseNet121 backbone and classification head
   8: Compile $M_i$ with AdamW optimizer (ρ, λ) and binary cross-entropy loss
   9: Define model checkpoint based on validation accuracy
   Repeat for epoch e=1 to E:
    • Train $M_i$ on D_train (batch size = B)
    • Evaluate $M_i$ on D_val
    • Save best weights if validation accuracy improves
   until epoch E is reached
   10: Load best saved weights for $M_i$
   11: Evaluate $M_i$ on D_val and record metrics (Acc, Prec, Rec, F1)
   12: Save best model $M_i$ and training history
End For

13: Aggregate validation metrics across all K folds
14: Return {$M_1$, $M_2$, $M_3$, $M_4$}, validation metrics, training histories
```

2.5. Model testing and evaluation

The best-performing model from the 4-fold cross-validation was evaluated on the test set, which consisted of 60 images. The primary evaluation tool was the confusion matrix, which captures the types of misclassifications and overall model performance [37]. There are four values used in the confusion matrix, as shown in Table 1. From those values, four evaluation metrics were derived: (i) accuracy measures the ratio of correct predictions to the total predictions; (ii) precision measures the proportion of true positives among predicted positive; (iii) recall measures the proportion of true positives among actual positives; and (iv) F1-score represents the harmonic mean of precision and recall [37]. Each metric is shown in (7) to (10).

Table 1. Confusion matrix values for binary classification [37]

Actual class	Prediction class	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{7}$$

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1 - score = \frac{2(Precision \times Recall)}{Precision+Recall} \tag{10}$$

Additionally, other measurements were also included to strengthen the research. Area under the receiver operating characteristic curve (AUC-ROC) was computed to evaluate classification performance across all decision thresholds, providing a threshold-independent measure of discriminative ability [38]. Confidence scores derived from the sigmoid activation function (6) were analyzed to assess individual classification certainty [39]. To provide a range of plausible population accuracy values, 95% confidence intervals (CI) for test accuracy were calculated using the Wilson score method [40]. CV stability was assessed by computing the mean and standard deviation of accuracy across the 4 folds, with the coefficient of variation (CV%=SD/Mean×100), to evaluate performance consistency [41].

Statistical evaluation was also performed using McNemar's test at α=0.05, to compare DenseNet and the baseline model on the same dataset, with the null hypothesis that both models have equal error rates [42]. The test statistic is:

$$X^2 = \frac{(b-c)^2}{(b+c)} \tag{11}$$

where b=cases where DenseNet correct and RF wrong, c=cases where RF correct and DenseNet wrong. Baseline model used in this study is RF model, with 100 trees, 10 max_depth and 2 min_samples_leaf, trained with identical 4-fold CV settings as DenseNet for fair comparison. RF was selected as a baseline due to its robustness and computational efficiency in classification tasks [43]. Before training the RF, the preprocessed data was flattened into 1D vectors, and reduced to 100 components with principal component analysis (PCA) [23]. If DenseNet's performance failed to exceed the RF baseline in accuracy, preprocessing and hyperparameters were re-evaluated.

2.5.1. Comparative analysis with state-of-the-art models

To evaluate the competitiveness of the proposed approach, this study compared recent handwriting-based dyslexia detection studies using the following criteria: classification performance, model complexity, computational efficiency, data requirements, and practical deployability (hardware requirements, implementation complexity). The comparison focuses on studies published between 2020 and 2025, specifically Patil *et al.* [16], Alqahtani *et al.* [18], DysDiTect [19], and Zaibi and Bezine [20]. These studies represent current state-of-the-art approaches with similar objectives but different methodological choices.

3. RESULTS AND DISCUSSION

3.1. Models' performance results

3.1.1. DenseNet model performance in training and testing

The DenseNet model training histories across 4 folds is illustrated in Figure 3. Each fold demonstrates consistent convergence patterns, with both training and validation accuracy (blue lines) increasing, while training and validation loss (red lines) generally decrease. Training accuracies steadily increased in each fold, ranging from around 80% to over 90%, while validation accuracies consistently exceeded 88%. However, relatively high training/validation losses (>50%) and fold-to-fold variations reflect the challenge of training deep networks on small datasets (200 images). The robust architecture of DenseNet121 can caused model to be overconfident, necessitating aggressive regularization to constrained feature learning capacity while preventing overfitting. Thus, making minor fluctuations and unsmoothed training histories.

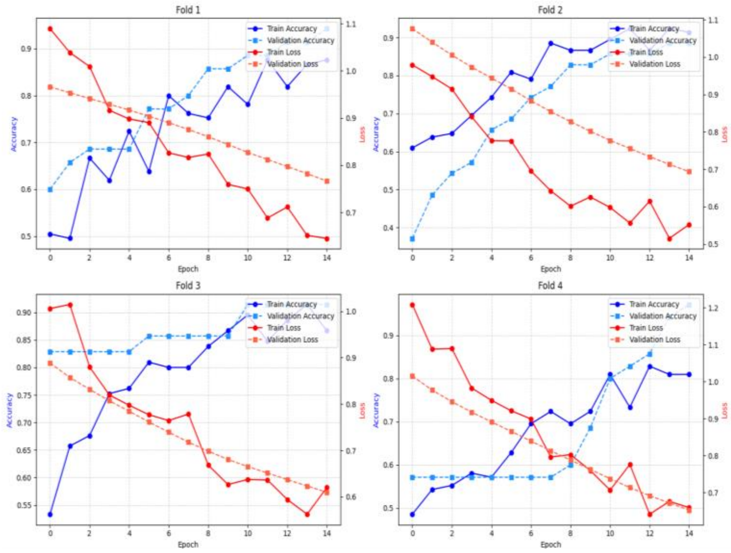


Figure 3. Training and validation history on each fold

The DenseNet model achieved 90.0% test accuracy, as well as the additional metrics, as shown in Table 2. The 95% confidence interval indicates reliable performance estimation despite the relatively small test set, with the interval width of 15.49% reflecting the inherent uncertainty associated with limited sample sizes in educational datasets. Low CV%=3.44% (<5%) demonstrated good stability, shown in mean training accuracy of 92.86% ± 3.19%, confirming robust performance across data splits. The AUC-ROC of 98.44% indicates strong discriminative capacity to separate classes across all classification thresholds. Analysis of prediction confidence scores revealed that dyslexic samples showed broader confidence ranges (8-49%) with consistently low scores, reflecting high model certainty in identifying dyslexic spatial patterns, while non-dyslexic samples exhibited narrower but overlapping ranges (42-64%) with some predictions falling below the 50% classification threshold. This asymmetric confidence distribution explains the model's conservative classification behavior and perfect precision—the model confidently identifies dyslexic patterns but shows uncertainty with some non-dyslexic samples that may exhibit atypical spatial features.

Table 2. Overall DenseNet metrics performances

Metrics	Values
Test accuracy	90%
Test precision	100%
Test recall	80%
Test F1-score	88.89%
Train accuracy	92.86%
Train precision	89.01%
Train recall	98.33%
Train F1-score	93.13%
95% confidence interval	[79.85%, 95.34%]
AUC-ROC	98.44%
CV variability	3.44%
Confidence score (dyslexia)	8-49%
Confidence score (non dyslexia)	42-64%

3.1.2. Comparative performance against baseline

Table 3 presents the metrics performances from proposed DenseNet and baseline RF models, with "Diff" column that represent metric differences between both models. During training, DenseNet demonstrated superior performance across all metrics, achieved 9.29% more accuracy, 19.28% more recall, 10.40 more F1-score and slight 0.32% more in precision. DenseNet also demonstrated superior stability with lower standard deviations across metrics compared to baseline RF (SD: 3.19% vs. 8.42%), indicating more consistent learning. The performance gap widened substantially in testing, where DenseNet outperformed RF by 26.67 % in accuracy, 33.33 % in precision, 26.67 % in recall, and 29.63 % in F1-score, further supporting that the DL approach not only learned more effectively but also generalized better to unseen data.

Table 3. Metrics performances comparison of DenseNet and baseline model

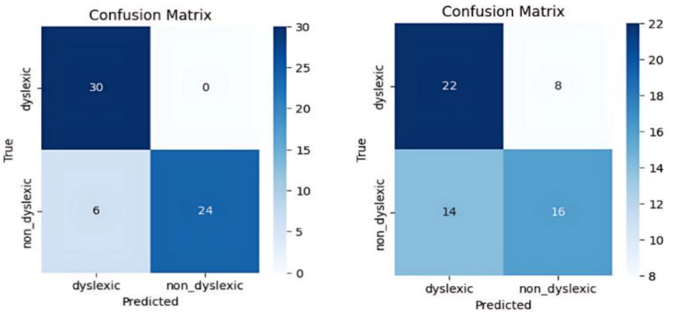
Metric	DenseNet (train)	RF (train)	DenseNet SD	RF SD	Diff (Train)	DenseNet (Test)	RF (Test)	Diff (Test)
Accuracy	92.86%	83.57%	3.19%	8.42%	+9.29%	90%	63.33%	+26.67%
Precision	89.01%	88.69%	8.01%	7.93%	+0.32%	100%	66.67%	+33.33%
Recall	98.33%	79.05%	2.89%	14.24%	+19.28%	80%	53.33%	+26.67%
F1-score	93.13%	82.73%	3.49%	8.97%	+10.40%	88.89%	59.26%	+29.63%

To ensure DenseNet model exceed the baseline RF model, the McNemar’s statistical test were applied (Table 4). McNemar’s test confirmed statistical significance (χ^2=11.25, p=0.000796), with DenseNet correctly classifying 18 additional cases that RF misclassified while RF only corrected 2 cases DenseNet missed (9:1 ratio). This 9:1 ratio of disagreement cases favoring DenseNet demonstrates substantial and statistically significant superiority over machine learning approaches. The p-value of 0.000796 indicates that the probability of observing this performance difference by random chance is less than 0.08%, providing strong evidence that the improvement is genuine rather than artifactual.

Table 4. McNemar’s test summary (DenseNet121 vs RF)

Statistic	Value
Both correct	36
DenseNet only correct (b)	18
RF only correct (c)	2
Both wrong	4
χ^2 (Chi-square)	11.25
p-value	0.000796
α (significance level)	0.05

Figure 4 presents the confusion matrix visualizations for the CNN-DenseNet and RF models. The confusion matrix in Figure 4(a) illustrates that DenseNet model’s DenseNet achieved perfect identification of all 30 dyslexic samples (zero false positives) but misclassified 6 non-dyslexic samples as dyslexic (false negatives), yielding 80% recall. This conservative bias, where the model preferentially erring toward dyslexia when uncertain, is reflected in overlapping non-dyslexia confidence scores (42-64%, some below 50%). In contrast, Figure 4(b) showed RF bidirectional confusion (8 false positives, 14 false negatives), misclassifying 27% of dyslexic and 47% of non-dyslexic cases, indicating fundamental feature representation limitations.



(a) (b)

Figure 4. CNN-DenseNet (a) baseline RF and (b) confusion matrix visualization

3.2. Discussions

3.2.1. Ablation analysis

To validate component contributions, ablation experiments was conducted by systematically removing key elements from the proposed pipeline.

1) Preprocessing pipeline impact: the scan-like image conversion, data augmentation, and standardization proved critical for robust performance. Removing these steps—retaining only basic resizing and normalization—degraded training accuracy from 92.86% to 75.57% (±14.67%) and test accuracy from 90% to 83.33%. More critically, training stability increasing by 360% (SD: 3.19%→14.67%), indicating inconsistent learning across folds. Test precision dropped from 100% to 77.78%, introducing false positives that undermine screening reliability. Raw handwriting images contain excessive noise,

inconsistent lighting, and background artifacts that interfere with spatial feature extraction. The preprocessing pipeline's operations (grayscale conversion, noise reduction, and binarization) standardize inputs into scan-like formats, enabling DenseNet to focus on spatial organization patterns rather than image quality confounds. This 6.67% accuracy improvement and 360% stability increase justify the preprocessing overhead as essential for deployment consistency.

2) Architecture choice: DenseNet121's dense connectivity patterns preserve spatial features, particularly important for capturing subtle disorganization patterns characteristic of dyslexic handwriting. Its 7.06 million parameters balance expressiveness with computational efficiency (208.57 ms/image), unlike hybrid models requiring sequential processing stages (Table 5).
3) Data augmentation: with only 100 original images, augmentation was essential for generalization [29], [30]. The augmentation increased effective training samples and helped the model learn rotation-invariant spatial features, though the small base sample size remains a limitation reflected in the 15.49% confidence interval width.

3.2.2. State-of-the-art comparison

Table 5 presents a comparative analysis between the proposed method against recent handwriting-based approaches. The proposed approach achieved competitive performance while offering distinct practical advantages in deployability and resource efficiency. Unlike hybrid or multi-stage architectures employed in prior works—such as the CNN-BiLSTM model by Patil *et al.* [16] and the CNN-positional-LSTM-attention network in DysDiTect—the proposed model utilizes a single DenseNet121 architecture with 7.06 million parameters. Despite this simpler configuration, our model achieved 6.8% higher accuracy than DysDiTect [19] and maintained competitive performance with Patil *et al.* [16], while demonstrating superior generalization stability through 4-fold cross-validation (CV%=3.44%). Although Alqahtani *et al.* [18] achieved the highest accuracy (99.33% with CNN-SVM), their character-level approach requires substantially more complex implementation. Similarly, Zaibi and Bezine [20] reported 99% accuracy using ensemble models, but required multiple model pipelines and specialized data acquisition tools.

Table 5. Comparing state-of-the-art models

Relevan studies	Classification performance	Model complexity	Computational efficiency	Dataset used	Practical deployability HW (hardware requirments) and Impl (implementation complexity)
Proposed method	Acc: 90%, Prec: 100%, Rec: 80%, F1: 88.89%, AUC: 98.44%	Single-architecture DenseNet121, total parameter: 7,058,017	208.57 ms/img	200 English paragraph images (140 for train in 4-fold CV, 60 for test), balanced classes 50:50	HW: 12 GB RAM, Core i3 2.00 GHz CPU, 128 MB GPU, 500 GB storage Impl: single model, preprocessing pipeline, direct extraction and classification in one model
Patil *et al.* [16]	Acc: 95.6%, Prec: 94.38%, Rec: 91.51%, F1: 92.61%	Hybrid (CNN-BiLSTM with CTC loss), total parameter: 8,743,247	(not stated)	IAM dataset: 1,539 pages, 657 individuals, character-level analysis	HW: (not stated) Impl: hybrid model, preprocessing and word segmentation, CNN + BiLSTM for extraction and sequential analysis
Alqahtani *et al.* [18]	CNN: 98.59%, CNN-RF: 98.44%, CNN-SVM: 99.33% (stating accuracy only)	Single and Hybrid (CNN, CNN-SVM, and CNN-RF), total parameter: (not stated)	(not stated)	176,673 English character images (70/15/15 split)	HW: (not stated) Impl: multiple models, preprocessing + segmentation, CNN + classifier
DysDiTect [19]	Handwriting only: Acc: 83.2%, Sens: 79.2%, Spec: 86.4%, AUC: 91.2% With grade: Acc: 85%, Sens: 83.3%, AUC: 89.7%	Hybrid CNN-positional-LSTM-attention, total parameter: (not stated)	Max 50 epochs with early stopping	100,000 Chinese characters, 1,064 children (483 DD, 581 TD), word-level	HW: (not stated) Impl: complex hybrid, character segmentation, sequential features, transfer learning
Zaibi and Bezine [20]	Best:99% (GB, RF), AdaBoost:97%, SVM-RBF:94% (stating accuracy only)	Ensemble (GB, RF, AdaBoost, SVM), total parameter: (not stated)	(not stated)	120 Arabic samples (ages 7-12), 12 handwriting tasks	HW: Wacom tablet + MovAlyzer software Impl: multiple ensemble models, 12-task protocol

A key distinguishing factor is the proposed method's paragraph-level analysis approach, compared to character-level or word-level methods that require significantly larger datasets: Alqahtani *et al*. [18], DysDiTect [19], and Patil *et al*. [16]. Paragraph-level analysis offers greater ecological validity for screening contexts where educators collect naturalistic writing samples without requiring time-intensive word segmentation. The perfect precision (100%) is particularly valuable for preliminary screening, ensuring that when the model flags potential dyslexia, it is avoiding false positives that cause unnecessary anxiety and inappropriate resource allocation in educational settings.

The proposed method demonstrates superior practical deployability with inference time of only 208.57 ms per image on modest hardware (Intel Core i3 CPU, 12 GB RAM, 128 MB GPU). In contrast, prior studies either did not report computational metrics or required more resource-intensive architectures. The single-model design eliminates the complexity of hybrid pipelines that require separate feature extraction, sequential processing, and ensemble classification stages. This straightforward implementation enables direct deployment via cloud platforms (Google Colab), where educators can upload handwriting photographs, apply the automated preprocessing pipeline, and obtain immediate preliminary screening results without specialized hardware or technical expertise, making it a practical and accessible alternative for real-world dyslexia screening in resource-constrained educational environments.

3.2.3. Limitations

Several technical limitations stem from the experimental design and dataset constraints. The small dataset (200 images) when used on robust DL model caused the need for aggressive regularization (L2, dropout, batch normalization) to prevent overfitting, which also constrained the model's feature learning capacity. This manifested as relatively high training/validation losses (>50%), less smooth learning curves, and potential underfitting despite achieving 90% test accuracy. The 15.49% confidence interval width reflects this sample size limitation, indicating that performance estimates carry inherent uncertainty. Additionally, some handwriting format variations—such as longer or shorter paragraphs, print or cursive styles—introduced uncontrolled variability that may have affected classification consistency. Some students produced brief single-paragraph samples while others wrote extensive multi-paragraph texts, creating heterogeneous spatial complexity across samples. The model's conservative classification bias resulted in six false negatives (non-dyslexic students misclassified as potentially dyslexic), likely representing cases with atypical spatial features such as naturally irregular handwriting, fatigue effects, or rushed writing. Paragraph-level spatial analysis alone cannot capture word-level spelling errors, letter reversals, or phonological patterns that provide complementary diagnostic information in clinical dyslexia assessment. Consequently, the model should be interpreted only as a preliminary assistive tool for research—not as a standalone diagnostic instrument in educational settings.

Although the full preprocessing, training, and testing pipeline is consolidated into a single Google Colab notebook—making the model technically deployable in real educational settings—the current implementation is not yet user-friendly for teachers, school staff, or parents without an IT background. Running the notebook still requires basic familiarity with Python, Google Colab, and file management. Without a graphical interface or automated input/output workflow, non-technical users may struggle to upload handwriting samples, interpret outputs, or troubleshoot runtime errors. As a result, even though the model is executable and reproducible, it is not immediately accessible for practical screening use and would require additional design work (e.g., a web interface or mobile application) to support real-world adoption in primary education contexts.

4. CONCLUSION

This study proposed a DenseNet121-based approach for preliminary dyslexia screening through paragraph-level handwriting analysis, achieving 92.86% train accuracy and 90% test accuracy, despite the small dataset. The model demonstrated statistically significant superiority over traditional ML RF baseline (McNemar's test: $\chi^2 = 11.25$, $p<0.001$) and competitive performance relative to more complex state-of-the-art methods, while maintaining practical advantages such as low computational cost (208.57 ms/image), a single-architecture workflow, and ease of deployment (modest hardware requirements and easy writing sample collection), making it accessible for resource-constrained educational settings. Given that handwriting variability occurs across various specific learning disabilities and can be influenced by multiple non-diagnostic factors, this tool is designed as an early screening support system rather than a definitive diagnostic instrument. The perfect precision ensures that flagged cases warrant follow-up assessment, while the 80% recall indicates that negative screenings should not preclude further evaluation when other dyslexic indicators are present. However, the small dataset (200 images) and free-form writing format introduce limitations that future work must address. Despite these constraints, the proposed method demonstrates the feasibility of deep learning-based dyslexia screening as a practical, efficient, and accessible preliminary assessment tool for educational contexts.

Future work should pursue three integrated directions. First, standardized writing protocols would reduce confounding variability and enable clinical validation studies that directly compare model outputs with professional diagnostic assessments. Second, expanding the dataset with larger and more balanced samples would strengthen generalization and support the exploration of advanced architectures—such as attention mechanisms or transformer-based models—to capture and integrate more detailed sequence-level graphomotor patterns from word- and character-level analyses into the paragraph-level representation. Third, the model should be translated into accessible screening tools, such as teacher-friendly interfaces or mobile applications for at-home early screening, to facilitate practical adoption by educators and parents.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Chelsea Zaomi Pondayu	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Widodo		✓		✓	✓	✓				✓		✓		✓
Murien Nugraheni	✓			✓	✓					✓		✓		

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset that supports the findings of this study is openly available in the GitHub repository *“Dyslexia_Detection”* by user dlsathvik04 at https://github.com/dlsathvik04/Dyslexia_Detection.

REFERENCES

[1] G. R. Lyon, S. E. Shaywitz, and B. A. Shaywitz, “A definition of dyslexia,” *Annals of Dyslexia*, vol. 53, no. 1, pp. 1–14, Jan. 2003, doi: 10.1007/s11881-003-0001-9.

[2] J. Miciak and J. M. Fletcher, “The critical role of instructional response for identifying dyslexia and other learning disabilities,” *Journal of Learning Disabilities*, vol. 53, no. 5, pp. 343–353, Feb. 2020, doi: 10.1177/0022219420906801.

[3] L. S. Siegel, D. P. Hurford, J. L. Metsala, M. R. Ozier, and A. C. Fender, “Thoughts on the definition of dyslexia,” *Annals of Dyslexia*, Aug. 2025, doi: 10.1007/s11881-025-00337-y.

[4] V. W. Berninger, K. H. Nielsen, R. D. Abbott, E. Wijsman, and W. Raskind, “Writing problems in developmental dyslexia: under-recognized and under-treated,” *Journal of School Psychology*, vol. 46, no. 1, pp. 1–21, Feb. 2008, doi: 10.1016/j.jsp.2006.11.008.

[5] L. Yang *et al.*, “Prevalence of developmental dyslexia in primary school children: a systematic review and meta-analysis,” *Brain Sciences*, vol. 12, no. 2, p. 240, Feb. 2022, doi: 10.3390/brainsci12020240.

[6] A. B. Sunil, A. Banerjee, M. Divya, H. K. Rathod, J. Patel, and M. Gupta, “Dyslexia: an invisible disability or different ability,” *Industrial Psychiatry Journal*, vol. 32, no. Suppl 1, pp. S72–S75, Nov. 2023, doi: 10.4103/ipj.ipj_196_23.

[7] V. Sainz, J. J. Álvarez-Arjona, and J. L. Gómez-Gutiérrez, “Self-Concept and academic performance in students with and without learning difficulties: a longitudinal study in an inclusive school setting,” *SAGE Open*, vol. 15, no. 3, Jan. 2025, doi: 10.1177/21582440251356072.

[8] N. D. Alqahtani, B. Alzahrani, and M. S. Ramzan, “Deep learning applications for dyslexia prediction,” *Applied Sciences (Switzerland)*, vol. 13, no. 5, p. 2804, Feb. 2023, doi: 10.3390/app13052804.

[9] N. P. G. Seshadri, B. K. Singh, and R. B. Pachori, “EEG based functional brain network analysis and classification of dyslexic children during sustained attention task,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4672–4682, 2023, doi: 10.1109/TNSRE.2023.3335806.

[10] R. Svaricek, N. Dostalova, J. Sedmidubsky, and A. Cernek, “INSIGHT: combining fixation visualisations and residual neural networks for dyslexia classification from eye-tracking data,” *Dyslexia*, vol. 31, no. 1, Jan. 2025, doi: 10.1002/dys.1801.

[11] R. Vaitheeshwari *et al.*, "Dyslexia analysis and diagnosis based on eye movement," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 4109–4119, 2024, doi: 10.1109/TNSRE.2024.3496087.

[12] Y. Alkhurayyif and A. R. W. Sait, "Deep learning-driven dyslexia detection model using multi-modality data," *PeerJ Computer Science*, vol. 10, p. e2077, Jun. 2024, doi: 10.7717/PEERJ-CS.2077.

[13] Y. S. G. Kim, A. Wolters, and J. won Lee, "Reading and writing relations are not uniform: they differ by the linguistic grain size, developmental phase, and measurement," *Review of Educational Research*, vol. 94, no. 3, pp. 311–342, Jul. 2024, doi: 10.3102/00346543231178830.

[14] E. Van Heuverswyn, C. Gosse, and M. Van Reybroeck, "Handwriting difficulties in children with dyslexia: poorer legibility in dictation and alphabet tasks, slowness in the alphabet task," *Dyslexia*, vol. 30, no. 2, Apr. 2024, doi: 10.1002/dys.1767.

[15] L. Bazen, M. van den Boer, E. H. de Bree, and P. F. de Jong, "Presentation matters: surface text features and text quality in written narratives of Dutch high school students with and without dyslexia," *Dyslexia*, vol. 30, no. 4, Aug. 2024, doi: 10.1002/dys.1786.

[16] S. P. Patil, R. S. Apare, R. H. Borhade, and P. N. Mahalle, "Automated dyslexia screening using children's handwriting in English language with convolutional neural network and bidirectional long short-term memory model," *Engineered Science*, vol. 32, 2024, doi: 10.30919/es1345.

[17] M. Baggett, L. L. Diamond, and A. Olszewski, "Dysgraphia and dyslexia indicators: analyzing children's writing," *Intervention in School and Clinic*, vol. 59, no. 5, pp. 319–330, Aug. 2024, doi: 10.1177/10534512231189449.

[18] N. D. Alqahtani, B. Alzahrani, and M. S. Ramzan, "Detection of dyslexia through images of handwriting using hybrid AI approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, pp. 942–951, 2023, doi: 10.14569/IJACSA.2023.0141099.

[19] H. W. Liu, S. Wang, and S. X. Tong, "DysDiTect: dyslexia identification using CNN-positional-LSTM-attention modeling with Chinese dictation task," *Brain Sciences*, vol. 14, no. 5, p. 444, Apr. 2024, doi: 10.3390/brainsci14050444.

[20] T. Zaibi and H. Bezine, "Early detection of learning disabilities through handwriting analysis and machine learning," *Procedia Computer Science*, vol. 246, no. C, pp. 3702–3712, 2024, doi: 10.1016/j.procs.2024.09.186.

[21] A. R. W. Sait and Y. Alkhurayyif, "Lightweight hybrid transformers-based dyslexia detection using cross-modality data," *Scientific Reports*, vol. 15, no. 1, May 2025, doi: 10.1038/s41598-025-01235-4.

[22] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Q. Weinberger, "Convolutional networks with dense connectivity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8704–8716, Dec. 2022, doi: 10.1109/TPAMI.2019.2918284.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. The MIT Press, 2016.

[24] V. Ganchovska and I. Krasteva, "Converting color to grayscale image using LabVIEW," in *2022 International Conference Automatics and Informatics (ICAI)*, Oct. 2022, pp. 320–323, doi: 10.1109/icai55857.2022.9960062.

[25] N. Uzakkyzy *et al.*, "Image noise reduction by deep learning methods," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 6, pp. 6855–6861, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6855-6861.

[26] G. Qiu, "An improved recursive median filtering scheme for image processing," *IEEE Transactions on Image Processing*, vol. 5, no. 4, pp. 646–648, Apr. 1996, doi: 10.1109/83.491340.

[27] K. Maliński and K. Okarma, "Analysis of image preprocessing and binarization methods for OCR-based detection and classification of electronic integrated circuit labeling," *Electronics*, vol. 12, no. 11, p. 2449, May 2023, doi: 10.3390/electronics12112449.

[28] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of Graphics Tools*, vol. 12, no. 2, pp. 13–21, Jan. 2007, doi: 10.1080/2151237x.2007.10129236.

[29] N. Harun, I. S. Isa, S. A. Ramlan, M. K. Osman, and M. I. F. Maruzuki, "Dysgraphia handwriting image augmentation for CNN model classification," in *2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE)*, Aug. 2024, pp. 209–213, doi: 10.1109/iccsce61582.2024.10696383.

[30] S. Ethiraj and B. K. Bolla, "Augmentations: an insight into their effectiveness on convolution neural networks," in *Advances in Computing and Data Sciences*, Springer International Publishing, 2022, pp. 309–322.

[31] G. Ran, "Comparative analysis of machine learning and deep learning in practical scenarios," *Highlights in Science, Engineering and Technology*, vol. 94, pp. 531–539, Apr. 2024, doi: 10.54097/r03tx683.

[32] A. Zafar *et al.*, "A comparison of pooling methods for convolutional neural networks," *Applied Sciences*, vol. 12, no. 17, p. 8643, Aug. 2022, doi: 10.3390/app12178643.

[33] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, vol. 1, pp. 448–456.

[34] N. Alsadi, S. A. Gadsden, and J. Yawney, "Intelligent estimation: a review of theory, applications, and recent advances," *Digital Signal Processing: A Review Journal*, vol. 135, p. 103966, Apr. 2023, doi: 10.1016/j.dsp.2023.103966.

[35] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense convolutional network and its application in medical image analysis," *BioMed Research International*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/2384830.

[36] N. Le *et al.*, "K-fold cross-validation: an effective hyperparameter tuning technique in machine learning on GNSS time series for movement forecast," Springer Nature Switzerland, 2024, pp. 377–382.

[37] S. Sathyanarayanan, "Confusion matrix-based performance evaluation metrics," *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

[38] Ş. K. Çorbacıoğlu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: a guide to interpreting the area under the curve value," *Turkish Journal of Emergency Medicine*, vol. 23, no. 4, pp. 195–198, Oct. 2023, doi: 10.4103/tjem.tjem_182_23.

[39] J. Gawlikowski *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. S1, pp. 1513–1589, Jul. 2023, doi: 10.1007/s10462-023-10562-9.

[40] B. H. Willis, D. Coomar, and M. Baragilly, "Tailored meta-analysis: an investigation of the correlation between the test positive rate and prevalence," *Journal of Clinical Epidemiology*, vol. 106, pp. 1–9, Feb. 2019, doi: 10.1016/j.jclinepi.2018.09.013.

[41] G. F. Reed, F. Lynn, and B. D. Meade, "Use of coefficient of variation in assessing variability of quantitative assays," *Clinical and Vaccine Immunology*, vol. 10, no. 6, pp. 1162–1162, Nov. 2003, doi: 10.1128/cdli.10.6.1162.2003.

[42] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional," *BMC Medical Research Methodology*, vol. 13, no. 1, Jul. 2013, doi: 10.1186/1471-2288-13-91.

[43] A. Jeamaon and C. Khemapatapan, "Development cyber risk assessment for intrusion detection using enhanced random forest," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 18, no. 4, pp. 429–442, Sep. 2024, doi: 10.37936/ecti-cit.2024184.256185.

BIOGRAPHIES OF AUTHORS

Chelsea Zaomi Pondayu is a final-year student in informatics education at Universitas Negeri Jakarta (2025), with a focus on AI, web development and educational technology. Her passion for AI innovation led her to become a finalist in the 2023 outstanding student competition at FT-UNJ, proposing an AI-based cancer diagnosis tool. She has completed internships at Indonesia's Ministry of Communication and Informatics (now Komdigi) and teaching practicums at SMK Negeri 17 Jakarta, where she taught web development and database systems to vocational students. Her research interests include: ML and DL. She can be contacted at email: chelseazaomi9@gmail.com or chelseazaomipondayu_1512621089@mhs.unj.ac.id.

Widodo is an assistant professor at Department of Informatics Education of Universitas Negeri Jakarta. Widodo earned a Ph.D. in computer science from University of Indonesia in January 2020, M.Sc. in computer science from University of Indonesia in 2004, and B.Sc. in information systems from Gunadarma University in 1999. His research interests include: privacy preserving data publishing, natural language processing, and classification and clustering. He can be contacted at email: widodo@unj.ac.id.

Murien Nugraheni specializes in AI, data mining, and expert systems, with research interests spanning decision support systems (DSS) and intelligent data analysis. She holds a bachelor of engineering (S.T.) from Universitas Ahmad Dahlan and a master of computer science (M.Kom.) from Universitas Gadjah Mada, equipping her with robust technical expertise in AI-driven solutions. Her work focuses on leveraging data mining techniques and knowledge-based systems to develop scalable decision-making tools for real-world applications. She can be contacted at email: muriennugraheni@unj.ac.id.