Jawi Lens an AI Based Digital Platform for Jawi Manuscript Recognition Transliteration and Malay Literary Heritage Preservation

Muhammad Irfan Bin Mhd Rusdi*, Ibrahim Bin Abdullah, Janudin Sardi @ Hj. Mohd Yusop

Academy of Language Studies, Universiti Teknologi MARA, Shah Alam, Malaysia

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.102100210

Received: 06 August 2026; Accepted: 11 August 2026; Published: 13 October 2026

ABSTRACT

Historical Malay Jawi manuscripts contain valuable records of religious, literary and intellectual traditions, but digitised images remain difficult to search and reuse without reliable transcription. This paper presents JAWI LENS as a proposed pilot platform integrating manuscript digitisation, optical character recognition (OCR), Jawi-to-Roman transliteration, expert validation and a searchable database. It documents the methodological framework for the next stage of grant-supported research. The formal pilot has not commenced, and empirical evaluation has not yet been completed. The proposed protocol compares a ResNet-18-BiLSTM-CTC recogniser with a fine-tuned TrOCR-base model using expert-verified text lines and manuscript-disjoint evaluation. Dataset scarcity will be addressed through controlled image augmentation, synthetic Jawi text generation, staged transfer learning and expert-validated pseudo-labelling. Recognition will be assessed using Character Error Rate (CER), Word Error Rate (WER), exact-line accuracy and degradation-stratified analysis. Transliteration will be evaluated separately on verified Jawi input and on OCR output to distinguish transliteration errors from recognition errors. Provisional project milestones are OCR CER of 10% or lower and WER of 20% or lower, and transliteration CER of 5% or lower and WER of 10% or lower on verified input. These are prospective planning criteria, not observed results or established field standards. A controlled contribution workflow will retain source images, alternative readings, reviewer decisions and correction histories. The contribution is a testable pilot design linking technical evaluation with scholarly traceability and access to Malay literary heritage, while explicitly defining the evidence needed before claims of system effectiveness can be made.

Keywords: Jawi manuscripts; optical character recognition; transliteration; pilot research; Digital Humanities; heritage preservation

INTRODUCTION

Historical Malay Jawi manuscripts provide evidence for the study of religious thought, literature, language and intellectual exchange. Digitisation can preserve access to their visual form, but a collection of page images does not automatically become a searchable textual resource. Researchers still need to identify relevant passages, interpret handwriting and connect transcriptions to their documentary sources. JAWI LENS is proposed to support these activities through an integrated workflow for image processing, recognition, transliteration and scholarly validation.

Jawi recognition requires attention to connected letter forms, dot placement and additional letters used to represent Malay sounds. Historical material also presents variation in handwriting, orthography, page layout and preservation. Character classification research provides useful foundations for Jawi processing (Razali et al., 2024), but classification of isolated or prepared characters does not establish the accuracy of continuous-line transcription from damaged manuscripts. Similarly, successful recognition in another script or dataset cannot substitute for evaluation on the manuscripts that a heritage platform intends to serve.

The practical motivation includes the corresponding author's exploratory use of ChatGPT to read uploaded manuscript PDFs and produce Roman-script renderings. As a Jawi specialist, the author inspected the output

against the manuscript and identified word-level mismatches. This activity combined document reading and transliteration; it did not establish a separately scored OCR output or identify the underlying recognition implementation. Because corpus counts, reference word totals and a reproducible scoring record are unavailable, the observations are reported qualitatively and no accuracy percentage is inferred. They motivate the formal pilot rather than constitute its results.

This article is a methodological framework and prospective pilot protocol, rather than an empirical report of a completed system. Its objectives are to specify a manageable manuscript corpus and annotation procedure; define two reproducible OCR configurations; establish separate OCR and transliteration evaluation; and describe a validation and database workflow appropriate for scholarly reuse. The framework also provides a basis for reporting the design and planned evaluation activities within the JAWI LENS grant project.

Four questions will guide the pilot. RQ1 asks how the proposed OCR configurations compare on previously unseen manuscripts. RQ2 asks how preprocessing, augmentation, synthetic data and transfer affect recognition under different degradation conditions. RQ3 asks how accurately the transliteration module converts verified Jawi text and how its performance changes when supplied with OCR output. RQ4 asks whether an image-linked review workflow can produce traceable corrections and support retrieval of verified passages. The present article specifies how these questions will be investigated; it does not claim that they have already been answered.

RELATED WORK AND RESEARCH RATIONALE

Jawi recognition and the unit of evaluation

Razali et al. (2024) constructed additional Arabic-derived character data for Jawi and evaluated deep-learning classification models. Their work supports the relevance of script-specific training material and augmentation. Its character-classification setting differs from the proposed JAWI LENS task, in which a model must transcribe an entire line while handling connected forms, spacing and damage. Reported classification accuracy should therefore not be presented as an expected line-level OCR result. This distinction informs the choice of sequence-based error measures and the requirement for manuscript-specific ground truth.

Sequence recognition and pretrained models

Shi et al. (2015) describe an architecture combining visual feature extraction, sequence modelling and transcription for image-based text recognition. Connectionist Temporal Classification (CTC) provides a way to train sequence models without character-by-character alignment (Graves et al., 2006). Residual networks provide the visual feature-extraction design adopted in the proposed baseline (He et al., 2015). JAWI LENS combines these established components into a specified ResNet-BiLSTM-CTC configuration; it does not claim a new neural architecture.

Li et al. (2023) introduce TrOCR, an encoder-decoder approach using pretrained image and text Transformers. This provides a second, technically distinct recognition configuration. Pretraining is a reason to investigate the model, not evidence that it already recognises historical Jawi adequately. The planned comparison will document the initial weights, vocabulary coverage, preprocessing and fine-tuning conditions so that any difference can be interpreted as a comparison between complete configurations rather than architecture alone.

Low-resource transcription and heritage platforms

Garces Arias et al. (2023) combine synthetic data, augmentation and Transformer-based recognition for handwritten Old Occitan. Their work motivates a staged approach to scarce annotations, while differences in script, handwriting and orthography limit direct transfer of their findings to Jawi. Synthetic examples in the proposed pilot will supplement real manuscript lines and will be evaluated against an exclusively real-image test set.

Sakhovskiy et al. (2026) present InkSight, which connects recognition with indexing and retrieval for historical Arabic manuscripts. This illustrates the research value of connecting textual extraction to access functions. JAWI

LENS takes a narrower pilot approach: verified Jawi transcription, Roman-script access and source-linked corrections. Its intended contribution is the integration of a Jawi-specific evaluation protocol with manuscript provenance and controlled human validation. The literature discussed here is a targeted methodological selection, not a systematic review or evidence that no other relevant system exists.

PROPOSED PILOT METHODOLOGY

Research status and design

The formal JAWI LENS pilot has not been implemented. A scored OCR corpus, completed model comparison and validated transliteration results are not yet available. Accordingly, all sampling quantities, configurations and thresholds below are proposed protocol settings. They will be recorded in a dated protocol before formal testing; subsequent changes will be documented with reasons. The study will combine quantitative recognition and transliteration evaluation with qualitative analysis of errors and reviewer feedback.

The primary recognition unit will be a manuscript text line paired with a verified Jawi transcription. Page-level processing will be evaluated separately because line recognition on manually checked crops does not measure the reliability of automatic page segmentation. The proposed workflow is shown in Figure 1. Each stage will retain links to the source manuscript and distinguish machine-generated text from expert-verified text.

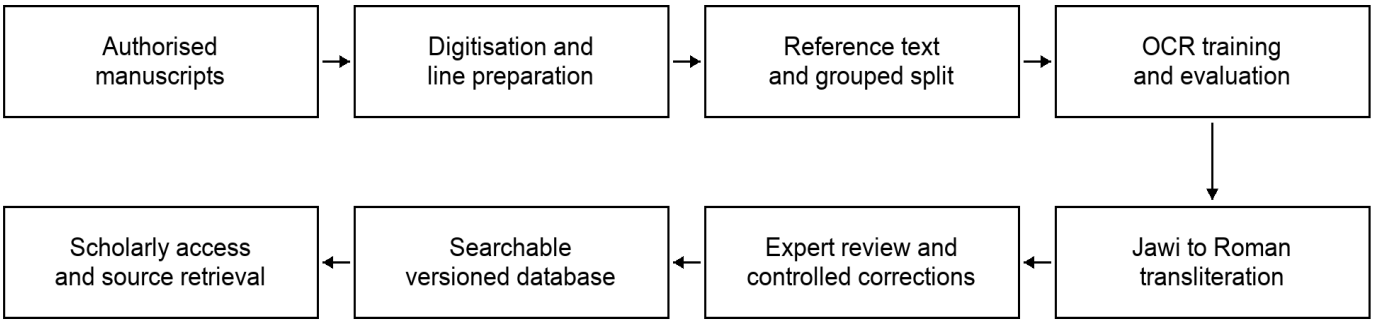


Figure 1 Proposed pilot workflow with separate OCR and transliteration evaluation

Corpus selection and data partitioning

An initial planning target is approximately 1,000 expert-verified lines drawn from at least ten manuscript or identifiable scribal groups, subject to access permissions, annotation capacity and image suitability. These quantities are recruitment targets, not an existing dataset or a statistical power calculation. Selection will seek variation in manuscript content, handwriting and preservation rather than only clear pages. The corpus register will report collection identifiers, selection criteria, rights, known dates, manuscript and page counts, line counts and reasons for exclusion. Unknown dates or scribal identities will remain explicitly unknown.

Where corpus composition permits, approximately 70% of lines will be allocated to training, 15% to validation and 15% to testing, with whole manuscript groups assigned together. Known shared hands and duplicate or closely related page images will be grouped to reduce leakage. The proportions are approximate because group integrity takes precedence over exact line counts. If too few independent groups are available, grouped cross-validation will be considered and the resulting uncertainty disclosed. A small pilot cannot establish generalisation across the full Malay manuscript tradition.

The split register will be frozen before augmentation, vocabulary adaptation or model selection. Test pages and their transcriptions will not be used to generate synthetic training text, tune transliteration dictionaries or obtain pseudo-labels. An expanded training pool must also exclude manuscripts and known hands reserved for testing. All comparisons will use the same eligible test items and report the number of manuscripts as well as lines.

Digitisation and reference transcription

Images will be obtained from authorised collections or captured under collection-approved handling procedures. For new captures, 300-600 dpi at the original page scale is a proposed starting range, with adequacy judged by the visibility of dots and fine strokes. Existing images will retain their recorded resolution without implying that enlargement creates additional detail. Original images will remain unchanged; processed derivatives will carry identifiers linking them to the originals.

Pages will be segmented into line crops and manually checked for clipped dots, overlapping lines and reading order. A Jawi specialist will produce a diplomatic transcription preserving the source spelling rather than silently modernising it. The annotation guide will define letters, punctuation, spaces, diacritics, ligatures, editorial marks and uncertain readings. Unicode normalisation will follow a fixed documented policy while retaining distinctions between different Jawi letters. Display direction will be separated from logical character order in stored text.

A second specialist will independently transcribe a stratified sample of at least 20% of training lines and all validation and test lines. Disagreements will be recorded before adjudication and summarised using character-level disagreement. Reference transcriptions will be finalised before viewing model predictions. Lines containing irrecoverable text will enter a separate challenge set rather than receive guessed ground truth. The number and degradation profile of excluded lines will be reported alongside the scored set, preventing improved scores through undisclosed removal of difficult material.

Degradation and preprocessing

Two annotators will assign each line a degradation category before model evaluation using the operational rubric in Table 1. Fading, bleed-through, stains, blur, skew and physical loss will also be recorded as separate attributes. Agreement and adjudicated labels will be reported. The categories describe this pilot's image conditions and will not be presented as a universal conservation scale.

Table 1 Proposed degradation rubric

Category	Operational criterion	Evaluation treatment
Mild	Main strokes and distinguishing dots are clear; minor background disturbance does not obstruct transcription.	Score against verified transcription.
Moderate	Local fading, blur or background interference obscures some features, but specialists can resolve the complete line from the image.	Score separately and retain damage attributes.
Severe	Extensive interference or stroke loss makes reading difficult; a complete reference remains possible after independent specialist review.	Score separately and report eligible line count.
Unresolvable	At least one text span cannot be read reliably after adjudication.	Retain in the challenge set; report exclusion from primary CER/WER.

The minimal-processing condition will apply only cropping, aspect-preserving resizing and required input normalisation. Candidate enhancement will include contrast adjustment, gentle denoising, deskewing and adaptive binarisation. Enhancement choices will be selected on validation data and then fixed. Original and processed crops will be inspected together to check that dots and thin strokes have not been removed. No restoration method will be allowed to replace missing writing with unverified inferred text in the reference corpus.

OCR configurations and training

The convolutional recurrent neural network (CRNN) baseline will use a ResNet-18 feature extractor with the classification head removed. Horizontal downsampling will be limited to a factor of four, followed by pooling over feature height, two bidirectional LSTM layers with 256 hidden units per direction and a linear character

projection trained with CTC. Images will remain in their original orientation. The extracted horizontal feature sequence will be traversed from right to left while labels remain in logical Unicode order. Training will check that the available sequence length permits alignment, including repeated characters.

The Transformer configuration will start from microsoft/trocr-base-handwritten and will be fine-tuned on the same real Jawi training lines. This is an initial checkpoint choice, not a Jawi-specific performance claim. Its input processor will use 384 by 384 pixels. The retained tokenizer will be tested for lossless encoding and decoding of the full Jawi inventory before training; failures will require documented vocabulary adaptation using training material only. Long lines that cannot be represented without harmful loss of detail will be flagged for a documented segmentation or input-resolution revision before the test protocol is frozen.

Table 2 specifies starting settings rather than optimised values. Both configurations will use identical manuscript partitions, common scoring rules and comparable tuning effort. Within-family comparisons will estimate preprocessing and augmentation effects. Between-family comparisons will also reflect different pretrained representations and input processing, so they will not isolate architecture as the sole cause of a difference.

Table 2 Proposed initial OCR configuration

Setting	**ResNet BiLSTM CTC**	**TrOCR base**
Image input	Height 64 pixels; preserve aspect ratio; pad batches to valid widths	384 by 384 pixels through the checkpoint processor
Initial weights	ResNet-18 ImageNet weights; new recurrent and output layers	microsoft/trocr-base-handwritten
Output	Training-derived Jawi character inventory and CTC blank	Existing tokenizer after Jawi round-trip validation
Training loss	CTC	Token cross-entropy
Optimiser and initial rate	AdamW; 0.0001	AdamW; 0.00005
Effective batch size	16 lines	16 lines using gradient accumulation if needed
Training limit	100 epochs; stop after 10 epochs without validation CER improvement	Same stopping rule
Primary decoding	Greedy CTC without external language model	Greedy autoregressive decoding
Replication	Three recorded random seeds	Three recorded random seeds

A separate service benchmark, E0, will formalise the author's ChatGPT-assisted workflow where document access permissions allow. For the line-level comparison, it will use the same held-out line images and fixed prompts requesting a Jawi transcription and a separate Roman rendering. Page-level PDF processing will be reported separately. The visible model or service label, access date, prompts, input files, conversation context and unedited outputs will be retained. This benchmark evaluates the observed service workflow; it will not attribute outcomes to an undisclosed OCR architecture. If separate Jawi output cannot be obtained, only combined Roman-output performance will be reported. Changes in service availability or model behaviour will be documented as reproducibility limitations.

The experiment log will retain exact model revisions, software versions, random seeds, learning-rate schedules, input limits, hardware, training duration and selected epochs. The physical batch size may be reduced for memory constraints while retaining the effective batch size. Any learning-rate tuning will use a predeclared small search around the starting values and validation CER only. Test-set results will not guide checkpoint selection.

Dataset expansion and transfer learning

Image augmentation will be applied only to training images. Initial candidates are rotations within three degrees, brightness and contrast changes, mild blur, background noise and simulated fading or bleed-through. Transformations that clip text, alter character identity or destroy the intended transcription will be rejected. A specialist will inspect samples from each augmentation family, with particular attention to dots and rare Jawi letters. The transformation ranges and sampling probabilities will be logged.

Synthetic lines will be rendered from approved training-partition Jawi text using fonts with verified contextual shaping. Font-based images do not reproduce all historical handwriting and will be labelled as synthetic. They will be introduced initially at one synthetic sample per real training line, with alternative ratios selected only through validation. Text from held-out manuscripts and undisclosed generated quotations will not enter the synthetic corpus. A rendering audit will check joining, dots and character coverage before degradation effects are applied.

Arabic-script transfer will be a separate experimental arm: a compatible recogniser will undergo intermediate training on licensed Arabic line-image/transcription pairs before Jawi fine-tuning. The external corpus and any Arabic-adapted checkpoint have not yet been selected; their provenance, licence, vocabulary and compatibility must be established before this arm begins. If suitable material is unavailable, the arm will be recorded as not undertaken. This condition avoids claiming that a transfer resource or tested model already exists.

Pseudo-labelling will follow the first supervised comparison. A model selected on validation data will propose transcriptions for previously unlabelled training-pool lines. Model confidence will prioritise candidates, but every admitted line must be checked and corrected by a specialist. Accepted, corrected and rejected counts, review time and model version will be recorded for each round. Confidence thresholds will be set using validation behaviour and will not be interpreted as calibrated probabilities without supporting checks. Difficult lines will also be sampled for manual annotation so that expansion does not concentrate exclusively on easy handwriting.

Jawi to Roman transliteration

The pilot transliteration module will use an explicit rule system and a versioned specialist-reviewed lexicon. A character-mapping baseline will first establish the effect of simple conversion. The proposed enriched module will apply longest-match lexical entries before fallback character rules and will flag unresolved words. Each dictionary entry will retain its Jawi form, Roman rendering, source context, accepted alternatives and review status. Unseen or ambiguous forms will not be silently converted into confident modern spellings.

A written transliteration convention will be agreed before reference creation and model evaluation. The primary Roman layer will represent the Malay reading supported by the source, with rules for vowel supply, word boundaries, proper names and Arabic-derived terms. The original Jawi transcription will remain available. Any modernised spelling intended for search will be stored separately and will not replace the primary scholarly layer. Admissible alternative Roman readings will be fixed during expert adjudication before system outputs are inspected.

Two evaluation tracks will be used. Track T1 will supply verified Jawi transcription to the transliteration module, isolating conversion performance. Track T2 will supply raw OCR predictions from the same lines, measuring the combined pipeline. Both will be compared with expert-verified Roman references. Primary scores will use a single adjudicated reference; a separately labelled sensitivity analysis may allow pre-approved alternatives. Abstentions and unresolved words will remain errors in full-coverage automatic scoring, while their frequency and review burden will also be reported.

Evaluation Protocol and Project Milestones

Measures and provisional thresholds

Character Error Rate will be calculated as CER = (S + D + I) / N, where S, D and I are character substitutions, deletions and insertions and N is the number of reference characters. WER will use the same calculation over whitespace-delimited words after the prescribed normalisation. Punctuation and diacritics will follow the annotation policy consistently. Scores will be reported as percentages using summed edit counts and reference lengths across the eligible test set; a separate manuscript-level summary will show variation between sources.

Exact-line accuracy will be the proportion of lines matching the reference completely. A separate undefined character-accuracy measure will not be used, because expressing 100% minus CER as another target would

duplicate the main criterion and may be misleading when insertions are numerous. Relative CER reduction will be calculated as (baseline CER minus treatment CER) divided by baseline CER, with no relative value reported when baseline CER is zero.

Table 3 defines provisional internal milestones. The OCR limits represent a planning tolerance of up to ten character edits per hundred reference characters and twenty word edits per hundred reference words. The stricter transliteration limits reflect the expectation that T1 receives verified text and does not need to resolve damaged images. These values are project planning choices, not literature-derived universal standards or forecasts. Their practical adequacy will be examined alongside manual correction time and expert judgement. Any revision must precede final testing and be retained in the protocol history.

Table 3 Prospective evaluation criteria with no results yet available

Component	Measures	Proposed milestone or reporting rule
OCR on real held-out lines	CER, WER and exact-line accuracy	CER ≤10% and WER ≤20%; report exact-line accuracy without an assumed target
Robustness	CER/WER by degradation and manuscript	Report every eligible stratum, sample size and change from baseline; no universal threshold assumed
T1 transliteration from verified Jawi	Roman CER and WER	CER ≤5% and WER ≤10%
T2 transliteration from OCR output	Roman CER and WER	Report absolute errors and the difference from T1; no unsupported combined threshold
Page processing	Missed or incorrectly cropped lines; reading-order errors	Report separately from recognition on checked crops
Human validation	Time per line; corrections; unresolved cases	Describe effort and remaining uncertainty; no productivity claim before measurement
Search and traceability	Correct source retrieval on predefined queries	Every released verified passage must link to its source image and review record

Controlled comparisons and uncertainty

Table 4 separates the principal experiments so that multiple data strategies are not introduced simultaneously. Each OCR comparison will retain the same real-data partitions and evaluation script. E7 and E8 will be compared independently with the validation-selected augmented configuration. E9 will add reviewed pseudo-labels to one configuration fixed before the expansion round. The most favourable test score will not be used to choose the final model.

Table 4 Planned experimental comparisons

ID	Configuration or comparison	Question addressed
E0	Fixed ChatGPT-assisted reading and Romanisation workflow	Service reference with separate Jawi and Roman outputs where available
E1	CRNN with minimal processing	Establish OCR baseline
E2	CRNN with selected enhancement	Enhancement effect relative to E1
E3	E2 with training augmentation	Augmentation effect relative to E2
E4	TrOCR with minimal processing	Establish Transformer configuration baseline
E5	TrOCR with selected enhancement	Enhancement effect relative to E4
E6	E5 with training augmentation	Augmentation effect relative to E5
E7	Selected E3 or E6 plus synthetic lines	Synthetic-data effect without added Arabic transfer
E8	Selected E3 or E6 plus intermediate Arabic training	Arabic-transfer effect without added synthetic lines

E9	Fixed selected configuration plus validated pseudo-labels	Effect and review cost of one expansion round
T1 and T2	Mapping baseline versus lexicon-enriched transliteration	Conversion gain and OCR error propagation

Results will include all three training seeds rather than only the best run. Where the number of independent manuscripts permits, paired bootstrap intervals will resample manuscript groups rather than treating neighbouring lines as independent observations. With very few groups, intervals will be labelled unstable and manuscript-level results will be prioritised over claims of statistical superiority. Error analysis will inspect dot confusion, joined forms, rare letters, damaged strokes, segmentation and orthographic ambiguity. OCR errors, transliteration-rule errors and legitimate alternative readings will be distinguished through expert review.

Grant reporting and progression decisions

The present article records a design-stage contribution. Subsequent reporting will distinguish four milestones: an authorised corpus and annotation guide; a frozen split and reference set; reproducible model and transliteration experiments; and an image-linked searchable demonstrator. Each milestone will be supported by dated evidence such as corpus counts, annotation records, configuration files, evaluation outputs and demonstration logs. Planned activities will not be reported as completed deliverables.

Meeting the numerical targets will indicate progress towards the pilot's internal objectives, not readiness for unrestricted scholarly reliance. Failure to reach them will still provide useful evidence about data needs, degradation and model choice. Progression to a larger study will consider performance across manuscripts, the ability to retain source links and the expert effort needed to correct outputs. Claims of improved research productivity or educational effectiveness will require a later user study.

Platform Access and Human Validation

The proposed database will store manuscript and page identifiers, line coordinates, image rights, raw OCR, verified Jawi, Roman transliteration, model version and review history. The interface will display text beside its source crop and visibly distinguish unreviewed, reviewed and verified records. Search will support both Jawi and Roman queries, with modernised search forms kept separate from the underlying transcription. Demonstrator checks will use a predefined set of passages and queries to determine whether the correct source can be retrieved.

Specialists will adjudicate difficult readings and approve reference material. A later controlled contribution pathway will allow trained community members to suggest corrections to selected, authorised records. Agreement by two contributors may prioritise a suggestion for review, but will not by itself confer verified status. Each accepted correction will retain the earlier version, the reason for change and a link to its source. Corrections from a frozen test set will be recorded for future corpus versions without entering the current training cycle.

Accessibility will be supported through Unicode text, adjustable display, keyboard-accessible review controls and export of permitted verified records. The project will aim to release evaluation scripts, annotation guidance and non-restricted dictionary contributions under appropriate licences after ownership and collection conditions are established. Manuscript images and restricted content will not be presumed openly reusable. Community participation and any evaluation involving users will follow applicable institutional procedures before recruitment. These are planned access and governance features rather than evidence of an existing public service.

DISCUSSION AND LIMITATIONS

The framework connects a technical recognition task to the evidential needs of manuscript scholarship. Its main methodological contribution is the separation of source transcription, transliteration and search-oriented normalisation, each linked to the original image. This allows a researcher to inspect how a machine reading became a verified record and prevents a convenient Roman rendering from obscuring the manuscript's Jawi form. If implemented and validated, these functions could support passage discovery, linguistic comparison and teaching using selected heritage materials.

The most immediate limitation is the absence of a completed pilot and benchmark corpus. The proposed recruitment target may be revised after a rights and annotation audit, and access to multiple independent hands may be limited. Neither the initial hyperparameters nor the internal performance thresholds have been validated for this corpus. The article therefore supports a research plan and grant-reporting framework, not an assertion that JAWI LENS currently achieves the stated accuracy levels.

Further limitations concern annotation uncertainty, representativeness and domain shift. Specialists may disagree over damaged writing, and limiting primary scoring to resolvable lines will underestimate the difficulty of the complete archive unless exclusions are reported. Font-generated synthetic lines may inadequately represent historical handwriting. Arabic transfer may fail to benefit Malay-specific orthography. Pretrained resources may also contain material whose provenance cannot be fully reconstructed; known overlaps will be checked and remaining uncertainty disclosed.

Transliteration introduces an additional interpretive layer. A single Roman rendering may not capture every defensible reading, particularly when vowel representation or historical vocabulary is ambiguous. Reporting a primary reference score alongside a pre-approved-alternative analysis will make this limitation visible. Expert review remains necessary for passages whose interpretation affects names, chronology or substantive scholarly arguments. The pilot will assess whether the proposed workflow supports that review, rather than assuming that automatic output is sufficient.

CONCLUSION

JAWI LENS is proposed as a pilot platform for historical Malay Jawi manuscript recognition, transliteration and heritage access. This article specifies a corpus strategy, two OCR configurations, staged methods for addressing limited data, separate transliteration tests and a provenance-preserving validation workflow. It also defines provisional quantitative milestones and the records needed for transparent grant reporting. The formal pilot has not commenced, and empirical evaluation has not yet been completed; no measured system accuracy is claimed. The next research stage is to secure and annotate the corpus, freeze the protocol, implement the planned comparisons and assess whether verified outputs can support a searchable manuscript demonstrator. These steps will establish the evidence required for a larger study of historical Jawi manuscripts.

ACKNOWLEDGEMENTS

The authors acknowledge the Ministry of Higher Education Malaysia for supporting the related research through the Fundamental Research Grant Scheme, research code FRGS/1/2023/WAB10/UITM/02/1. The authors also acknowledge Universiti Teknologi MARA and the Academy of Language Studies for the institutional context supporting the JAWI LENS research initiative.

Data Availability

No benchmark dataset or empirical results are reported in this article. Future access to manuscript images and verified transcriptions will depend on collection permissions. The project intends to document the evaluation protocol, annotation guidance and releasable research resources when the pilot is implemented.

REFERENCES

1. Garces Arias, E., Pai, V., Schöffel, M., Heumann, C., & Aßenmacher, M. (2023). Automatic transcription of handwritten Old Occitan language. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 15416–15439). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.953
2. Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (pp. 369–376). https://doi.org/10.1145/1143844.1143891
3. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv. https://doi.org/10.48550/arXiv.1512.03385

4. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2023). TrOCR: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 13094–13102. https://doi.org/10.1609/aaai.v37i11.26538
5. Razali, S., Muchtar, K., Rinaldi, M. H., Nurdin, Y., & Rahman, A. (2024). Augmentation of additional Arabic dataset for Jawi writing and classification using deep learning. Jurnal Rekayasa Elektrika, 20(1). https://doi.org/10.17529/jre.v20i1.33722
6. Sakhovskiy, A., Ulitin, I., Bojarskaja, E., Kokh, V., Murtazin, R., Novopoltsev, M., & Budennyy, S. (2026). InkSight: Towards AI-aided historical manuscript analysis. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations) (pp. 271–281). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.eacl-demo.20
7. Shi, B., Bai, X., & Yao, C. (2015). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. arXiv. https://doi.org/10.48550/arXiv.1507.05717