Journal of Human Centered Technology

https://www.humentech.utm.my E-ISSN: 2821-3467

# Improving Right to Left Cursive Handwritten Text Recognition in Historical Manuscripts Using Learnable Edge Features and Channel Attention

Bilal Abdulrahman[1,2], Farhan Mohamed[1,*]

[1]Department of Emergent Computing, Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Bahru, 81310, Malaysia
[2]Department of Computer Science, College of Applied Science, University of Samarra, Samarra, 34010, Iraq

*Corresponding Author farhan@utm.my

**Cite:** https://doi.org/10.11113/humentech.v5n2.140 Research Article

**Abstract:**
Recent Handwritten Text Recognition (HTR) systems for Arabic and other right-to-left historical manuscripts have advanced exploration through convolutional neural network (CNN), recurrent, and Transformer-based models. However, degradation, weak diacritics, unstable baselines, and visually similar cursive letterforms still limit the recognition robustness. This paper presented an edge-aware line-level HTR framework that extends a CNN-Transformer baseline with a learnable edge-extraction channel and Squeeze-and-Excitation (SE) channel attention. The edge module emphasized stroke boundaries and character contours, while SE attention recalibrated feature responses to suppress background artifacts and preserve informative ink patterns. The resulting sequence was modeled by a Transformer encoder and trained using Connectionist Temporal Classification (CTC) with an auxiliary decoder cross-entropy loss. The experiments on the Kalima Arabic manuscript line-image dataset, using Books 1-8 with 86 pages and 1,759 annotated text lines, reduced character error rate from 6.40% to 4.10% and word error rate from 27.43% to 20.43%. These results show that combining learnable structural cues with channel-wise attention has improved robustness for degradation-prone historical manuscript collections.

**Keywords:** R-L handwritten text recognition; Historical manuscripts; CNN; Learnable edge extraction; Channel attention; Connectionist Temporal Classification

## 1. INTRODUCTION

Right-to-left (R-L) handwritten manuscripts are among the most enduring records of intellectual life across the Islamic world. For over a millennium, they have carried authoritative religious texts and their commentaries alongside major works in law, theology, philosophy, medicine, optics, astronomy, geography, mathematics, literature, and the arts (1). Yet their value is not limited to the main text. Many manuscripts preserve marginal notes, ownership statements, waqf records, seals, and reading certificates paratextual traces that allow scholars to study provenance, transmission chains, teaching circles, and the social history of knowledge (2). Their physical and visual properties also speak: codicological evidence such as pape inks, and bindings, together with paleographic cues like script style, ligatures, ductus, and diacritics, can help date and localize copies and illuminate how writing conventions evolved. In this sense, a manuscript is not merely a container of content; it is a crafted historical object that carries both meaning and memory (3, 4).

Despite this significance, access to historical right-to-left handwritten manuscripts remains limited in practice. Manual transcription and cataloging require specialized expertise and substantial time, which restricts discovery, search, and large-scale analysis (5). Digitization has therefore become essential for preservation and broader scholarly access (6). However, scanning alone does not make collections truly usable: without reliable machine-readable text, even rich digital archivesremain difficult to query, compare, or analyze computationally (7, 8). Figure 1 illustrates a typical manuscript page whose script variability and physical condition reflect the realities that recognition systems must confront (9).

Handwritten Text Recognition (HTR) for historical right-to-left handwritten sources is particularly demanding because right-to-left handwritten script combines structural complexity with high visual ambiguity (10). Characters frequently share the same skeletal form and differ primarily by dots or diacritics; handwriting varies across regions, time periods, and individual scribes; and manuscript pages commonly exhibit bleed-through, fading, stains, tears, skew, and uneven illumination. These factors blur stroke boundaries and suppress fine dot patterns, complicating both segmentation and recognition (11). A s a result, building robust HTR systems is not only a technical objective but also a practical requirement for scalable cultural heritagedigitization (12, 13). HTR pipelines can operate at different segmentation levels as shown in Figure 2, character level, word level, or line level each with distinct assumptions and modes (10).



Received 12 March 2026; Accepted 03 June 2026
**Available online** 06 August 2026

Figure 1. R-L handwritten manuscript (9).

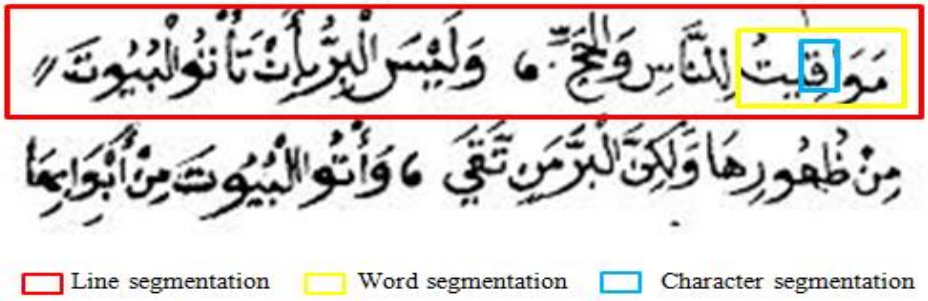


Figure 2. Segmentation levels (10).

Character-level recognition (14, 15) treats the task as classifying individual characters after segmenting them from the image. While appealing in its simplicity, it is often brittle for right-to-left handwritten manuscripts because characters are typically connected within words, diacritics may drift above or below the baseline, and degraded strokes make clean character segmentation unreliable. In practice, segmentation errors at this level tend to cascade, producing disproportionate degradation in the final transcription. Word-level recognition shifts the unit of prediction to the whole word. This can reduce the need for perfect character boundaries, but it introduces new challenges: word boundaries are not always visually clear in handwritten R-L language, and the effective vocabulary becomes very large due to rich morphology and orthographic variation especially in historical texts. Consequently, word-level systems (16-18) may struggle to generalize to rare forms and previously unseen spellings. Line-level recognition takes an entire text line as the modeling unit, typically assuming that a line can be extracted more reliably than individual characters or perfectly separated words. This approach preserves longer context, which helps disambiguate visually similar glyphs and stabilize predictions, while avoiding many of the compounding errors caused by fine-grained segmentation in noisy manuscripts (19, 20). For these reasons, line-level recognition . ,remains dominant in high-accuracy HTR systemsparticularly for historical materials where layout artifacts and degradation make detailed segmentation fragile. A common line-level pipeline encodes a line image into a sequence of visual features using a convolutional backbone, then models long-range dependencies using recurrent networks or more recently Transformer encoders. Transformers are especially effective at leveraging context without recurrence, but their success depends heavily on the quality and discriminability of the extracted visual features, which are often compromised in degraded manuscript imagery (21).

This study investigates how explicit stroke-structure modelling and channel-wise attention can strengthen R-L line-level HTR under realistic manuscript noise. We propose an enhanced convolutional neural network (CNN) –Transformer architecture that augments a baseline pipeline with: (1) a learnable edge extractor that produces a trainable stroke-boundary representation, providing a more adaptive alternative to fixed filters such as Sobel; and (2) Squeeze-and-Excitation (SE) blocks that recalibrate feature channels using global context to emphasize informative cues such as strokes and dot patterns. To quantify the contribution of these mechanisms, two variants were compared: Baseline CNN–Transformer: grayscale input, standard CNN blocks, Transformer encoder, and Connectionist Temporal Classification (CTC) -based recognition and Edge-aware SE-CNN–Transformer (proposed): grayscale input augmented with a learnable edge channel, SE-enhanced CNN blocks, a Transformer encoder, and joint training with CTC plus an auxiliary decoder objective.

Finally, we integrate a lightweight character-level 3-gram language model during beam search decoding (shallow fusion) to improve spelling consistency when visual evidence is uncertain—an especially valuable constraint in historical R-L manuscripts where minor diacritic ambiguity can alter meaning and where degradation can obscure crucial details.

## 2. RELATED WORK

Recent research on R-L handwritten text recognition (HTR) in historical documents has accelerated in response to the expanding digitization of manuscript collections and the persistent difficulty of transcribing degraded cursive R-L script. Transformer-based sequence modeling has further strengthened end-to-end recognition pipelines, particularly in OCR systems that incorporate detection and recognition as connected stages. Waly *et al.* (22) presented an end-to-end system combining CNN feature extraction with Transformer-based sequence modeling, supported by a text detection module. They reported precision, recall, and F-measure of 81.66%, 78.82%, and 79.07% for detection, and achieved a CER of 7.91% with a WER of 31.41% on handwritten R-L language. The discrepancy between relatively low CER and higher WER highlights a persistent challenge for historical handwriting: small local ambiguities often caused by weak diacritics or degraded strokes can accumulate across a line and destabilize whole-word correctness even when most characters are individually plausible.

Another influential direction reframes line and word isolation as a detection problem to improve robustness against touching components and complex layouts. Hakim *et al.* **Error! Reference source not found.** employed YOLO-based transfer learning for line and word extraction, followed by deep recognition models that combine recurrent CTC-based components with convolutional networks enhanced by Squeeze-and-Excitation blocks. Their approach achieved strong detection performance (F1-scores of 98.1% for line detection and 94.38% for word detection) and reported a recognition CER of 8.27%. Although detection-driven segmentation reduces failures associated with heuristic splitting, it remains a two-stage dependency: recognition is still constrained by detection quality, and residual CER indicates that ambiguity under ink degradation and diacritic loss remains difficult to eliminate.

Related heritage-focused efforts also underscore the importance of robust localization and normalization under deterioration. Al-Homed *et al.* **Error! Reference source not found.** addressed degraded manuscript catalog cards using deep detection and classification methods such as Faster R-CNN, reporting 92.5% classification accuracy and outperforming OCR-based baselines. While this task differs from line transcription, it reinforces the general observation that historical document analysis often hinges on reliable region extraction and geometric normalization before recognition becomes feasible.

More recent work has increasingly emphasized dataset contributions and Transformer-centric recognition tailored to R-L language. Hakim *et al.* **Error! Reference source not found.** introduced a word-level annotated dataset curated for historical R-L language detection and recognition and evaluated CNN–BLSTM and Transformer-based recognizers integrated with an R-L language model. Their reported results include CER 9.22% with WER 22.72% on KALIMA and CER 5.18% with WER 17.42% on IFN/ENIT, demonstrating the practical value of language-model support for stabilizing predictions when visual evidence is uncertain. At the same time, the broader fragmentation of historical data—differences in script, era, scanning quality, and annotation conventions—continues to make generalization across collections a central difficulty.

Chan *et al*.**Error! Reference source not found.** proposed HATFORMER, a Transformer encoder–decoder architecture adapted to R-L language using a ViT-based image processor and an R-L language-specific tokenizer, together with training strategies designed for low-resource historical conditions. They reported a CER of 8.6% on a large public historical dataset and 4.2% on a large private modern dataset, indicating that carefully adapted Transformer pipelines can generalize across domains. Nonetheless, this work also highlights a critical limitation shared by many Transformer-based systems: performance remains highly dependent on the quality of extracted visual features, and when degradation erases stroke boundaries or diacritics, attention mechanisms alone may not reliably recover missing cues.

Complementary studies explored practical deployment routes and broader restoration pipelines. Faizullah *et al.* **Error! Reference source not found.** improved Tesseract for ancient R-L language handwriting through transfer learning, reporting an average CER of 14.02% and WER of 41.39% with overall accuracy of 87.89%, providing a pragmatic option when established OCR engines are required, albeit with lower performance than specialized deep HTR systems. Miloud *et al.* **Error! Reference source not found.** investigated restoration and recognition of ancient R-L language manuscripts using an attention-based bidirectional LSTM variant and reported very high accuracy, while also emphasizing the substantial effort required to curate and annotate heritage datasets; however, differences in datasets and evaluation setups complicate direct comparisons with line-level CER/WER benchmarks. Similarly, Alqahtani and Alfahmi **Error! Reference source not found.** presented a historical R-L language OCR pipeline based on preprocessing, segmentation, and CNN recognition, reporting 99% character recognition accuracy, but also reflecting a recurring trade-off: segmentation-heavy pipelines can perform well when boundaries are recoverable, yet they may become brittle in manuscripts with severe bleed-through, broken baselines, or dense marginalia.

Accordingly, the novelty of this study lies in integrating a learnable edge representation with SE-based channel recalibration inside a controlled CNN-Transformer line-recognition framework, rather than relying only on generic visual features or segmentation-heavy pipelines. This directly addresses the identified gap in robust feature discrimination under degraded right-to-left manuscript conditions.

## 3. METHOD

This study adopts a unified, disciplined framework for R-L language handwritten text recognition in historical manuscripts, designed to ensure reproducibility and clean attribution of architectural choices. The proposed model is composed of two main components: a feature-extraction backbone and a sequence-modeling head, together forming an end-to-end line-level HTR system as illustrated in Figure 3. The backbone transforms the input line image into a compact feature map and then into a one-dimensional feature sequence, while the sequence head models long-range dependencies and produces character distributions over time. The end-to-end pipeline is designed to handle the ambiguity inherent in cursive R-L language handwriting, where character identity often depends on neighboring context and joining states.

The system is trained under a joint loss combining Connectionist Temporal Classification (CTC) (30-32) and cross-entropy (CE) (33, 34) with labels moothing. This design allows the model to exploit both alignment-free recognition and auto-regressive decoding, while also incorporating explicit structural features through the learnable edge extractor.

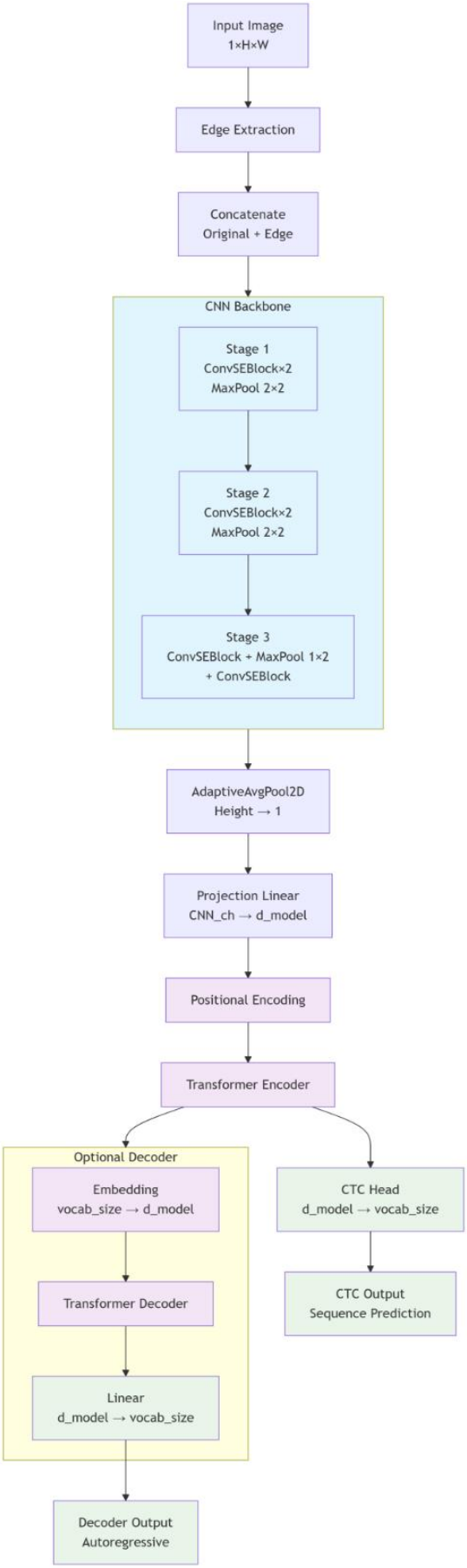

Figure 3. Architecture of proposed model.

### 3.1 Dataset

Experiments are conducted on the Kalima historical R-L language manuscript dataset **Error! Reference source not found.**, which provides line-level annotations and transcriptions suitable for end-to-end evaluation of handwriting recognition under realistic manuscript conditions. Each sample consists of a grayscale line image and its corresponding ground-truth text. The dataset consists of page images from multiple books covering varied topics, periods, and handwriting styles, with careful manual annotation and independent verifications shown in Figure 4. Each page is accompanied by ground-truth that locates text lines and words and provides transcriptions, enabling end-to-end evaluation of detection and HTR pipelines. For this work, we adopt the portion spanning Books 1–8, which totals 86 pages (71 for training and 15 for testing) and 1,759 text lines (1,443 training lines and 316 test lines).

Figure 4. Typical examples of pages from Kalima dataset (35).

### 3.2 Preprocessing and Data Preparation

Input images are resized to a fixed height while preserving aspect ratio, and within each mini-batch, images are padded to a common width to enable efficient batching. Text labels are normalized and used to build a character vocabulary from the training split. Character-to-index and index-to-character mappings are constructed to support both the CTC branch and the auxiliary decoder branch. Since Arabic is written right-to-left, label handling and decoding post-processing preserve correct directionality so that predicted outputs match the natural reading order.

To improve robustness against common manuscript distortions, the training pipeline applies realistic augmentation, including random brightness and contrast adjustments, mild perspective and affine transformations, Gaussian blur, and random erasing; elastic warping is used when available. Validation preprocessing remains deterministic (tensor conversion and normalization only) to ensure that evaluation reflects generalization rather than augmentation effects.

### 3.3 Edge-Aware SE-CNN–Transformer

The proposed model begins with a learnable edge extractor that predicts an edge-like map from the grayscale line image using a lightweight CNN with two 3×3 convolutions, batch normalization, and a sigmoid activation. The learned edge map replaces fixed filters (e.g., Sobel enables the model to discover the most informative structural gradients directly from data (36). The predicted edge map is concatenated with the original grayscale image to form a two-channel input, explicitly emphasizing stroke boundaries and fine contours that are critical in degraded manuscripts and for distinguishing visually similar Arabic characters. Table 1 shows a shape flow through the proposed HTR model.

In Table 1, B denotes the batch size, while H and W are the input line-image height and width, respectively. The symbol C refers to the number of feature channels (e.g., 64, 128, and 256), and T is the temporal sequence length obtained after collapsing the spatial feature map into a 1D sequence (approximately $T \approx W/8$ after the pooling schedule). The notation $|V|$ represents the vocabulary size (i.e., the number of output symbols/classes, including the CTC blank where applicable), and U denotes the target/output token length used by the optional decoder branch (variable per sample). The backbone is built from three stages of convolutional blocks. In the full model, each block is implemented as a ConvSEBlock that combines Conv–BatchNorm–ReLU with a Squeeze-and-Excitation (SE) (37, 38) unit for channel-wise recalibration using global context. Spatial resolution is progressively reduced via max pooling (39), resulting in an effective width reduction of

approximately 1/8. This configuration captures local stroke cues (enhanced by the edge channel) and higher-level texture/layout evidence (stabilized by SE attention).

Table 1. Shape flows through proposed HTR model.

| Stage | Layer / Operation | Shape |
|---|---|---|
| Input | Grayscale Image | B × 1 × H × W |
| Edge Extract | Learnable Edge | B × 1 × H × W |
| Concatenation | Original + Edge | B × 2 × H × W |
| Stage 1 | ConvSEBlock ×2 (2→64) | B × 64 × H × W |
| | MaxPool2d (2×2) | B × 64 × H/2 × W/2 |
| Stage 2 | ConvSEBlock ×2 (64→128) | B × 128 × H/2 × W/2 |
| | MaxPool2d (2×2) | B × 128 × H/4 × W/4 |
| Stage 3 | ConvSEBlock (128→256) | B × 256 × H/4 × W/4 |
| | MaxPool (1×2) | B × 256 × H/4 × W/8 |
| Height Pooling | ConvSEBlock (256→256) | B × 256 × H/4 × W/8 |
| | AdaptiveAvgPool2d (height→1) | B × 256 × 1 × W/8 |
| Sequence Prep | Squeeze & Permute → (B, T, C) | B × T × 256 (T ≈ W/8) |
| Projection | Linear (256→256) | B × T × 256 |
| Positional Encoding | Add PE along T | B × T × 256 |
| Transformer Encoder | L=4 layers, heads=4, ff=512, p=0.2 | B × T × 256 |
| CTC Head (main) | Linear (256→\|V\|) | B × T × \|V\| |
| CTC Output | Sequence logits over time | B × T × \|V\| |
| Optional Decoder | Embedding (\|V\|→256) | B × U × 256 |
| | Transformer Decoder (causal + cross-attn) | B × U × 256 |
| Loss (train) | Linear (256→\|V\|) | B × U × \|V\| |
| | CTC + λ·CE (decoder, optional) | — |
| Inference | Greedy/Beam (+ optional 3-gram LM), RTL post-proc | B × decoded tokens |

A multi-layer Transformer encoder (4 layers, 4 attention heads, feed-forward dimension 512, dropout 0.2) processes the feature sequence to model long-range dependencies across the entire line. This is essential for R-L language recognition, where character identity can depend on neighboring context and joining states **Error! Reference source not found.**. The encoder produces contextualized representations that integrate local shape information with global text structure. Encoder outputs are fed into two parallel heads. The first is a CTC classification head implemented as a linear projection to vocabulary logits over time, enabling alignment-free recognition and robustness when character boundaries are ambiguous. The second is an auxiliary Transformer decoder branch trained with teacher forcing using masked self-attention and cross-attention to encoder memory. The decoder is supervised using cross-entropy with label smoothing, improving sequence coherence and reducing visually motivated errors. The overall training objective is following Equation (1) **Error! Reference source not found.**:

$$Loss = CTC + \lambda\, CE_{decoder} \qquad (1)$$

where λ=0.5; λ controls the relative contribution of the auxiliary decoder cross-entropy term.

This hybrid design preserves the strengths of CTC while boosting sequence-level consistency through decoder supervision. At inference, transcriptions are produced using greedy decoding or prefix beam search from the CTC head. To improve spelling stability in noisy manuscript settings, a lightweight character 3-gram language model is trained on training transcripts and integrated into beam search via shallow fusion, controlled by beam width and LM weight. Final outputs are post-processed to ensure correct right-to-left ordering.

### 3.4 Performance Evaluation

The proposed model is evaluated using Character Error Rate (CER) and Word Error Rate (WER), computed via Levenshtein edit distance between the predicted transcription and the ground-truth text, which is standard practice in handwritten text recognition. CER measures how frequently the system makes character-level edits (substitutions, deletions, insertions), while WER captures correctness at the word level and is therefore more sensitive to errors that disrupt complete tokens. Improvements from the edge-aware components are expected to appear primarily as fewer

substitutions among visually similar Arabic characters, whereas gains from the Transformer encoder and beam-search decoding with a lightweight language model are expected to reduce long-range inconsistencies and produce more coherent sequences. Formally, CER and WER are defined as listed in Equation (2) and (3) (42-44):

$$CER = \frac{S + D + I}{N} \qquad (2)$$

$$WER = \frac{S_w + D_w + I_w}{N_w} \qquad (3)$$

$S_w D_w I_w N_w$ Here, S denotes the number of substitution errors (a reference character is replaced by an incorrect character in the prediction), D denotes the number of deletion errors (a reference character is missing in the prediction), and I denote the number of insertion errors (an extra character is added in the prediction). N is the total number of characters in the ground-truth reference. Analogously, S_w, D_w, and I_w represent word-level substitutions, deletions, and insertions, and N_w is the total number of words in the ground-truth reference. Lower CER and WER indicate better recognition performance.

## 4. Experiments

### 4.1 Experimental Setup

All experiments were implemented in PyTorch **Error! Reference source not found.**, using a fixed and reproducible configuration. The primary hyperparameters are summarized in Table 2. To enhance robustness to real-world document artifacts, training incorporated progressive resizing, starting at an input height of 96 px and switching to 128 px mid-training to preserve fine strokes and diacritics in later epochs. In addition, realistic data augmentation was applied during training, including brightness/contrast jitter, mild perspective and affine/shear transformations, Gaussian blur, and random erasing; elastic warping was used when available. Right-to-left (RTL) handling was enforced through consistent label processing and output post-processing to preserve correct Arabic language reading direction **Error! Reference source not found.**. For decoding, we used greedy decoding or prefix beam search; unless stated otherwise, all reported results use beam search with a character 3-gram language model integrated via shallow fusion to improve spelling consistency. To avoid occasional instability in CTC training when target sequences exceed available time steps, batches violating the CTC length constraint were automatically skipped.

Table 2. Model training and decoding system parameters.

| Parameter | Value | Description |
|---|---|---|
| Epochs | 100 | Number of training epochs |
| Batch_size | 4 | Batch size |
| Accum_steps | 2 | Gradient accumulation steps |
| Base_lr | 1e-3 | Base learning rate |
| Weight_decay | 1e-4 | Weight decay for regularization |
| Grad_clip | 3.0 | Gradient clipping value |
| Img_h_start | 96 | Starting image height |
| Img_h_final | 128 | Final image height (progressive resize) |

### 4.2 Compared Models

We evaluate two architectures under identical training and decoding conditions to isolate the effect of explicit structural cues and channel-wise attention. The Baseline CNN–Transformer uses single-channel grayscale input, and a conventional CNN backbone constructed from standard Conv–BN–ReLU blocks, without learnable edge extraction and without Squeeze-and-Excitation (SE) attention. The Edge-aware SE-CNN–Transformer augments the baseline with (i) a learnable edge extraction module that provides an additional structural channel emphasizing stroke boundaries, and (ii) SE attention blocks that recalibrate channel responses based on global context. Aside from these targeted modifications, the remaining components (Transformer encoder, decoding strategy, and optimization schedule) are kept fixed to ensure a controlled comparison.

This design choice is intended to isolate the contribution of the edge channel and SE attention while avoiding confounding changes in optimization, data augmentation, or decoding. Statistical significance testing was not conducted because each model was trained and evaluated once under the same controlled experimental protocol. Therefore, the reported improvements should be interpreted as controlled CER/WER differences rather than inferential statistical claims; repeated-run significance analysis is left for future work.

## 5. RESULTS AND DISCUSSION

Performance is evaluated using CER and WER, where lower values indicate better transcription quality. To ensure a fair comparison, the baseline CNN–Transformer and the proposed model are trained under identical preprocessing, augmentation, optimization, and decoding settings. Figure 5 reports the learning dynamics by tracking CER and WER across epochs for each model. In the baseline (Figure 5 (a)), error rates decrease steadily but plateau at a higher level, reflecting persistent confusions between visually similar Arabic characters and reduced robustness to degradation artifacts. By contrast, the proposed model (Figure 5 (b)) exhibits a consistently lower trajectory and smoother convergence, indicating improved discriminative feature learning and fewer character-level substitutions that propagate to word-level mistakes.

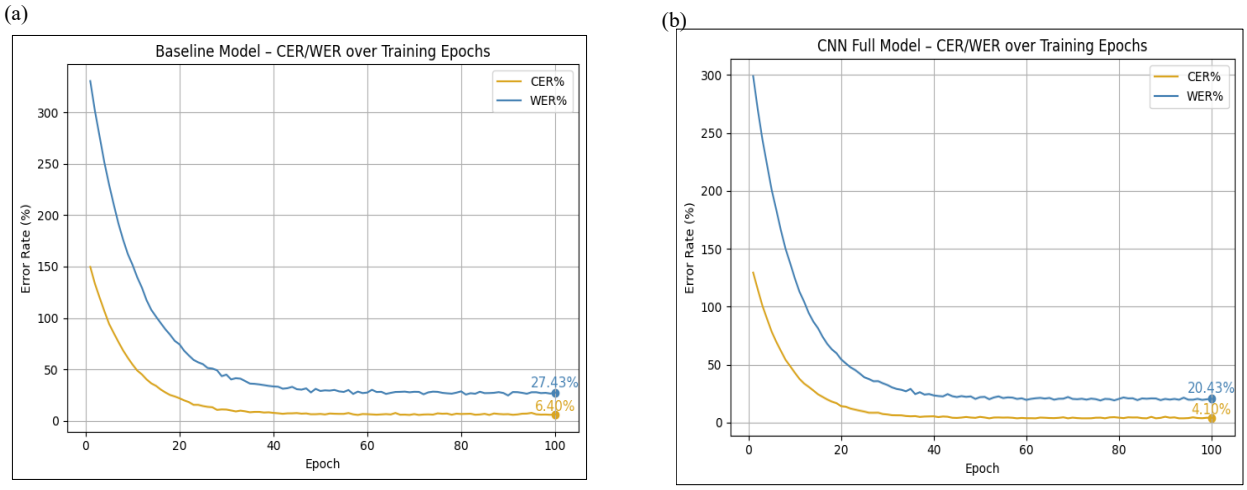


Figure 5. Training CER and WER curves for (a) baseline CNN–Transformer and (b) proposed Edge-aware SE-CNN–Transformer.

In the same context, Figure 6 summarizes the end-to-end outcome by comparing the final CER and WER achieved by both systems on the evaluation split, confirming the net benefit of the proposed architectural enhancements. Under the controlled protocol, the proposed edge-aware SE-CNN–Transformer consistently outperformed the baseline. On the primary test partitions, CER decreased from 6.40% to 4.10%, corresponding to an absolute improvement of 2.30 percentage points and a relative reduction of 35.9%. Similarly, WER decreased from 27.43% to 20.43%, yielding an absolute improvement of 7.00 percentage points and a relative reduction of 25.5% (Figure 6). These results demonstrate that incorporating learnable edge cues and SE-based channel recalibration provides substantial gains over a conventional CNN–Transformer pipeline when all other factors are held constant.

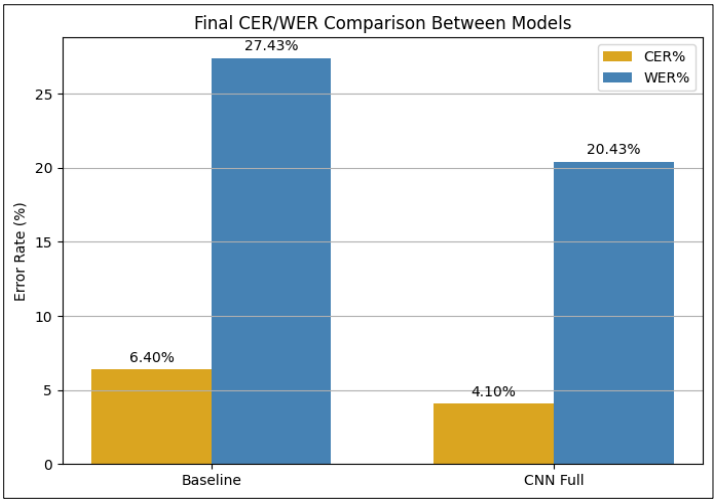


Figure 6. Performance comparison between the baseline CNN–Transformer and the proposed Edge-aware SE-CNN–Transformer models.

To contextualize performance relative to representative recent systems, Table 3 reports CER/WER values from related studies alongside our two variants. The full model achieves the strongest CER and a clearly improved WER compared to the baseline, supporting the effectiveness of explicit structural modeling and attention mechanisms for Arabic language manuscript HTR. Because the cited studies differ in dataset splits, preprocessing, and decoding protocols, Table 3 should be interpreted as contextual evidence rather than a strictly identical benchmark. The primary claim of this paper is therefore based on the controlled baseline-versus-proposed comparison conducted under the same training and decoding settings.

Table 3. Comparison of best achieved results across studies.

| Model / Study | Data Type | CER (%) ↓ | WER (%) ↓ |
|---|---|---|---|
| Hakim & Belaid (23) | Word-level (Kalima) | 8.27 | N/A |
| Hakim et al. (25) | Line-level (Kalima) | 9.22 | 22.72 |
| Bouchal et al. (35) | Line-level (Kalima) | 7.53 | 30.78 |
| Our baseline CNN–Transformer | Line-level (Kalima) | 6.40 | 27.43 |
| Our Edge-aware SE-CNN–Transformer | Line-level (Kalima) | **4.10** | **20.43** |

The observed improvements align with the intended roles of the proposed components. The learnable edge extractor explicitly emphasizes stroke boundaries and fine contours, which are critical in degraded manuscript settings and can reduce substitutions among visually similar Arabic characters. SE attention further improves robustness by recalibrating feature channels, suppressing activations caused by background texture and bleed-through artifacts while amplifying channels that correlate with discriminative handwriting cues. These cleaner, more informative features benefit the Transformer encoder, which leverages long-range context across the line to resolve ambiguities introduced by Arabic language joining behavior and context-dependent glyph shapes.

In addition, beam search decoding with a lightweight character 3-gram language model provides sequence-level regularization, improving spelling consistency and reducing linguistically implausible outputs in noisy cases. Overall, the controlled comparison indicates that combining explicit structural modeling (learnable edges) with channel-wise attention (SE) yields consistent accuracy gains in line-level R-L language HTR and is particularly suitable for historical manuscript recognition scenarios where ink degradation and stroke discontinuities are common.

A limitation of the current study is that the reported gains are evaluated on the selected Kalima line-image split only; therefore, cross-dataset robustness and the individual contribution of each component should be further verified in future work through ablation and external validation experiments.

## 6. CONCLUSION

This paper presented an edge-aware line-level HTR system for degraded right-to-left historical manuscripts. By adding a learnable edge-extraction channel and SE-based feature recalibration to a CNN-Transformer baseline, the proposed model improved the discriminability of stroke boundaries, dot patterns, and visually similar cursive characters under manuscript noise. On the Kalima line-image split, the model reduced CER from 6.40% to 4.10% and WER from 27.43% to 20.43% under the same training and decoding conditions. For digital humanities and archival systems, this framework can support more reliable transcription, search, and metadata enrichment for degradation-prone manuscript collections. Future work will include repeated-run significance testing, more detailed ablation, cross-dataset validation, and stronger language-prior integration while preserving provenance and controllability.

## AUTHORSHIP CONTRIBUTION STATEMENT

Bilal Abdulrahman: conceptualization, methodology, software, formal analysis, investigation; Farhan Mohamed: review, editing.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## DECLARATION OF COMPETING INTEREST

The authors declare no conflict of interest.

## DECLARATION OF GENERATIVE AI

The authors used ChatGPT (OpenAI) during the preparation of this manuscript to improve grammar, paraphrase selected passages, and enhance clarity and readability. After using the tool, the authors carefully reviewed and edited the content and take full responsibility for the content of the publication.

## REFERENCES

(1) Waley MI. Manuscripts and their importance in Islamic cultural and religious studies (Internet). MAYDAN: Teaching & Learning. 2018 Feb 28 (cited 2026 Jun 11). Available from: https://themaydan.com/2018/02/manuscripts-importance-islamic-cultural-religious-studies
(2) de Bakker A. Collections, collectors, and a breviary leaf at McGill (Ms. Medieval 0239). J Alamire Found. 2025; 17(2):227–241. https://doi.org/10.1484/J.JAF.5.152940.
(3) Ayele YS, Saez NO, Bezie BT, Taye B, Demssie DAS, Dires DA. Material history of Ethiopic manuscripts: Original repair, damage, and anthropogenic impact. Arts. 2025; 14(6):173. https://doi.org/10.3390/arts14060173.
(4) Gacek A. Arabic manuscripts: A vademecum for readers. Vol. 98. Leiden: Brill; 2009. https://doi.org/10.1163/ej.9789004170360.i-350.
(5) Grallert T. Open Arabic periodical editions: A framework for bootstrapped scholarly editions outside the Global North. Digit Humanit Q. 2022; 16(2):1–20. https://doi.org/10.63744/w9dk2dm7m92p.
(6) Girdhar N, Coustaty M, Doucet A. Digitizing history: transitioning historical paper documents to digital content for information retrieval and mining—a comprehensive survey. IEEE Trans Comput Soc Syst. 2024; 11(5):6151–6180. https://doi.org/10.1109/tcss.2024.3378419.
(7) Nockels J, Gooding P, Terras M. The implications of handwritten text recognition for accessing the past at scale. J Doc. 2024; 80(7):148–167. https://doi.org/10.1108/jd-09-2023-0183.
(8) Humphries M, Leddy LC, Downton Q, Legace M, McConnell J, Murray I, Spence E. Unlocking the archives: using large language models to transcribe handwritten historical documents. Hist Methods. 2025:1–19. https://doi.org/10.1080/01615440.2025.2500309.
(9) Khayyat MM, Elrefaei LA. A deep learning-based prediction of Arabic manuscripts handwriting style. Int Arab J Inf Technol. 2020; 17(5):702–712. https://doi.org/10.34028/iajit/17/5/3.
(10) Abdulrahman Tuama B, Mohamed F. A systematic literature review of deep learning methods for handwritten text recognition in historical Arabic manuscripts. Eng Technol Appl Sci Res. 2025; 15(4):25772–25782. https://doi.org/10.48084/etasr.12123.
(11) Alrehali B, Alsaedi N, Alahmadi H, Abid N. Historical Arabic manuscripts text recognition using convolutional neural network. Proc 6th Conf Data Science and Machine Learning Applications (CDMA). IEEE; 2020. 37–42. https://doi.org/10.1109/cdma47397.2020.00012.
(12) Saeed M, Chan A, Mijar A, Habchi G, Younes C, Wong CW, Khater A. Muharaf: Manuscripts of handwritten Arabic dataset for cursive text recognition. Adv Neural Inf Process Syst. 2024; 37:58525–58538. https://doi.org/10.48550/arXiv.2406.09630.
(13) Aljishi F, Mughaus R, Luqman H, Parvez MT. A comparative study of four handwritten text recognition models in Arabic script. Ing Syst Inf. 2024; 29(6):2243. https://doi.org/10.18280/isi.290614.
(14) Qaroush A, Jaber B, Mohammad K, Washaha M, Maali E, Nayef N. An efficient, font independent word and character segmentation algorithm for printed Arabic text. J King Saud Univ Comput Inf Sci. 2022; 34(1):1330–1344. https://doi.org/10.1016/j.jksuci.2019.08.013.
(15) Jindal A, Ghosh R. Text line segmentation in Indian ancient handwritten documents using faster R-CNN. Multimed Tools Appl. 2023; 82(7):10703–10722. https://doi.org/10.1007/s11042-022-13709-y.
(16) Omayio EO. Word segmentation by component tracing and association (CTA) technique. J Eng Res. 2022. https://doi.org/10.36909/jer.15207.
(17) Zhang C. Improved word segmentation system for Chinese criminal judgment documents. Appl Artif Intell. 2024; 38(1):2297524. https://doi.org/10.1080/08839514.2023.2297524.
(18) Mahajan S, Rani R, Trehan K. DELIGHT-Net: DEep and LIGHTweight network to segment Indian text at word level from wild scenic images. Int J Multimed Inf Retr. 2023; 12(2):29. https://doi.org/10.1007/s13735-023-00293-6.
(19) Droby A, Kurar Barakat B, Alaasam R, Madi B, Rabaev I, El-Sana J. Text line extraction in historical documents using mask R-CNN. Signals. 2022; 3(3):535–549. https://doi.org/10.3390/signals3030032.
(20) Abuain W. Text-line segmentation techniques for Arabic-handwritten documents: A review. Stat Optim Inf Comput. 2025; 14(1):329–339. https://doi.org/10.19139/soic-2310-5070-2466.
(21) Elharrouss O, Al-Maadeed S, Alja'am JM, Hassaine A. A robust method for text, line, and word segmentation for historical Arabic manuscripts. In: Data analytics for cultural heritage: current trends and concepts. Cham: Springer; 2021. p. 147–172. https://doi.org/10.1007/978-3-030-66777-1_7.
(22) Waly A, Tarek B, Feteha A, Yehia R, Amr G, Gomaa W, Fares A. Invizo: Arabic handwritten document optical character recognition solution. https://doi.org/10.48550/arXiv.2502.05277.
(23) Hakim B, Belaid A. Text line and word detection and recognition of historical Arabic manuscripts (Internet). Research Square [Preprint]. 2023 (cited 2026 Jun 11). https://doi.org/10.21203/rs.3.rs-2883455/v1
(24) Al-Homed LS, Jambi KM, Al-Barhamtoshy HM. A deep learning approach for Arabic manuscripts classification. Sensors. 2023; 23(19):8133. https://doi.org/10.3390/s23198133.
(25) Hakim B, Belaid A. Deep learning for accurate recognition of Arabic handwritten words in historical documents. Procedia Comput Sci. 2024; 244:57–65. https://doi.org/10.1016/j.procs.2024.10.178.
(26) Chan A, Mijar A, Saeed M, Wong CW, Khater A. Hatformer: Historic handwritten Arabic text recognition with transformers. https://doi.org/10.48550/arXiv.2410.02179.
(27) Faizullah S, Ayub MS, Alghamdi T, Ali TS, Khan MA, Nabil E. Revolutionizing historical document digitization: LSTM-enhanced OCR for Arabic handwritten manuscripts. Int J Adv Comput Sci Appl. 2024; 15(10):1–10. https://doi.org/10.14569/ijacsa.2024.01510120.
(28) Miloud K, Abdelmounaim ML, Mohammed B, Ilyas BR. Restoration of ancient Arabic manuscripts: A deep learning approach. Stud Eng Exact Sci. 2024; 5(2):e7722. https://doi.org/10.54021/seesv5n2-183.
(29) Alqahtani AA, Alfahmi SS. System for detection and recognition of historical Arabic manuscripts. Research Square. 2025. arXiv:2410.02179v2. https://doi.org/10.21203/rs.3.rs-5936450/v1

(30) Buoy R, Iwamura M, Srun S, Kise K. Explainable connectionist-temporal-classification-based scene text recognition. J Imaging. 2023; 9(11):248. https://doi.org/10.3390/jimaging9110248.
(31) Wang JY, Jang JSR. Training a singing transcription model using connectionist temporal classification loss and cross-entropy loss. IEEE ACM Trans Audio Speech Lang Process. 2022; 31:383–396. https://doi.org/10.1109/taslp.2022.3224297.
(32) Kubra KT, Umair M, Zubair M, Naseem MT, Lee CS. Detection and recognition of bilingual Urdu and English text in natural scene images using a convolutional neural network–recurrent neural network combination with a connectionist temporal classification decoder. Sensors. 2025; 25(16):5133. https://doi.org/10.3390/s25165133.
(33) Mao A, Mohri M, Zhong Y. Cross-entropy loss functions: theoretical analysis and applications. Proc Int Conf Mach Learn (ICML). PMLR; 2023. 23803–23828. https://doi.org/10.48550/arXiv.2304.07288.
(34) Li Y, Zou Y, Glorioso P, Altman E, Fisher MPA. Cross entropy benchmark for measurement-induced phase transitions. Phys Rev Lett. 2023; 130(22):220404. https://doi.org/10.1103/physrevlett.130.220404.
(35) Bouchal B, Belaid A, Meziane F. Towards accurate recognition of historical Arabic manuscripts: A novel dataset and a generalizable pipeline. ACM Trans Asian Low-Resour Lang Inf Process. 2025. https://doi.org/10.1145/3744243.
(36) Ranjan R, Avasthi V. Edge detection using guided Sobel image filtering. Wirel Pers Commun. 2023; 132(1):651–677. https://doi.org/10.1007/s11277-023-10628-5.
(37) Jin X, Xie Y, Wei XS, Zhao BR, Chen ZM, Tan X. Delving deep into spatial pooling for squeeze-and-excitation networks. Pattern Recognit. 2022; 121:108159. https://doi.org/10.1016/j.patcog.2021.108159.
(38) Huang M, Wan N, Zhu H. Reconstruction of structural acceleration response based on CNN-BiGRU with squeeze-and-excitation under environmental temperature effects. J Civ Struct Health Monit. 2025; 15(3):985–1003. https://doi.org/10.1007/s13349-024-00859-w.
(39) Matoba K, Dimitriadis N, Fleuret F. Benefits of max pooling in neural networks: Theoretical and experimental evidence. Trans Mach Learn Res. 2023. Available from: https://openreview.net/pdf?id=YgeXqrH7gA.
(40) Yang J, Shokouhifar M, Yee L, Khan AA, Awais M, Mousavi Z. DT2F-TLNet: A novel text-independent writer identification and verification model using a combination of deep type-2 fuzzy architecture and transfer learning networks based on handwriting data. Expert Syst Appl. 2024; 242:122704. https://doi.org/10.1016/j.eswa.2023.122704.
(41) Chen Y, Xu K, Zhou P, Ban X, He D. Improved cross entropy loss for noisy labels in vision leaf disease classification. IET Image Process. 2022; 16(6):1511–1519. https://doi.org/10.1049/ipr2.12402.
(42) Mutawa AM, Allaho MY, Al-Hajeri M. Machine learning approach for Arabic handwritten recognition. Appl Sci. 2024; 14(19):9020. https://doi.org/10.3390/app14199020.
(43) Li DL, Lee SK, Liu YT. Printed document layout analysis and optical character recognition system based on deep learning. Sci Rep. 2025; 15(1):23761. https://doi.org/10.1038/s41598-025-07439-y.
(44) Khallouli W, Uddin MS, Sousa-Poza A, Li J, Kovacic S. Leveraging transformer-based OCR model with generative data augmentation for engineering document recognition. Electronics. 2024; 14(1):5. https://doi.org/10.3390/electronics14010005.
(45) Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. Vol. 32. Curran Associates; 2019. https://doi.org/10.48550/arXiv.1912.01703.
(46) Najam R, Faizullah S. Analysis of recent deep learning techniques for Arabic handwritten-text OCR and post-OCR correction. Appl Sci. 2023; 13(13):7568. https://doi.org/10.3390/app13137568.