AI-Based Sindhi Handwritten Alphabets Classification with Web-Based Development

Mudasir Murtaza, Farhad Ali, Muhammad Taha
Department of Computer Science Quaid-e-Awam University of Engineering, Science and Technology Nawabshah, Pakistan
*Correspondence: mudasirmurtaza20@gmail.com, farhadalishaikh06@gmail.com, muhammadtahafarooque@gmail.com

Citation | Murtaza. M, Ali. F, Taha. M, "AI-Based Sindhi Handwritten Alphabets Classification with Web-Based Development", IJIST, Vol. 07 Special Issue. pp 187-195, May 2025.
Received | April 21, 2025 **Revised |** May 18, 2025 **Accepted |** May 20, 2025 **Published |** May 22, 2025.

Handwriting recognition has advanced significantly for many widely-used scripts, but low-resource languages like Sindhi have received little attention. In this study, we propose the development of a powerful AI model designed to classify handwritten Sindhi alphabets. To address challenges such as varied handwriting styles and the lack of publicly available datasets, the model leverages a manually curated, diverse dataset, advanced CNN architectures, and data augmentation techniques. To encourage further research, the dataset will be made available publicly in two versions: raw and augmented.

The main contributions of this study include achieving approximately 93% training accuracy, 96% validation accuracy, and a loss rate of less than 1%. Additionally, we have created valuable open-source datasets for Sindhi handwriting recognition. Although a web-based application is planned for future development, these accomplishments lay a solid foundation for digitizing Sindhi texts, creating educational tools, and helping preserve the Sindhi language.

Keywords— Sindhi, Handwriting Recognition, Convolutional Neural Networks, Data Augmentation, Web-Based Application, Machine Learning, Open-Source.

Introduction:

Handwriting recognition is a key area of research in pattern recognition and machine learning, with applications in digitization, education, and accessibility. While significant progress has been made in recognizing scripts like English, Arabic, and Chinese, less attention has been given to Sindhi, mainly due to the lack of large-scale datasets and specific recognition models. Although Sindhi is spoken by millions, its complex script, with intricate ligatures and structural similarities to Arabic and Urdu, poses major challenges for automated text recognition.

Most existing handwriting recognition research has focused on high-resource languages, leaving Sindhi technology underdeveloped. The primary obstacles include the absence of comprehensive annotated datasets and the limited studies addressing the unique characteristics of the Sindhi script. This research seeks to bridge this gap by developing a robust dataset of handwritten Sindhi alphabets and making it openly available to the research community. We also implement a CNN-based model to achieve accurate classification. The final goal is to deploy this model as a web-based application, enabling real-time classification to support document digitization and language education.

Objectives:

This research sets out to close the resource disparity in Sindhi handwriting recognition by:

- Establishing a strong, varied dataset of handwritten Sindhi alphabets.
- Conceiving and testing a CNN-based classification model.
- Providing foundations for an Internet-based application with real-time recognition capability.
- Releasing the dataset openly for use in future academic and applied research.

1.2 Novelty and Contributions:

This work makes the following novel contributions to the area:

- A high-quality, manually curated open-source dataset for the entire Sindhi script—something previously lacking.
- Use of state-of-the-art CNN models on a low-resource script with high accuracy.
- Employment of aggressive preprocessing and data augmentation methods customized to the complexities of handwriting in Sindhi.
- Construction of a system for real-world application through web-based deployment, facilitating wider accessibility and usefulness.

Related Work and Literature Review:

Convolutional Neural Networks (CNNs) have become the most popular method for handwriting recognition due to their excellent feature extraction and classification capabilities. Pioneering designs like LeNet have shown high accuracy in character recognition for scripts such as Arabic and Pashto, and these have been followed by more advanced models like AlexNet, VGG, and ResNet [1][2]. Additionally, it has been shown that transfer learning, which leverages pre-trained models, significantly enhances performance, especially in scenarios with limited data [3].

Challenges in Low-Resource Languages:

Languages like Sindhi and Urdu face challenges such as limited datasets and varied handwriting styles. Previous research has tackled these issues by creating custom datasets and applying data augmentation techniques [4][5]. While some progress has been made in Sindhi recognition, particularly for numeral recognition and limited character sets [6][7], comprehensive solutions for the entire Sindhi alphabet are still largely unexplored.

Preprocessing and Data Augmentation:

Effective preprocessing, including grayscale conversion, noise reduction through Gaussian filtering, and normalization, is essential for standardizing input data [8]. Data

augmentation techniques such as rotation, scaling, and flipping help artificially expand the dataset, improving the model's robustness [9]. For Sindhi script, which contains many dots, diacritics, and tightly spaced strokes, preprocessing operations are very important. Grayscale conversion reduces image data without compromising the contrast between ink and background. Gaussian filtering reduces background noise, improving stroke clarity. Normalization facilitates consistent input, which is very necessary because handwritten Sindhi characters contain a high intra-class variance.

Gaps and Opportunities:

Although CNN-based recognition has made significant progress, Sindhi handwriting recognition has not yet benefited from these advancements due to limited research. There is an opportunity to use modern CNN architectures and transfer learning techniques to develop an effective system for Sindhi character classification. This work also explores integrating the model into a web-based application for real-time interaction, expanding its practical use.

Proposed Methodology:

The proposed methodology is divided into several phases to ensure the systematic development, evaluation, and deployment of the system.

Dataset Creation:

- **Data Collection**: Handwritten samples of Sindhi alphabets are collected from various individuals with different handwriting styles to ensure a diverse dataset. The dataset contains samples that were gathered in a controlled setting to optimize consistency and clarity, as seen in Figure 1. All of the Sindhi alphabets are represented in the final dataset. An overview of the digital alphabet images used in this study is given in Figure 2.
- **Bias and Representation Consideration:** When obtaining handwriting samples, we tried to maintain demographic balance by representing the various age groups, genders, and educational levels. Nevertheless, there would still be some sampling bias originating from regional availability and volunteer participation. This is a limitation we accept and look forward to enhancing demographic representation in subsequent releases of the dataset.
- **Digitization**: The samples are scanned or photographed and converted into high-quality digital images.
- **Labeling**: Each image is carefully labeled with its corresponding Sindhi alphabet to create a well-structured dataset.
- **Open-Sourcing**: The raw and augmented datasets will be made publicly available to promote further research in Sindhi handwriting recognition.

Data Preprocessing:

- **Image Processing**: Images are converted to grayscale and resized to a fixed size (e.g., 64×64 pixels).
- **Normalization**: Pixel values are normalized to a 0–1 range.
- **Noise Reduction**: Techniques such as Gaussian filtering or thresholding are used to remove noise from the images.

Model Design and Training:

- **CNN Architecture**: A CNN-based model is implemented, starting with a basic architecture (e.g., LeNet) and extending to more advanced models (e.g., VGG, ResNet).
- **Training Strategy**: 10% was set aside for testing, 20% for validation, and 70% for training. This widely used split maintains strong validation and testing sets while guaranteeing adequate training data. In light of the relatively small dataset size and the significance of avoiding overfitting, this distribution balances the evaluation of generalization and model learning. Optimizers like Adam and loss functions such as categorical cross-entropy are used.
- **Model Evaluation**: Performance is monitored using metrics like accuracy, precision, recall, and F1 score on the validation set.

Performance Optimization:

- **Data Augmentation**: Techniques like rotation, scaling, and flipping are applied to enhance the dataset's diversity.
- **Transfer Learning**: Pretrained models are used to reduce training time and improve accuracy.
- **Comparative Analysis**: Different model architectures and hyperparameters are tested to find the best-performing model.

Deployment:

- **Web-Based Application**: A user-friendly interface is created using Flask or Streamlit, allowing users to upload handwritten samples for real-time classification.
- **Open-Source Release**: Both the raw and augmented datasets, along with the trained model, will be made available for researchers.

To ensure clarity, each step in the pipeline from data collection to deployment is outlined. Figure 3 illustrates the overall methodology used for the system's design and development.

Figure 1. Dataset Collection Sample

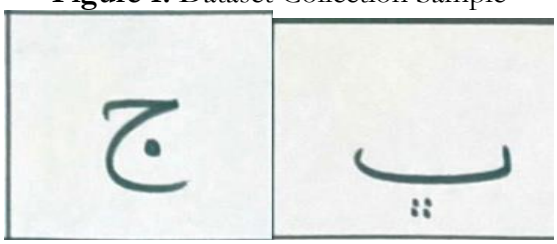

Figure 2. Alphabet Images

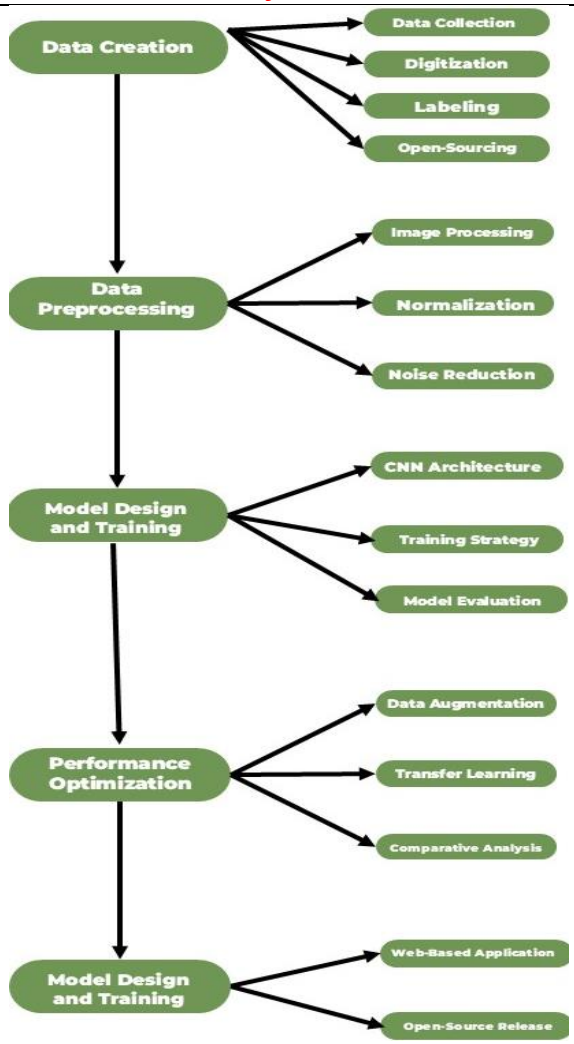

Figure 3. Methodology Flowchart

Potential Challenges and Mitigation Strategies:

Challenge	Mitigation Strategy
Dataset Diversity and Size	Collect data from diverse age groups, genders, and writing styles to ensure a comprehensive dataset.
Noisy or Inconsistent Data	Employ preprocessing techniques (e.g., Gaussian filtering, thresholding) to standardize and clean the images.
Similar-Looking Alphabets	Utilize advanced CNN architectures (e.g., ResNet) or attention mechanisms to capture subtle differences between characters.
Limited Dataset Size	Apply extensive data augmentation techniques (rotation, flipping, scaling) to artificially expand the dataset.
High Computational Requirements	Leverage cloud computing platforms (e.g., Google Colab, AWS) to access GPUs/TPUs for faster training and optimization.

Results:

In the results section, we present the key resources developed for this research, followed by an evaluation of the performance of our CNN model.

Open-Source Handwritten Sindhi Dataset:

Two datasets were developed and made publicly available to facilitate future research. The Raw Dataset contains unprocessed handwritten samples, whereas the Augmented Dataset comprises images altered through rotation, flipping, and scaling for enhanced generalizability. These can be accessed on Kaggle for public use.

- **Raw Dataset** (Contains original scanned images of handwritten Sindhi alphabets.): [https://www.kaggle.com/datasets/mudasirmurtaza/sindhi-alphabets]
- **Augmented Dataset** (Includes images processed and expanded using techniques such as rotation, flipping, and scaling.): [https://www.kaggle.com/datasets/mudasirmurtaza/sindhi-handwritten-alphabets-augmented-dataset]

CNN Model Performance:

The CNN model was trained using the augmented dataset. It achieved:

- Training Accuracy: ~93%
- Validation Accuracy: ~96%
- Validation Loss: < 1.0

5.2. Explanation of Graphs:

Figure 4 shows the trend of accuracy during training. Training and validation accuracy grew steadily with every epoch, ultimately reaching equilibrium. Most importantly, the validation accuracy was higher than training accuracy, indicating good generalization and no overfitting, which is a welcome result considering the complexity of the Sindhi script. Figure 5 plots the curves of training and validation loss. The loss reduced steadily for training and validation sets. The proximity of these curves suggests that the model was learning well and not underfitting or overfitting. The last loss plateaued below 1.0, further establishing the model's resilience.

Observation:

The higher validation accuracy over training accuracy points to strong generalization, which is uncommon in deep learning models and reflects the quality of preprocessing and dataset diversity. These findings illustrate that preprocessing methods (grayscale, filtering) and augmentation strategies made a significant impact on model performance. This is particularly promising for such a complex script as Sindhi, which contains numerous visually similar characters. The high accuracy of validation assures the model's applicability for real-world use, especially in educational or digitization software for the Sindhi language.

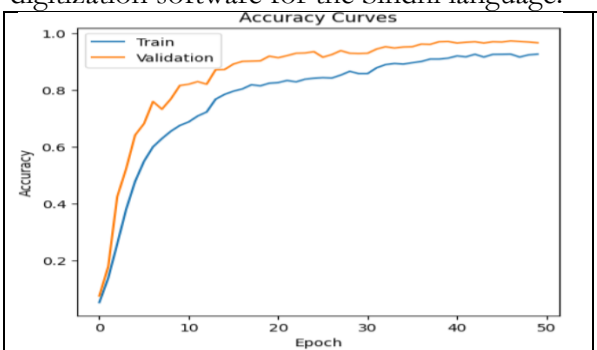

Figure 4. Training and Validation Accuracy Graph

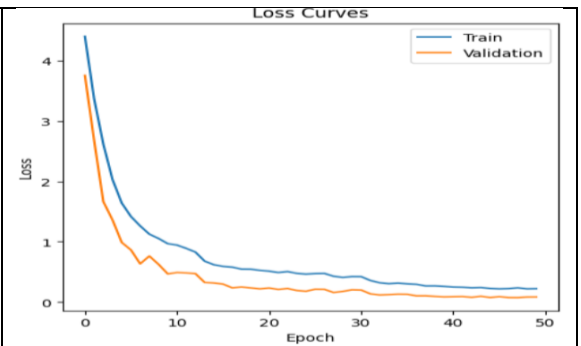

Figure 5. Training and Validation Loss Graph

Discussion:

Although it is not possible to make direct comparisons with the existing literature because of the novelty of our research on the complete Sindhi alphabet, it is useful to place our findings within the larger body of handwriting recognition in low-resource languages.

Earlier research on handwritten Sindhi character recognition, e.g., by Chandio and Leghari [1], was on isolated characters or digits and attained accuracies of about 90%. Likewise, Pashto handwritten recognition with CNNs [2] attained an accuracy of 93%, whereas Urdu character classification research has achieved between 88% to 94% based on dataset quality [10],[4].

In contrast to these, our model attained a validation accuracy of ~96%, although it was tackling the entire alphabet set of 52 characters and dealing with high intra-class variability. This suggests that our data-oriented strategy, coupled with efficient preprocessing and augmentation, played a major role in better recognition accuracy.

In addition, as opposed to prior art that tends to emphasize static offline models, our work is intended for real-time use as part of a web-based application. This field-worthy deployment aspect further aligns our work and contrasts it with education and cultural preservation use cases in real life.

Future Work:

In the future work section, we outline our future directions, focusing on dataset expansion, model optimization, and system integration.

Dataset Expansion:

We plan to collect more samples and continuously update the dataset to create a more diverse resource, supporting further research and improvements in Sindhi language recognition.

Model Optimization and Enhancement:

We will continue enhancing the CNN model to improve accuracy and generalization, exploring advanced techniques such as hyperparameter tuning and architecture modifications.

Model Evaluation and Performance Metrics:

The model will be evaluated using accuracy, F1 score, recall, precision, and confusion matrices to assess how well the trained model classifies the handwritten alphabet.

Integration into Web Application:

To ensure easy accessibility and enable real-time classification and recognition, the trained model will be integrated into a web-based application.

Conclusion:

This paper presents a systematic approach for Sindhi handwritten letter recognition by generating and making available two distinct datasets. Using state-of-the-art CNN architectures and preprocessing techniques, our model achieved approximately 93% training accuracy and 96% validation accuracy, with the loss dropping below 1%. These results highlight the model's strong generalization, even though Sindhi is a complex script.

Moving forward, we will continue improving the CNN model to enhance accuracy and generalization. Additionally, the open-source availability of the dataset encourages further research and facilitates the development of better recognition methods for handwriting classification in low-resource languages.

Finally, we plan to integrate the improved model into a fully operational web-based application, enabling real-time classification and recognition.

References:

[1] A. A. Chandio and M. Leghari, "Deep learning-based isolated handwritten Sindhi character recognition," *Indian J. Technol.*, vol. 6, no. 2, pp. 12–19, 2020, [Online]. Available: https://www.researchgate.net/publication/343170746_Deep_learning-based_isolated_handwritten_Sindhi_character_recognition

[2] H. A. Muhammad Sadiq Amin, Siddiqui Muhammad Yasir, "Recognition of Pashto Handwritten Characters Based on Deep Learning," *Sensors*, vol. 20, no. 20, p. 5884, 2020, doi: https://doi.org/10.3390/s20205884.

[3] Q. U. A. Akram and S. Hussain, "Improving Urdu Recognition Using Character-Based Artistic Features of Nastalique Calligraphy," *IEEE Access*, vol. 7, pp. 8495–8507, 2019, doi: 10.1109/ACCESS.2018.2887103.

[4] S. K. & J. J. P. C. R. Syed Yasser Arafat, Nabeel Ashraf, Muhammad Javed Iqbal, Iftikhar Ahmad, "Urdu signboard detection and recognition using deep learning," *Multimed. Tools Appl.*, vol. 81, pp. 11965–11987, 2022, doi: https://doi.org/10.1007/s11042-020-10175-

2.
[5] A. A. Sanjrani *et al.*, "Extended framework for Sindhi numerals OCR using gradient orientation histograms," *J. Intell. Fuzzy Syst.*, vol. 43, no. 2, pp. 2045–2056, 2022, doi: 10.3233/JIFS-219304/ASSET/A3AA616C-998F-4BB9-96EA-7C931A6C1491/ASSETS/GRAPHIC/10.3233_JIFS-219304-IMG2.JPG.
[6] J. Baber et al, "Urdu handwritten character recognition using deep learning," *J. Inf. Commun. Technol. Res.*, vol. 3, no. 1, pp. 67–74, 2020.
[7] M. K. S. A. Naveed, Ahmed Soomro, Leezna Saleem, "OHSCR: Benchmarks dataset for offline handwritten Sindhi character recognition," *Sir Syed Univ. J. Res.*, vol. 4, no. 1, pp. 11–20, 2024, doi: 10.33317/ssurj.618.
[8] A. H. J. Asghar Ali Chandio, Mehwish Leghari, Mehjabeen Leghari, "Multi-Font and Multi-Size Printed Sindhi Character Recognition using Convolutional Neural Networks," *Pakistan J. Eng. Appl. Sci.*, vol. 25, 2019, [Online]. Available: https://journal.uet.edu.pk/ojs_old/index.php/pjeas/article/view/1635
[9] S. H. Fazli Khaliq, Muhammad Shabir, Inayat Khan, Shafiq Ahmad, Muhammad Usman, Muhammad Zubair, "Pashto Handwritten Invariant Character Trajectory Prediction Using a Customized Deep Learning Technique," *Sensors*, vol. 23, no. 13, p. 6060, 2023, doi: https://doi.org/10.3390/s23136060.
[10] M. R. B. Sayma Shafeeque A. W. Siddiqui, Rajashri G. Kanke, Ramnath M. Gaikwad, "Review on Isolated Urdu Character Recognition: Offline Handwritten Approach," *Ijraset J. Res. Appl. Sci. Eng. Technol.*, 2023, doi: https://doi.org/10.22214/ijraset.2023.55164.