VOL. 24, NO. 3, 2025, 337-343
elektrika.utm.my
ISSN 0128-4428

ELEKTRIKA
Journal of Electrical Engineering

Enhancing Contextual Arabic Handwritten Characters Recognition with CNNs: A Comparative Study on Augmentation Strategies and Dataset Scaling

Fatima Aliyu Shugaba[1], **Usman Ullah Sheikh**[1]*, **Mohd Afzan Othman**[1], **Nurulaqilla Khamis**[1] and **Muhammad Habibullah Abdulfattah**[1]

[1]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310, Johor Bahru, Malaysia

*Corresponding author: usman@fke.utm.my

Abstract: Arabic handwritten character recognition (AHCR) faces significant challenges due to the cursive nature of the script, positional variations of characters, and inconsistencies in existing datasets which hinder robust model training and generalization. Existing AHCR systems often rely on datasets with unreliable annotations, limited character sets, and inconsistent forms, restricting their real-world applicability. This study presents a comparative evaluation of four convolutional neural network (CNN) experiments developed to enhance contextual Arabic handwritten character recognition. Beginning with a baseline model trained on the HMBD dataset, we progressively examine the impact of data augmentation and dataset scaling by correcting mislabeled samples and incorporating additional positional forms, resulting in an extended dataset with 114 contextually diverse classes. While prior work such as the CNN-5 model, reported a 91.96% accuracy on the original HMBD dataset, our final model, trained on an enhanced dataset with semantic and structural improvements, achieved a higher test accuracy of 92.24%, along with precision and F1 scores of 92.48% and 92.24%, respectively, outperforming CNN-5 despite the increased complexity of class structure. The results underscore the importance of both data integrity and model architecture, and offer a robust framework for developing scalable and reliable handwritten Arabic OCR systems. This study benchmarks state-of-the-art CNNs and provides a reproducible pipeline that bridges performance with real-world applicability.

Keywords: Arabic Handwritten Character Recognition, Contextual Forms, Dataset Correction, Semantic Dataset

Article History: received 3 May 2025; accepted 17 June 2025; published 22 December 2025
Digital Object Identifier 10.11113/elektrika.v24n3.739

1. INTRODUCTION

Arabic handwritten character recognition (AHCR) has attracted significant attention in the field of pattern recognition and optical character recognition (OCR) due to the linguistic and structural complexity of the Arabic script. Unlike Latin-based scripts, Arabic is inherently cursive, with characters changing shape based on position in word (initial, medial, final, or isolated), leading to increased challenges in segmentation and classification [1, 2]. Furthermore, many existing datasets suffer from issues such as inconsistent annotations, mislabeled samples, and inadequate representation of all contextual forms, limiting their effectiveness in training robust recognition systems.

Several deep learning-based models have been proposed in recent years to address these challenges, notably convolutional neural networks (CNNs) which have shown exceptional performance in image classification tasks, including handwritten text recognition [3, 4]. However, the performance of such models depends on the network architecture and significantly on the quality and diversity of the training data [5-7]. Such datasets provide a wide range of samples and variations, enabling the models to learn robust patterns and generalize well to new data.

Despite recent advances, current AHCR models often overlook the critical importance of addressing dataset completeness, particularly the inclusion of all positional forms and correction of mislabeled samples. This leads to models that struggle with generalization and practical deployment.

Our contribution mainly lies in the holistic approach that combines semantic enrichment with rigorous label correction and augmentation. Unlike prior studies that focused on model design or augmentation, we emphasize structural completeness by incorporating all positional forms and correcting labeling inconsistencies in the HMBD dataset. This enables the model to better capture contextual character variations, enhancing recognition accuracy and robustness.

In the Arabic domain, many models report high accuracy on datasets like AHCD [8], AIA9K [9], Hijja [4], HACDB [10], and HMBD [11], but these often lack semantic completeness or exhibit data leakage due to overlapping or mislabeled samples [7].

The AHCD, AIA9K, and Hijja datasets focus on

isolated characters (28 Arabic characters) without positional variations. While HACDB offers a distinctive contribution by providing all positional forms of the Arabic characters, including overlapping shapes, providing a valuable resource particularly for segmentation-based recognition tasks, it presents notable limitation that restricts its applicability to real world scenarios. Specifically, it excludes dots and certain marks (like Hamzah) to simplify the classification problem and reduce the number of classes. However, dots are essential features whereas many characters share the same base shape and are only distinguishable by the presence, number, and position of dots (e.g., ب (baa), ت (taa), and ث (thaa)). This simplification reduces the number of classes but also limits its linguistics fidelity and practical utility.

In [11], Balaha et al. introduced the HMBD dataset along with the AHCR-DLS framework, reporting high accuracy by applying extensive augmentation, generating over 865,000 samples from the original dataset. However, the quality and class completeness were not rigorously addressed. While HMBD has served as a cornerstone for Arabic handwritten character recognition studies, it has limitations, including mislabeled samples and incomplete coverage of character forms.

Research on AHCR has evolved significantly over the past decades, driven by the growing need for digital archiving, linguistic analysis, and intelligent document processing. Traditional rule-based and template-matching methods [12-14] laid the groundwork for early AHCR systems but struggled to handle variations in handwriting styles, diacritic placements, and character positioning.

With the emergence of machine learning and deep learning, CNNs have become the dominant architecture due to their ability to learn spatial hierarchies and local features directly from raw input images [15, 16]. Numerous studies have leveraged CNNs to improve AHCR accuracy, such as [17], who explored real-time recognition tools, and [18], who combined CNNs with hybrid optimization strategies.

In this study, we evaluate and compare the performance of CNN-based Arabic character recognition systems across four experimental setups. These include: a baseline model trained on the standard HMBD dataset (105 characters), an augmented variant, a structurally improved version of HMBD with 114 classes (including previously excluded/missing middle and end forms of some character), and a final model combining both augmentation and dataset enhancements. Our approach prioritizes both dataset integrity and structural completeness, with the goal of enhancing recognition performance while maintaining realistic dataset sizes. In addition, this paper compares these models and benchmarks the best result against the proposed CNN-5 model in [19] trained on HMBD with augmentation.

Recent works including Mezghani et al. [20] and Abu Al-Haija [21] have highlighted the importance of semantically rich datasets and class diversity. Additionally, comprehensive reviews by Alhamad et al. [7], Alsurori et al. [22], and Ahmad et al. [23] emphasized dataset structure, contextual form inclusion, and the trade-off between augmentation and dataset realism as critical factors for improving generalizability.

Building on these insights, the present study introduces an enhanced dataset with 114 contextual classes and benchmarks CNN performance across four experimental configurations, highlighting the necessity of structural completeness and accurate annotation.

1.2 Dataset

HMBD is a public dataset introduced by Balaha et al. in [11]. It contains 54,115 black and white image samples of both handwritten Arabic characters (105), and digits (10), making it 115 classes, written by 125 volunteers. All images have 300 x 300 dimensions and are sorted into individual folders corresponding to 115 total classes. It can be found at the following address: https://github.com/hossambalaha/hmbd-v1.

2. METHODOLOGY

This section describes the data preprocessing, CNN architecture, training strategy, and evaluation metrics employed in all four experiments. The implementation leverages Keras and TensorFlow libraries and is guided by established deep learning practices in handwritten character recognition.

2.1 Data Preprocessing and Augmentation

In this study, we focus solely on Arabic characters. Therefore, after isolating folders of characters only, the number of samples amount to 49,101 images. Upon manual inspection, we identified missing positional forms (particularly middle and end forms) for certain characters. We addressed these omissions and expanded the 105-character classes to 114, ensuring complete contextual representation. We also observed and corrected mislabeled samples, cleaned out unclear and corrupted images, and collected additional samples to expand the overall classes. The extended HMBD now contains 57,317 images of handwritten Arabic characters only, sorted into individual folders labeled with the Unicode representation of Arabic character corresponding to each class. Finally, all images were standardized to 64×64 pixel resolution, converted to grayscale, and normalized. The dataset was split into: 80% for training, 10% for validation and 10% for testing.

For experiments using augmentation (Exp 2 & 4), transformations such as rotation (±10 degrees), width/height shift (10%), shear (0.1), and zoom (10%) were applied on the training data using image data generator, to simulate real-world handwriting variability and enhance the model's generalization.

2.2 CNN Architecture

The model architecture follows a deep sequential CNN design with three convolutional layers, batch normalization, ReLU activation, max-pooling, and fully connected dense layers. The network structure is summarized in Table 1.

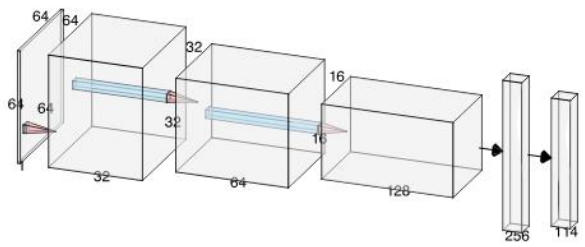

Figure 1. Baseline CNN Architecture for AHCR

The CNN model comprises:

- **Input Layer**: (64×64×1)
- **Conv2D (32 filters, 3x3 kernel)** → BatchNorm → ReLU → MaxPooling (2x2)
- **Conv2D (64 filters, 3x3 kernel)** → BatchNorm → ReLU → MaxPooling (2x2)
- **Conv2D (128 filters, 3x3 kernel)** → BatchNorm → ReLU → MaxPooling (2x2)
- **Flatten → Dense (256 units) → Dropout (0.5)**
- **Dense (#Classes: 105 or 114)** → Softmax activation

Convolutional and pooling output shapes are computed by:

$$O_{conv} = \left(\frac{I - k + 2p}{s}\right) + 1 \quad O_{pool} = \left(\frac{I - ps}{s}\right) + 1 \quad (1)$$

Where:

- I: input size
- k: kernel size
- p: padding
- s: stride
- ps: pooling size

Final classification uses softmax:

$$P(y = j|x) = \frac{e^{zj}}{\sum_{k=1}^{K} e^{zk}} \quad (2)$$

With cross-entropy loss for optimization:

$$\mathcal{L}\ (y,\ \hat{y}) = -\sum_{j=1}^{K} y_j \log\ (P(y = j|x)) \quad (3)$$

Table 1. Model Summary

Layer Type	Output Shape	Parameters
Conv2D (32)	(64, 64, 32)	320
MaxPooling2D	(32, 32, 32)	0
Conv2D (64)	(32, 32, 64)	18,496
MaxPooling2D	(16, 16, 64)	0
Conv2D (128)	(16, 16, 128)	73,856
MaxPooling2D	(8, 8, 128)	0
Flatten	(8192)	0
Dense (256)	(256)	2,097,408
Dropout (0.5)	(256)	0
Output Dense	(105)	26,985

2.3 Training Configuration

In order to optimize training stability and ensure effective convergence, we adopted a training setup designed specifically for the complexities of AHCR. The Adam optimizer, with its default parameters (β_1=0.9, β_2=0.999), was selected due to its adaptive learning capabilities and robustness in managing non-stationary objectives, which are typical in deep learning tasks involving diverse handwriting styles. Categorical cross entropy was used as the loss function, consistent with the multi-class classification objective required for 114 distinct Arabic character classes.

A learning rate of 0.001 was initialized to balance learning speed and minimize the risk of overshooting minima. To further refine convergence, we employed the 'ReduceLROnPlateau' function, allowing the learning rate to decrease automatically when validation loss stagnated. This mechanism helps the model escape potential plateaus and facilitates more precise adjustments during training. A batch size of 32 was chosen to maintain a balance between computational efficiency and learning stability, particularly given the variability in character structure and handwriting quality.

The training was limited to a maximum of 60 epochs, but early stopping with a patience of 10 epochs was included to stop training once no significant improvement in validation accuracy was detected, thereby preventing unnecessary use of resources as well as overfitting. Furthermore, model checkpoint was utilized to ensure that the best-performing model, based on validation metrics, was preserved. Altogether, this setup was aimed at achieving both fast convergence and strong generalization, which are critical for building a reliable handwritten Arabic character recognition system.

2.4 Overview of Experimental Setup

The experiments were structured to isolate the individual and combined effects of data augmentation and dataset extension. Each experiment adhered to a consistent CNN architecture and training pipeline for a fair evaluation.

2.5 Evaluation Metrics

For the model evaluation we used accuracy, precision, recall, and F1 score as in the following equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN + \mathcal{E}} \quad (4)$$

$$Precision = \frac{TP}{TP + FP + \mathcal{E}} \quad (5)$$

$$Recall = \frac{TP}{TP + FP + \mathcal{E}} \quad (6)$$

$$F1\ Score = 2.\frac{Precision\ .Recall}{Precision + Recall} \quad (7)$$

Where $\mathcal{E}$ is a small constant to prevent division by zero, *TP* (True Positive) denotes number of samples correctly predicted as belonging to a given class, *TN* (True Negative) denotes number of samples correctly predicted as not belonging to a given class, *FP* (False Positive) denotes number of samples incorrectly predicted as belonging to a given class, and *FN* (False Negative) denotes number of samples belonging to a class but incorrectly predicted as another class.

3. RESULTS AND DISCUSSION

The results of our four experimental configurations are analyzed in this section to determine the influence of dataset augmentation and class expansion on the recognition of handwritten Arabic characters. Using a consistent CNN architecture and training strategy, we observed clear performance distinctions across the configurations. A summary table and detailed evaluation is presented to support the conclusions drawn from these experiments in Table 2 and Table 3.

3.1 Quantitative Results Summary

Experiment 1, which served as the baseline using the standard 105-class HMBD dataset without augmentation, revealed the model's limited generalization capacity despite high training accuracy. Augmenting this same dataset (Experiment 2) significantly improved test performance, confirming the value of data variability. Extending the class structure to 114 without augmentation (Experiment 3) showed that cleaning the data alone can provide strong gains. Ultimately, experiment 4, which integrated both structural enrichment and augmentation, achieved the best results, with a test accuracy of 92.24% and strong macro precision, recall, and F1 scores.

3.2 Confusion Matrix and Per-Class Evaluation

Analysis of confusion matrices on the test data showed high confidence in recognizing common characters such as 'م', 'ا', and 'ن'. The confusion matrices for Experiment 2 and Experiment 4 both demonstrate strong overall classification performance, with a clear concentration of correct predictions along the diagonal. However, a closer comparison between the two shows that Experiment 4's confusion matrix exhibits a cleaner, sharper diagonal with fewer scattered off-diagonal errors.

This improvement suggests that while data augmentation in Experiment 2 helped the model generalize better compared to the baseline, it was not enough to eliminate certain ambiguities caused by unclear character forms and occasional mislabeled samples in the original dataset.

The presence of mislabeled samples in the original HMBD dataset, as visually demonstrated in Figure 4, introduces noise that misguides model training and evaluation. Without addressing these errors, any reported accuracy becomes an unreliable measure of true performance. Our work rectifies this by manually inspecting and correcting such mislabels, thereby enabling more trustworthy model evaluation and comparison.

In Experiment 2, although the model successfully classified the majority of samples, some confusion remained between visually similar characters, particularly those differentiated by small diacritics or minor structural differences, such as (ب) and (ت), or (خ) and (ح).

Experiment 4, by contrast, benefited not only from augmentation but also from a corrected and structurally complete dataset. The addition of missing contextual forms gave the model a broader and more representative understanding of character variations. The confusion matrix showed a noticeable reduction in misclassification areas, and the diagonal becomes more dominant, indicating that the model is better at distinguishing between closely related classes. These improvements confirm that context-aware dataset construction plays a critical role in advancing Arabic handwritten character recognition.

Overall, this comparison underlines that, while augmentation can extend the variability of the training data, true generalization and high precision come from pairing it with a clean, semantically rich dataset. The enhancements seen in Experiment 4 highlight the critical role of dataset quality in achieving dependable handwritten Arabic character recognition results.

3.3 Generalization and Robustness

The best model (Experiment 4) reported test loss of 0.2861 and test accuracy of 92.24%. the precision, recall, and F1-score are 92.48%, 92.24%, and 92.24% respectively. These results affirm the critical role of dataset structure, balance, and contextual representation in developing accurate, reliable Arabic OCR systems.

3.4 Summary and Comparative Conclusion

The comparative evaluation across the four experiments highlights the cumulative impact of data augmentation and dataset scaling on Arabic handwritten character recognition. Our final configuration (Experiment 4), which integrated augmentation with a structurally enriched 114-class dataset, achieved a test accuracy of 92.24%, alongside precision and F1 scores of 92.48% and 92.24%, respectively.

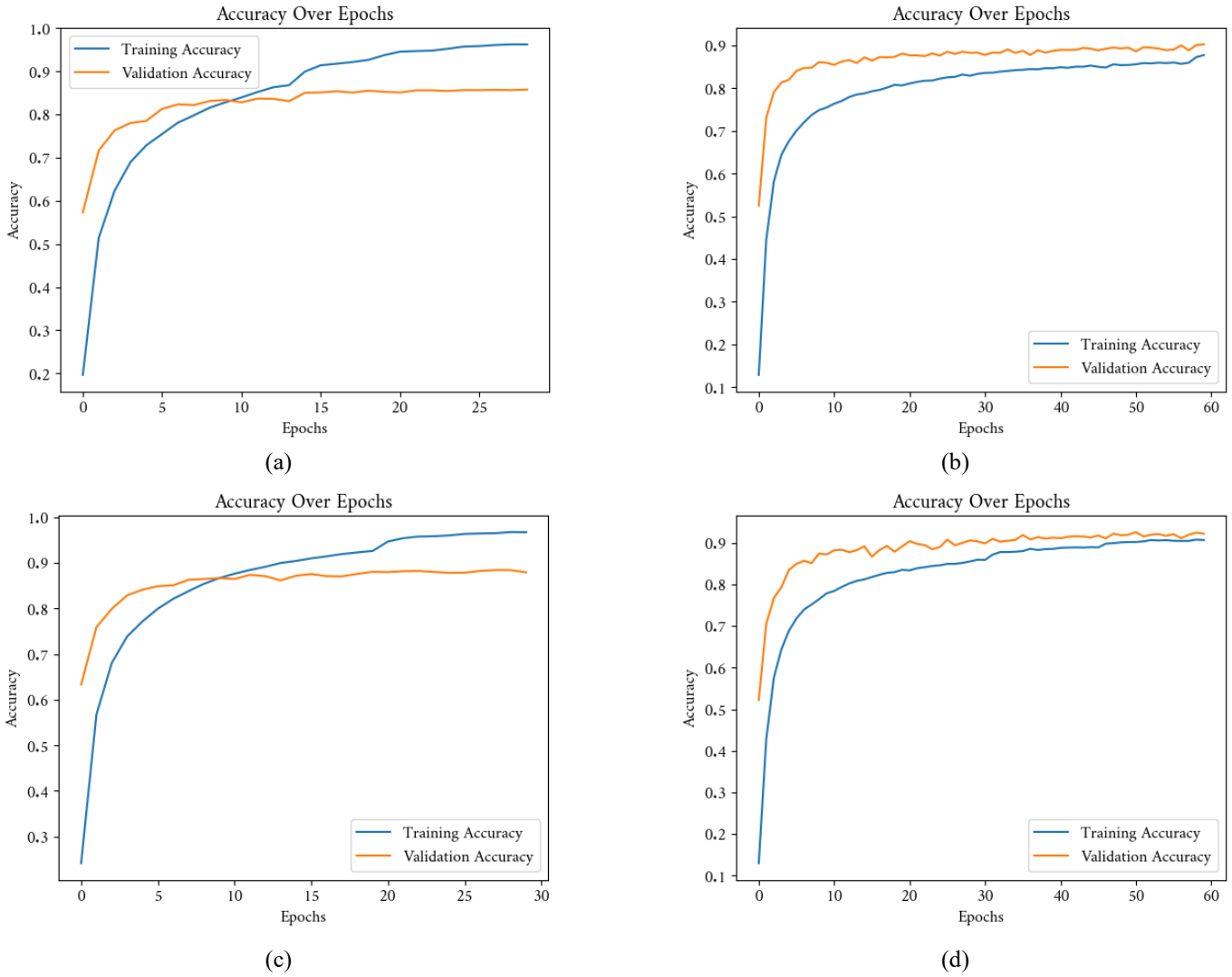

Figure 2. Accuracy for 60 epochs for experiments 1-4 (a-d)

Table 2. Observations and Insights on the Training Results

Experiment	Training Accuracy (%)	Validation Accuracy (%)	Observation
1	96.22	85.77	Suffered from overfitting. Did not generalize well
2	87.74	90.26	Data augmentation significantly improved generalization
3	96.72	87.93	Expansion boosted performance even without augmentation
4	90.68	**92.20**	Achieved the best results. Augmentation along with clean and structural integrity yields the most robust model

Table 3. Performance Metrics on Testing Data

Experiment	Dataset Classes	Augmentation	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Exp 1	105	No	82.43	82.95	82.43	82.43
Exp 2	105	Yes	89.72	89.90	89.72	89.71
Exp 3	114	No	87.45	87.82	87.45	87.45
Exp 4 (Final)	114	Yes	**92.24**	**92.48**	**92.24**	**92.24**

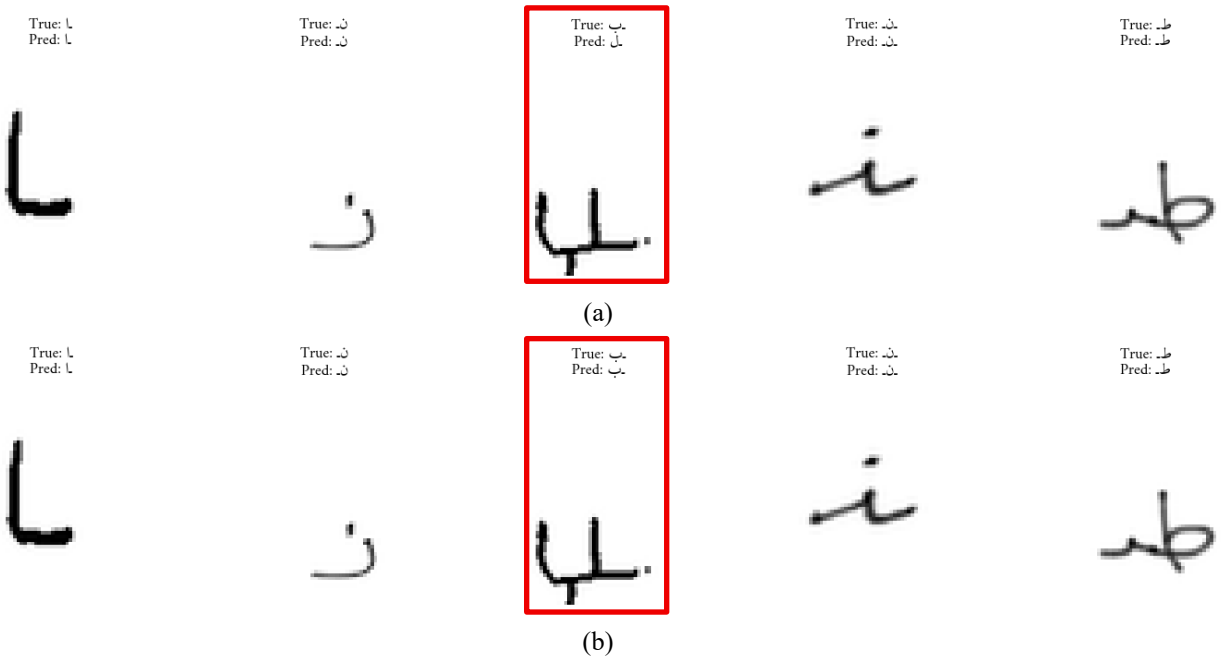

Figure 3. Visualization of Test Predictions for experiments 3 and 4 (a and b)

Figure 4. Mislabelled Image from the HMBD dataset.

This performance underscores a strong generalization capability and robust recognition of Arabic characters across contextual forms.

Compared to the benchmark study by Balaha et al. [19], which reported 91.96% accuracy using the CNN-5 model trained on the original version of the HMBD dataset (54,115 samples of characters and Arabic digits), our approach offers a more semantically complete and practically deployable alternative. Despite using significantly higher classes, our model outperforms the benchmark in terms of dataset fidelity and form diversity, reflecting improved applicability for real-world OCR scenarios where quality and semantic coverage matter as much as quantity.

4. CONCLUSION

This study systematically evaluated the effects of data augmentation and dataset expansion on the performance of CNNs for Arabic handwritten character recognition. Four experimental configurations were analyzed, each isolating the variables of dataset structure and augmentation to assess their influence on model performance. Results demonstrated that both augmentation and contextual form inclusion are essential for building accurate and generalizable recognition systems.

Our results emphasize that:

- Generalization benefits more from semantic completeness than dataset size.
- Contextual forms are essential for realistic Arabic character recognition.
- Combining moderate augmentation with dataset correction yields more robust models.

Thus, dataset curation and augmentation must go hand-in-hand for advancing AHCR systems.

This work presents a comprehensive comparative study on enhancing Arabic handwritten character recognition. By restructuring and cleaning the HMBD dataset, expanding character classes to include all positional forms, and applying strategic augmentation, we achieve higher test accuracy (92.24%) compared to CNN-5. Our findings stress the importance of dataset integrity alongside network design, suggesting future work focus on further semantic enrichment and domain-specific data augmentation.

In future work, we intend to explore hybrid architectures such as CNN-LSTM and Transformer-based models to further improve the sequence modeling of Arabic scripts. We also plan to extend the dataset to cover full word-level recognition with diacritics and incorporate attention-based mechanisms for form-aware feature weighting. Finally, we aim to release the corrected and extended 114-class dataset to support reproducibility and further advancements in Arabic handwritten text recognition.

ACKNOWLEDGMENT

The authors extend their appreciation to Universiti Teknologi Malaysia (UTM) for providing the essential support and resources that made this research possible.

REFERENCES

[1] M. T. Parvez, "Arabic handwritten text recognition using structural and syntactic pattern attributes," Ph.D., King Fahd University of Petroleum and Minerals (Saudi Arabia), Saudi Arabia, 2010.

[2] M. S. K]horsheed, "Off-Line Arabic Character Recognition – A Review," *Pattern Analysis & Applications,* vol. 5, no. 1, pp. 31-45, 2002/05/01 2002, doi: 10.1007/s100440200004.

[3] H. M. Balaha, H. A. Ali, and M. Badawy, "Automatic recognition of handwritten Arabic characters: a comprehensive review," *Neural Computing and Applications,* vol. 33, pp. 3011-3034, 2021.

[4] N. Altwaijry and I. Al-Turaiki, "Arabic handwriting recognition system using convolutional neural network," *Neural Computing and Applications,* vol. 33, no. 7, pp. 2249-2261, 2020.

[5] N. Alrobah and S. Albahli, "Arabic handwritten recognition using deep learning: A Survey," *Arabian Journal for Science and Engineering,* vol. 47, no. 8, pp. 9943-9963, 2022.

[6] R. Ahmed *et al.*, "Offline arabic handwriting recognition using deep machine learning: A review of recent advances," in *Advances in Brain Inspired Cognitive Systems: 10th International Conference, BICS 2019, Guangzhou, China, July 13–14, 2019, Proceedings 10*, 2020: Springer, pp. 457-468.

[7] H. A. Alhamad *et al.*, "Handwritten recognition techniques: a comprehensive review," *Symmetry,* vol. 16, no. 6, p. 681, 2024.

[8] A. El-Sawy, M. Loey, and H. El-Bakry, "Arabic handwritten characters recognition using convolutional neural network," *WSEAS Transactions on Computer Research,* vol. 5, no. 1, pp. 11-19, 2017.

[9] M. Torki, M. E. Hussein, A. Elsallamy, M. Fayyaz, and S. Yaser, "Window-based descriptors for Arabic handwritten alphabet recognition: a comparative study on a novel dataset," *arXiv preprint arXiv:1411.3519,* 2014.

[10] A. Lawgali, M. Angelova, and A. Bouridane, "HACDB: Handwritten Arabic characters database for automatic character recognition," in *European workshop on visual information processing (EUVIP)*, 2013: IEEE, pp. 255-259.

[11] H. M. Balaha, H. A. Ali, M. Saraya, and M. Badawy, "A new Arabic handwritten character recognition deep learning system (AHCR-DLS)," *Neural Computing and Applications,* vol. 33, pp. 6325-6367, 2021.

[12] A. Cheung, M. Bennamoun, and N. W. Bergmann, "An Arabic optical character recognition system using recognition-based segmentation," *Pattern recognition,* vol. 34, no. 2, pp. 215-233, 2001.

[13] S. Alma'adeed, C. Higgins, and D. Elliman, "Off-line recognition of handwritten Arabic words using multiple hidden Markov models," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2004: Springer, pp. 33-40.

[14] H. S. Beigi, "An overview of handwriting recognition," in *Proceedings of the 1st annual conference on technological advancements in developing countries, Columbia University, New York*, 1993, pp. 30-46.

[15] R. Ahmed *et al.*, "Novel deep convolutional neural network-based contextual recognition of Arabic handwritten scripts," *Entropy,* vol. 23, no. 3, p. 340, 2021.

[16] N. Alrobah and S. Albahli, "A hybrid deep model for recognizing arabic handwritten characters," *IEEE Access,* vol. 9, pp. 87058-87069, 2021.

[17] M. A. Alzubaidi, M. Otoom, and N. S. Ahmad, "Real-time assistive reader pen for Arabic language," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP),* vol. 20, no. 1, pp. 1-30, 2021.

[18] A. T. Sahlol, M. Abd Elaziz, M. A. Al-Qaness, and S. Kim, "Handwritten Arabic optical character recognition approach based on hybrid whale optimization algorithm with neighborhood rough set," *IEEE Access,* vol. 8, pp. 23011-23021, 2020, doi: 10.1109/ACCESS.2020.2970438.

[19] H. M. Balaha *et al.*, "Recognizing arabic handwritten characters using deep learning and genetic algorithms," *Multimedia Tools and Applications,* vol. 80, pp. 32473-32509, 2021.

[20] A. Mezghani, S. Kanoun, M. Khemakhem, and H. El Abed, "A database for arabic handwritten text image recognition and writer identification," in *2012 international conference on frontiers in handwriting recognition*, 2012: IEEE, pp. 399-402.

[21] Q. A. Al-Haija, "Leveraging ShuffleNet transfer learning to enhance handwritten character recognition," *Gene Expression Patterns,* vol. 45, p. 119263, 2022.

[22] M. H. Alsurori, A. A. Mohsen, G. Al-Badani, M. Al-Nahari, T. Faisal, and S. Alkasem, "Review on Arabic Handwritten Recognition Using Deep Learning and Machine Learning," in *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, 2023: IEEE, pp. 1-8.

[23] A. T. Al-Taani and S. T. Ahmad, "Recognition of Arabic handwritten characters using residual neural networks," *Jordanian Journal of Computers and Information Technology,* vol. 7, no. 2, 2021.