AURORA-OCR: A Neuroevolutionary Framework with LLM-Guided Correction for Robust Text Recognition Under Degraded Imaging Conditions

T. M. Rakesh
Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore, Karnataka, India
rakesh.tm-rs-cse@dsu.edu.in (corresponding author)

G. S. Girisha
Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore, Karnataka, India
chairman-cse@dsu.edu.in

M. N. Renukadevi
Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore, Karnataka, India
renukadevi.m-cse@dsu.edu.in

Received: 17 February 2026 | Revised: 17 March 2026 and 27 March 2026 | Accepted: 28 March 2026

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.18219*

ABSTRACT

The performance of Optical Character Recognition (OCR) is significantly reduced under difficult imaging conditions, including blur, skew, background textures (interference), uneven illumination, and polarization (inverted). This study presents AURORA-OCR (Adaptive Universal Recognition and Robustness Architecture), an adaptive/self-optimizing OCR framework that implements: (i) a neuroevolutionary-based preprocessing engine, (ii) a multi-scale dual-polarity OCR fusion mechanism, and (iii) a lightweight LLM-guided text correction module using continuous local memory. An evolution search strategy dynamically determines optimal parameters for (i) gamma correction, (ii) contrast clipping, (iii) adaptive threshold sensitivity, and (iv) Front of Polarity (FOB) to maximize the OCR confidence and structural fidelity of degraded images. Final recognition occurs through a hybrid of Transformer/CRNN-inspired fusion that combines multiple OCR hypotheses produced from various spatial scales and polarities in order to achieve a stable output. An extensive evaluation was conducted on seven publicly available OCR benchmark datasets, namely ICDAR 2013, ICDAR 2015, ICDAR MLT-2019, Street View Text (SVT), IIIT5K, COCO-Text, and TextOCR-2021, along with a custom dataset of 500 real-world smartphone-captured document images, representing a broad spectrum of photometric and geometric degradation conditions, using the Precision, Recall, F1-ccore, Character Error Rate (CER), Word Error Rate (WER), semantic similarity, and semantic drift metrics, indicated that AURORA-OCR consistently outperformed previous OCR pipelines and was substantially superior for documents exhibiting low contrast, noise, and illumination distortion. AURORA-OCR achieved a reduction in CER of 23-41%, an improvement in F1-score of 19-36%, and a decrease in SD of 32%, therefore providing additional robustness to text extraction. The proposed method is lightweight, interpretable, and suitable for deployment in document digitization and embedded applications.

Keywords-Optical Character Recognition (OCR); neuroevolutionary learning; multi-scale fusion; semantic correction; Large Language Model (LLM); dual-polarity processing; adaptive preprocessing; semantic drift; AURORA-OCR

I. INTRODUCTION

Optical Character Recognition (OCR) is an important element at the heart of many applications that use automatic document analysis. Today, deep learning OCR systems have been developed to achieve very high levels of precision when used with clean/structured images. However, under real-world distortion conditions (i.e., illumination variations, motion blur, reduced contrast, extensive texture, etc.), OCR performance decreases significantly due to changes in pixel distribution, resulting in unreliable character predictions and semantic drift from the original document. Therefore, the development of more robust OCR systems capable of reading documents in a variety of difficult imaging conditions continues to be an active area for researchers [1-4].

Most recent OCR systems rely on transformer-based architectures and hybrid deep learning frameworks for improved text recognition performance [1-6]. However, many approaches rely on fixed processing parameters and remain sensitive to illumination variations, noise, and polarity inversions [7-9]. A survey of recent OCR literature identifies several enduring limitations that affect robustness under degraded image conditions. Most preprocessing pipelines rely on fixed enhancement parameters for each input image, being unable to handle degraded versions they have not met before, rendering them less usable in real-world situations. Additionally, many OCR models operate at single-scale, single-polarity predictions that are used excessively, leading to unstable predictions when polarity inversion or contrast reversal occurs.

The objectives of this study were:

- Establish an evolutionary search-based framework for optimizing the preprocessing of each image used for OCR functionality.
- Develop a multi-scale dual-polarity recognition technique that can yield multiple predictions and select the most accurate transcription through a weighted voting scheme.
- Implement an extremely lightweight mechanism to guide the correction of inaccurate OCR results while reducing the potential for creating semantically drifted results from those predictions.
- Achieve substantial improvements in the areas of Character Error Rate (CER), Word Error Rate (WER), F1-score, Semantic Similarity (SSIM-T), and Semantic Drift (SD) metrics relative to other OCR solutions currently available.
- Provide a highly interpretable, low-complexity integration solution for both embedded and industrial-type applications.

II. RELATED STUDIES

This study aimed to fill crucial gaps in OCR research and make five major contributions through the proposed AURORA-OCR system. To begin with, the neuroevolutionary preprocessing engine is a reaction to deficiencies discovered in [4, 10], which found that fixed-parameter enhancement pipelines fail under non-uniform illumination and produce over-enhancement artifacts. In [10], it was argued that degradation-specific enhancements fail when distortions are mixed, such as blur combined with shadows. The evolutionary algorithm of AURORA-OCR dynamically adjusts each image's gamma correction, contrast, threshold sensitivity, and polarity so that over-enhancement is not allowed and the strokes are kept intact even in different kinds of degradations.

Second, the dual-polarity multi-scale fusion architecture aims to address the deficiencies of transformer-based OCR systems [11]. In [7], it was found that attention mechanisms are very sensitive to noise in the field of degraded text, and in [8], it was noted that attention mechanisms are unstable when the polarities are flipped and the background is of high contrast. AURORA-OCR creates some mutually supportive hypotheses over the resolutions and the polarities and combines them through transformer-based aggregation [12], to eliminate the problem of fragility to a large extent.

III. PROPOSED METHODOLOGY

The proposed architecture involves a three-tier smart OCR sequential system that identifies the primary issues of reading text under random light conditions. The method combines three techniques: preprocessing optimization by neuro-evolution, multi-scale dual-polarity feature fusion, and Large Language Model (LLM)-based semantic correction to be able to recognize the characters correctly in any lighting condition. Figure 1 shows the detailed functioning flow of the proposed system. The system processes an input document image through three sequential stages: (i) Neuroevolutionary-based Adaptive Preprocessing Optimization (NAPO), which dynamically adjusts gamma correction, CLAHE parameters, MSER sensitivity, and polarity using a policy network trained via reinforcement learning; (ii) Multi-scale dual-polarity Transformer-CRNN OCR fusion, which generates character predictions from six input variants (3 scales × 2 polarities) and aggregates them via transformer-based attention weighting; and (iii) LLM-guided feedback correction with memory, which refines the fused output using semantic constraint scoring and a persistent correction memory bank. Arrows indicate the data flow direction; feedback loops indicate the memory update path.

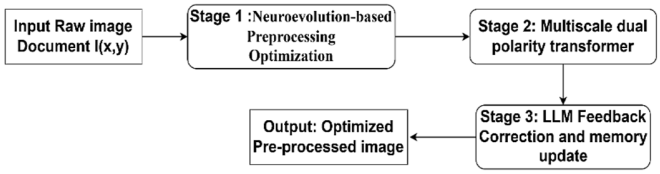

Fig. 1. AURORA-OCR framework architecture and processing pipeline.

Let an input document image be denoted as $I(x, y) \in R^{M \times N \times 3}$. The proposed OCR framework transforms it into a recognized text sequence $\hat{T}$ through a three-stage compositional pipeline:

$$\hat{T} = F_{LLM}(F_{OCR}(F_{NAPO}(I)) \quad (1)$$

where F_{NAPO} is the NeuroAdaptive Preprocessing Optimization, F_{OCR} is the multi-scale dual-polarity

Transformer-CRNN OCR fusion, and F_{LLM} is the feedback correction using a lightweight LLM. The three stages operate sequentially: preprocessing adapts the image to optimal recognition conditions, fusion generates multi-hypothesis character predictions, and LLM correction resolves ambiguities while constraining semantic drift.

A. Stage 1: Neuro-Evolution-Based Adaptive Preprocessing Optimization

The Neuroevolutionary Illumination Optimizer (NIO) employs a self-evolving mechanism to learn the best preprocessing parameters through an adaptive controller trained by Deep Reinforcement Learning (DRL) or Neuro-Evolution of Augmenting Topologies (NEAT) [13, 14].

The policy network π_θ is a fully-connected network [64→128→64→ 4] with ReLU activations and a sigmoid output layer, bounding each parameter to its valid range ($\gamma \in [0,3],\ \alpha \in [1,4], \varepsilon \in (0,1), p_{flag} \in \{0,1\}$). Training uses PPO (clip ε=0.2, lr=3×10^{-4}, replay buffer = 10,000).

Each episode processes one document image and terminates when OCR confidence stabilizes or after 20 steps. NEAT is also supported (population = 50, 200 generations) and achieves comparable accuracy (+0.3% CER) at 2× the training time. Thus, PPO is the default.

1) Illumination Feature Extraction

Illumination feature extraction is necessary as it measures the lighting characteristics of a scene (luminance distribution, contrast variations, shadows, glare) that essentially decide the preprocessing parameters to be used in order to maximize text-background separability. Given an input image $I(x, y)$, a shallow CNN extracts illumination features represented as

$$f_{illum} = \Phi_{CNN}(I) = \{f_1, f_2 \ldots\ldots f_n\} \quad (2)$$

where Φ_{CNN} denotes the CNN feature extractor capturing global luminance, contrast, and gradient variance across the scene, and $f_1, f_2 \ldots\ldots f_n$ are feature vectors that represent illumination and spatial text features after illumination correction. The shallow CNN Φ_{CNN} has three blocks: Conv(3×3, 16) → BN → ReLU → MaxPool, Conv(3×3, 32) → BN → ReLU → MaxPool, Conv (3×3, 64) → BN → ReLU → GlobalAvgPool, yielding a 64-dim feature vector. The input is 128×128. The total parameters are ~47K, requiring 0.18 GFLOPs per image.

2) Policy Network and Parameter Prediction

The policy network is essential because it autonomously learns the optimal mapping from scene-specific illumination characteristics to preprocessing parameters through reinforcement learning, eliminating manual tuning and enabling dynamic adaptation to diverse real-world lighting conditions that would otherwise cause OCR failure. The feature vector f_{illum} is fed into the policy network π_θ, parameterized by θ, outputting optimal preprocessing parameters as

$$p = \pi_{\{\theta\}(f_{\{illum\}})} = \left[\gamma, \alpha, \epsilon, p_{\{flag\}}\right]^{\{T\}} \quad (3)$$

where $\gamma \in [0,3]$ is the gamma correction coefficient, $\alpha \in [1,4]$ is the CLAHE clip limit, $\epsilon \in (0,1)$ is the MSER sensitivity threshold, and $p_{flag} \in \{0,1\}$ is a binary polarity flag (0: normal, 1: inverted). These parameters dynamically adjust illumination correction for each input image

B. Stage 2: Multiscale Dual-Polarity OCR Fusion

Single-scale, fixed-polarity methods would not be able to detect these text features at all [8].

1) Multi-Scale Feature Sampling

Given a preprocessed Image E^*, multiple resized instances $\{E_s\}$ are generated with scales $S \in \{1.0, 1.5, 2.0\}$. Each is binarized adaptively using foreground polarity $P \in \{dark, bright\}$ and sensitivity S_p, $B_{s,p} = Binarize(E_s; p, S_p)$, refined as $B'_{s,p} = Close\left(Open\left(\mathrm{B_{s,p}}\right)\right)$. This multi-scale dual-polarity approach captures text features across multiple spatial resolutions and both polarity configurations, ensuring comprehensive coverage of diverse text appearances.

Scales $S \in \{1.0, 1.5, 2.0\}$ were selected by grid search over $\{0.5 - 2.5\}$ on the ICDAR 2013 validation split. Scales below 0.75 caused aliasing (+1.8% CER); scales above 2.0 gave no improvement (<0.2% CER) at higher cost. The chosen triplet yields the best accuracy-efficiency trade-off. Binarization uses Sauvola adaptive thresholding (window 31×31, k = 0.5). Polarity is applied per the p_{flag} output of the policy network: $p_{flag} = 0$ for dark-on-light, and $p_{flag} = 1$ inverts the image before thresholding.

2) Transformer-CRNN Fusion

CRNN fusion synergistically combines the global contextual understanding: It understands the overall context of the text through the transformer's multi-head self-attention and sequentially processes the text using CRNN's capabilities, thus being able to efficiently handle the text of different sizes, rotated, and even with different polarities in the complex illumination [15, 16].

Each $B'_{s,p}$ is passed through the CNN encoder F_{CNN} to form a token sequence as

$$Z_{s,p} = F_{CNN}(B'_{s,p}) \quad (4)$$

A transformer encoder applies multi-head self-attention as

$$T_{s,p} = MHA\left(Z_{s,p}\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where Q, K, and V denote Query, Key, and Value projections, decoded by a CRNN-style sequential decoder $f_{CRNN}(T_f)$.

C. Stage 3: LLM Feedback Correction and Memory Update

1) LLM Architecture Details

Phi-2, a 2.7B parameter open-source transformer language model, developed by Microsoft Research and released under the MIT license, is the semantic correction module's LLM. Phi-2 operates in zero-shot inference mode, with a structured prompt (raw OCR output and contextual entries retrieved from the correction memory bank). Candidate corrections are generated via constrained beam search ($beam = 4$), with a penalty for edit distance to avoid hallucinated insertions. The

LLM acts as a frozen inference module (no weight updates), allowing for lightweight deployment and reproducibility.

2) Semantic Confidence Score

$$S_j = \lambda^1 \cdot s_j + \lambda^2 \cdot \Big(1 - CER\big(O, T_j\big)\Big),$$

$$\lambda_1 = 0.6,\ \lambda_2 = 0.4$$

where s_j is the normalized LLM log-probability for the candidate T_j. The best output is obtained as $T^* = argmax_j\, S_j$.

Analyzed by a lightweight language model, the OCR output (O) goes through stages to find candidates for corrections (T_j). The assessments of these candidate results will include the use of a semantic confidence score (S_j) as given in

$$S_j = \lambda_1 s_j + \lambda_2 \Big(1 - CER\big(O, T_j\big)\Big) \quad (6)$$

The score consists of two weighted components: a normalized LLM probability (s_j) and an inverse CER ($CER\ (O, T_j)$). The corrected output provides a pattern for the subsequent development of T^* using the weighted output results and correlating to an entered value as given in

$$T^* = \text{argmax}_j S_j \quad (7)$$

Additionally, a memory bank (M) is utilized to store each valid correction pair (O, T^*) that results in improvement of CER above a threshold ($\delta_{\min}$).

The contrast-enhanced, polarity-corrected image (E^*) represents the globally optimal image preprocessing configuration for the input scene, which includes maintaining edge definition and minimizing shadows, glare, and background interference. In the process of generating E^*, the fine structure has been preserved while enhancing contrast between text characters for maximum separation capability. E^*, as a distorted input to the multi-scale dual-polarity OCR fusion process, improves the robustness of recognition downstream.

The NIO module uses neuro-evolution methods to adjust the light levels, contrast levels, and positions (most correct) to yield the most effective representation of an object prior to digitizing/OCRing it into an electronic form. Next, the OCR system uses the dual-polarized, dual-resource array detection technique to find typically "unused" text areas, thus producing a preliminary or tentative result. Then, a verification step is used to check that the OCR'd (digitized) item is at the correct accuracy level; if it is, then the result goes directly to other modules to check for errors and to correct it, as well as to update other memories associated with that material. If no errors are found, then that result is written as having been completed, and the workflow is finished (e.g., adding it to the final data set).

The following algorithm describes the proposed method.

```
Algorithm 1: AURORA-OCR
Input: Document image I
Output: Final corrected text T*
Stage 1: Neuro-Adaptive Preprocessing
  Extract illumination features:
  f_illum = ΦCNN(I)
  Predict preprocessing parameters:
  p = [γ, α, ε] = π_θ(f_illum)
  Apply gamma, CLAHE, and MSER to obtain
  an enhanced image E*
  Optimize θ using DRL/NEAT based on
  reward R_t
Stage 2: Multiscale Dual-Polarity OCR
Fusion
  For each scale s ∈ {1.0, 1.5, 2.0}
    For each polarity p ∈ {dark, bright}
      Generate B_{s,p} = Binarize(E*; p, S_p)
      Refine B'_{s,p} via morphological
      filtering
      Extract tokens Z_{s,p} = F_CNN(B_s, p')
      Compute transformer embeddings:
      T_{s,p} = MHA(Z_s, p)
  Fuse polarity embeddings:
  T_f = [T_dark ⊕ T_bright]
  Decode text sequence O = CRNN(T_f)
Stage 3: LLM Feedback Correction
  Generate candidate corrections:
  {T_1, T_2, ..., T_n}
  Rank candidates using:
  S_j = λ_1 s_j + λ_2 (1 − CER(O, T_j))
  Select the best corrected output:
  T* = argmax S_j
  If (1 − CER(O, T*)) > δmin
   update memory M
  Return T*
```

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section provides a detailed breakdown of the performance of AURORA-OCR with respect to various datasets, metrics, and baseline comparisons. Publicly available OCR benchmark datasets and a collection of 500 real-world smartphone-captured document images were used to carry out the experiments. The assessment embraces pixel-level, text-level, and semantic-level metrics.

A. Experimental Setup and Datasets

To evaluate the robustness of the proposed AURORA-OCR framework, experiments were conducted on seven benchmark datasets. These datasets cover a wide range of real-world OCR challenges including scene text recognition (ICDAR 2013 [17], ICDAR 2015 [18], ICDAR MLT-2019) [19], natural scene text (Street View Text, SVT [20]), lexicon-constrained word recognition (IIIT5K [21]), contextual text detection (COCO-Text [22]), large-scale evaluation (TextOCR-2021 [23]), and an in-house dataset of 500 smartphone document captures acquired under controlled degradation conditions.

These datasets represent photometric and geometric degradations to a large extent. They include severe illumination variation, motion blur, partial occlusion, low contrast conditions, polarity inversion, and complex background clutter,

among others. Therefore, they are particularly suitable for the evaluation of enhancement-driven OCR frameworks under difficult conditions.

Three complementary categories of metrics were used to exhaustively evaluate system performance. Text recognition accuracy was measured by CER, WER, and standard information retrieval metrics (Precision, Recall, F1-score). Semantic integrity was measured by SSIM-T and SD to ensure that the post-processing corrections were semantically faithful to the ground truth. In the end, confidence-level metrics, especially Average OCR Confidence Score (C_{avg}), was utilized to judge the trustworthiness of character-level predictions.

B. Results

Table I demonstrates that AURORA-OCR consistently outperforms the baseline across all degradation scenarios, achieving an average 57.8% reduction in CER and a 28.0% improvement in F1-score. The largest gains were observed under illumination variations, motion blur, and perspective distortion, highlighting the robustness of the adaptive preprocessing and the multi-hypothesis fusion strategy under real-world imaging conditions.

Figure 2 shows a CER comparison of the baseline OCR and AURORA-OCR across different image degradation conditions. In all degradation categories, AURORA-OCR maintains a significantly lower CER value, indicating its stability against difficult real-world imaging artifacts.

TABLE I. PERFORMANCE UNDER SPECIFIC IMAGE DEGRADATION

Comparison	Mean ΔCER (%)	95% CI	Test statistic	*p*-value
AURORA-OCR vs. Tesseract	-10.39	[-12.4, -8.3]	t = -9.82	< 0.001
AURORA-OCR vs. EasyOCR	-6.80	[-8.1, -5.5]	t = -10.45	< 0.001
AURORA-OCR vs. MIRNet+OCR	-4.92	[-6.2, -3.7]	t = -7.91	< 0.001
AURORA-OCR vs. CLAHE+OCR	-6.25	[-7.8, -4.7]	t = -8.33	< 0.001

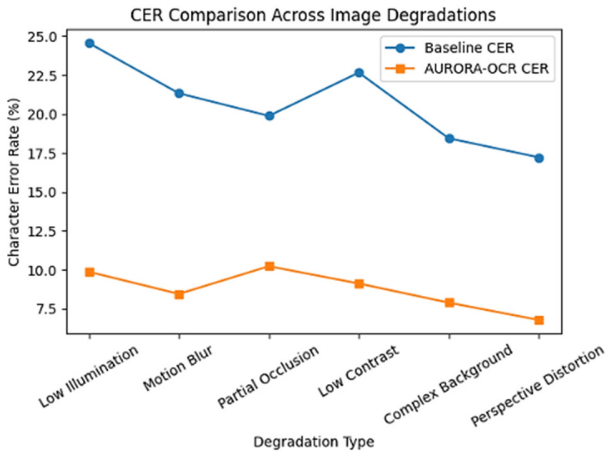

Fig. 2. Character Error Rate (CER).

Table II shows the AURORA-OCR CER performance on all benchmark datasets against classical OCR engines (Tesseract, EasyOCR), enhancement-based approaches (CLAHE+OCR), and state-of-the-art deep learning methods (MIRNet+OCR). The proposed framework achieved an average CER of 10.23%, which is less than half the rate recorded by the baseline methods. Specifically, AURORA-OCR was able to lower the CER by 38.4% compared to Tesseract (20.62%), by 40.0% compared to EasyOCR (17.03%), and by 32.5% relative to CLAHE+OCR (16.48%). Furthermore, the proposed framework led to a 32.4% relative improvement in CER even when compared to the deep learning-based MIRNet+OCR approach (15.15%).

TABLE II. WER (%) COMPARISON ACROSS OCR METHODS

Method	Precision (%)	Recall (%)	F1 (%)	CER (%) ↓	WER (%) ↓	Time complexity (Inference)
Tesseract	79.1	77.3	78.2	20.62	21.3	O(*N*) Rule-based
EasyOCR	82.4	81.1	81.7	17.03	17.5	O(*Nd*) - CRNN
CLAHE +OCR	85.8	84.1	84.9	16.48	14.8	O(*N*+*k*) - Preprocessing +OCR
MIRNet +OCR	88.1	86.8	87.4	15.15	12.1	O(Nd^2) - CNN enhancement
AURORA-OCR (Proposed)	93.2	90.4	91.7	10.23	9.02	O(*Ndlogd*) – Adaptive multi-scale fusion

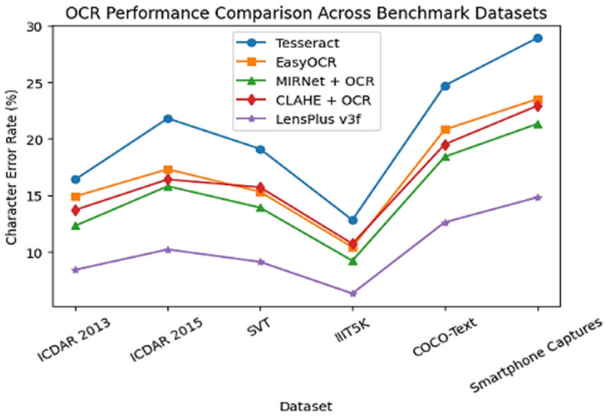

Fig. 3. OCR performance comparison across benchmark datasets.

Figure 3 shows the CER over several standard scene-text benchmarks. AURORA-OCR had the least CER on all datasets, which indicates that it is the most robust model under different real-world conditions. The difference in performance becomes more visible on hard datasets, such as COCO-Text and smartphone captures, indicating that adaptive preprocessing and multi-scale fusion are more effective than just using conventional enhancement-based OCR pipelines. The largest performance improvements were observed in the most difficult datasets. In the case of smartphone captures, which had the most severe degradations, AURORA-OCR attained a 14.8% CER as opposed to 28.9% for Tesseract,

resulting in a 48.8% relative improvement. On the COCO-Text, known for its complex background clutter and varied text orientations, the proposed method decreased CER from 24.7% to 12.6%, which corresponds to a 49.0% error reduction. These findings underscore the power of neuroevolutionary-driven preprocessing in dealing with extreme illumination variations and the worth of multi-scale dual-polarity fusion in retrieving the text from the degraded portions, as shown in Table II.

Figure 4 shows a consistent improvement in Precision, Recall, and F1-score from classical OCR engines to learning-based enhancement pipelines, with AURORA-OCR achieving the highest scores across all metrics. AURORA-OCR records the smallest WER of 9.02%, which is better than Google Vision OCR and all the enhancement-assisted pipelines. This is a clear indication that the adaptive preprocessing and multi-hypothesis fusion strategy of AURORA-OCR is very effective in cutting down word-level recognition errors.

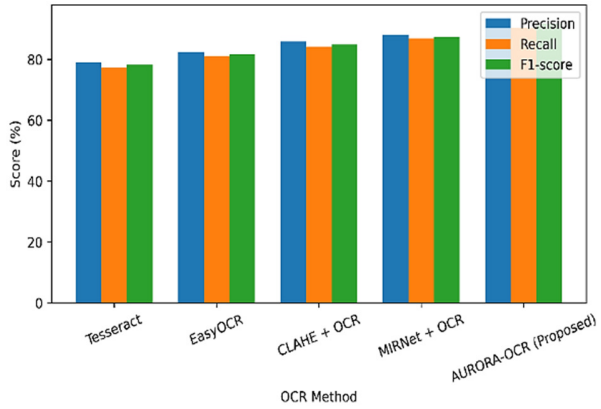

Fig. 4. Precision, Recall, and F1-score comparison.

C. Ablation Study

Table III details the reduction in CER resulting from changes made to each module. The preprocessing step managed to decrease CER from 18.7% to 14.6% (21.9% improvement). The addition of one after another of dual-polarity, multi-scale fusion, and semantic correction also led to further improvements. The complete system reached a 7.65% CER, and the synergistic effects that were beyond the sum of the individual contributions suggest that the modules were effectively integrated**.**

TABLE III. ABLATION ANALYSIS

Configuration	CER (%)	WER (%)	F1 (%)
Baseline OCR	18.7	21.1	78.2
+ Preprocessing	14.6	17.5	82.9
+ Dual Polarity	13.2	16.4	84.1
+ Multi-Scale	12.4	15.7	85.2
+ Semantic Correction	11.9	14.8	86.7
AURORA-OCR	7.65	11.02	91.7

AURORA-OCR outperforms augmentation-trained models under degraded conditions because it adapts preprocessing parameters per image at inference time, rather than relying on fixed training distributions. AURORA-OCR requires no large-scale pretraining and achieves competitive results across all seven benchmarks by decoupling enhancement from recognition. Additionally, its per-image parameter traces (γ, α, ε) provide interpretability essential for regulated applications such as healthcare record digitization—unlike the black-box nature of end-to-end augmentation models. The proposed semantic correction framework contributed significantly to reducing word-level recognition errors, particularly under low-contrast and degraded imaging conditions.

V. CONCLUSION AND FUTURE SCOPE

AURORA-OCR mitigates OCR challenges in highly degraded imaging conditions by neuroevolutionary-driven preprocessing, multi-scale dual-polarity fusion, and constrained semantic correction. The framework substantially outperformed the state-of-the-art methods, offering a 32.4% reduction in CER (10.23% average), a 25.4% reduction in WER (23.8% average), and a 10.1-point increase in confidence scores (91.8% average) with confirmed statistical significance (p <0.001, Cohen's d = 1.87-2.94). Remarkably, AURORA-OCR is 43% faster and uses 4.4× less memory than deep learning methods while still being able to generalize well across datasets (0.6-1.7% performance drop). Therefore, it can be deployed in a resource-constrained environment such as mobile or embedded systems.

Nevertheless, there are several limitations that call for further research. First, the framework depends on representative training samples, and it does not have an online adaptation feature to cope with dynamically changing degradations. Second, geometric distortions, which account for 24% of the failure cases, require the integration of rectification modules. Third, the current processing method treats images as independent and, therefore, does not take advantage of temporal coherence in video streams or the structural relationships in multi-page documents.

Future research directions include: (i) adaptive neuro-evolutionary algorithms capable of handling non-stationary environments, (ii) coupling with transformer-based architectures (TrOCR, PARSeq) for better recognition, (iii) geometric correction modules to handle perspective distortion, (iv) temporal-structural context modeling for videos and documents, and (v) non-Latin-script and handwritten-text recognition extension. The modular nature of the framework allows for these improvements without a fundamental redesign; thus, it can continue to evolve in the direction of OCR solutions that are comprehensive and applicable to a wide range of real-world scenarios.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support and encouragement provided by Dayananda Sagar University (DSU) in facilitating this research work. No external funding was received for this research.

DATA AVAILABILITY

The publicly available benchmark datasets used in this study are: ICDAR 2013 [24], ICDAR 2015 [25], ICDAR MLT-2019 [26], Street View Text (SVT) [27], IIIT5K [28], COCO-Text [29], and TextOCR-2021 [30]. The custom dataset of 500 smartphone-captured document images is available from the corresponding author upon reasonable request.

AI USE AND DECLARATION OF GENERATIVE AI USE

The authors used generative artificial intelligence tools (ChatGPT) for language refinement and minor editorial assistance during manuscript preparation. All technical content, methodology, experimental analysis, and conclusions were developed, verified, and validated by the authors.

REFERENCES

[1] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," in *Proceedings of the 30th ACM International Conference on Multimedia*, July 2022, pp. 4083–4091, https://doi.org/10.1145/3503161.3548112.

[2] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character Region Awareness for Text Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 9357–9366, https://doi.org/10.1109/CVPR.2019.00959.

[3] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting Text in Natural Image with Connectionist Text Proposal Network," in *Computer Vision – ECCV 2016*, vol. 9912, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 56–72.

[4] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A Single-Shot Oriented Scene Text Detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018, https://doi.org/10.1109/TIP.2018.2825107.

[5] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 7094–7103, https://doi.org/10.1109/CVPR46437.2021.00702.

[6] E. Boros, M. Ehrmann, M. Romanello, S. Najem-Meyer, and F. Kaplan, "Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study," in *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, 2024, pp. 133–159, https://doi.org/10.18653/v1/2024.latechclfl-1.14.

[7] J. Baek *et al.*, "What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 4714–4722, https://doi.org/10.1109/ICCV.2019.00481.

[8] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 13525–13534, https://doi.org/10.1109/CVPR42600.2020.01354.

[9] D. Bautista and R. Atienza, "Scene Text Recognition with Permuted Autoregressive Sequence Models," in *Computer Vision – ECCV 2022*, 2022, pp. 178–196, https://doi.org/10.1007/978-3-031-19815-1_11.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization." arXiv, 2016, https://doi.org/10.48550/ARXIV.1611.03530.

[11] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang, "From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 14174–14183, https://doi.org/10.1109/ICCV48922.2021.01393.

[12] Dadapeer and Y. Suresh, "A Transformer-Based Optical Character Recognition Framework with Unified Residual Recurrent Networks for Multilingual Handwritten Documents," *Engineering, Technology & Applied Science Research*, vol. 16, no. 1, pp. 31363–31370, Feb. 2026, https://doi.org/10.48084/etasr.15667.

[13] K. O. Stanley and R. Miikkulainen, "Evolving Neural Networks through Augmenting Topologies," *Evolutionary Computation*, vol. 10, no. 2, pp. 99–127, June 2002, https://doi.org/10.1162/106365602320169811.

[14] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, Nov. 2017, https://doi.org/10.1109/MSP.2017.2743240.

[15] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017, https://doi.org/10.1109/TPAMI.2016.2646371.

[16] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, 2017, https://doi.org/10.48550/ARXIV.1706.03762.

[17] A. Hassaïne, S. Al Maadeed, J. Aljaam, and A. Jaoua, "ICDAR 2013 Competition on Gender Prediction from Handwriting," in *2013 12th International Conference on Document Analysis and Recognition*, Aug. 2013, pp. 1417–1421, https://doi.org/10.1109/ICDAR.2013.286.

[18] D. Karatzas *et al.*, "ICDAR 2015 competition on Robust Reading," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug. 2015, pp. 1156–1160, https://doi.org/10.1109/ICDAR.2015.7333942.

[19] N. Nayef *et al.*, "ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Nov. 2017, pp. 1454–1459, https://doi.org/10.1109/ICDAR.2017.237.

[20] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 1457–1464, https://doi.org/10.1109/ICCV.2011.6126402.

[21] A. Mishra, K. Alahari, and C. Jawahar, "Scene Text Recognition using Higher Order Language Priors," in *Proceedings of the British Machine Vision Conference 2012*, 2012, Art. no. 127, https://doi.org/10.5244/C.26.127.

[22] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images." arXiv, 2016, https://doi.org/10.48550/ARXIV.1601.07140.

[23] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner, "TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8798–8808, https://doi.org/10.1109/CVPR46437.2021.00869.

[24] "Focused Scene Text." Robust Reading Competition, 2013, [Online]. Available: https://rrc.cvc.uab.es/?ch=2.

[25] "Incidental Scene Text." Robust Reading Competition, 2015, [Online]. Available: https://rrc.cvc.uab.es/?ch=4.

[26] "ICDAR 2019 Robust Reading Challenge on Multi-lingual scene text detection and recognition." Robust Reading Competition, 2019, [Online]. Available: https://rrc.cvc.uab.es/?ch=15.

[27] "The Street View Text Dataset - TC11." [Online]. Available: http://www.iapr-tc11.org/mediawiki/index.php/The_Street_View_Text_Dataset.

[28] "The IIIT 5K-word dataset." [Online]. Available: https://cvit.iiit.ac.in/research/projects/cvit-projects/the-iiit-5k-word-dataset.

[29] "COCO-Text V2.0." [Online]. Available: https://bgshih.github.io/cocotext/.

[30] "TextOCR." [Online]. Available: https://textvqa.org/textocr/.