Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

BEYOND BENCHMARK ACCURACY: EVALUATING WRITER DEPENDENT DOMAIN SHIFT IN HANDWRITTEN 0-999 DIGIT RECOGNITION

Alina Fatima[*1], Samina Muqeem[2], Nisha Irshad[3], Syed Mustafa Raza Naqvi[4]

[*1]*Independent AI Researcher*
[2,4]*Sir Syed University of Engineering and Technology*
[3]*Habib University*

[*1]alina.16.af@gmail.com, [2]smuqeem@ssuet.edu.pk, [3]nisha.irshad@habib.edu.pk,
[4]2023F-BCV-005@ssuet.edu.pk

DOI:https://doi.org/10.5281/zenodo.22915733

Keywords
Cross writer generalization, Digit specific CNN, Image classification, Handwritten digit recognition, writer dependent domain shift.

Article History
Received: 30 January 2026
Accepted: 15 March 2026
Published: 31 March 2026

Copyright @Author
Corresponding Author: *
Alina Fatima

Abstract
Handwritten digit recognition systems can achieve high accuracy under the defined evaluation while showcasing weaker performance when applied to handwriting of unseen writers. In this research paper we investigated writer dependent domain shift in handwritten 3 digit recognition and tested whether the training duration influences cross writer generalization. A digit specific convolutional neural network was implemented on cross writer experiments where two writers were used for training and the third writer was excluded for testing. The model was tested using multiple durations and random seeds. Then we evaluated on five completely new independent writers. Even though the model achieved 95.3% accuracy on the benchmark evaluation, the exact three digit accuracy fell down to 30.2% in one of the most challenging configuration. During the cross writer experiment the performance was varied due to writer's identity. Increasing the training duration only produced little improvements. Then Independent writers evaluation was conducted which resulted the accuracy of 66.78% on exact three digit, whereas the writer level performance ranged from 42.78% to 86.11%. The individual digit position accuracy was higher at 86-87%, which indicates that many complete number error resulted from mistakes in one or more digit positions. These finding establishes that receiving high baseline accuracy does not alone equate the robust writer independent generalization. These findings highlight the importance of evaluating handwriting recognition systems on unseen writers.

1. INTRODUCTION:

One of the most well-known problems in pattern recognition and computer vision is handwritten digit recognition. The application of this technology is in optical character recognition, document digitization, postal automation, financial document processing and other systems that have to convert handwritten information into machine readable format [1], [6]–[9]. The MNIST handwritten digit dataset is the most widely used for evaluating digit recognition algorithms [1]. Convolutional Neural Networks have shown promising performance on handwritten digit recognition by learning from the visual data[1], [7-9]. Nonetheless, High recognition accuracy under the established benchmark does not usually equate to

reliable performance as the model was shown handwritten digit images from handwriting that was not present in MNIST dataset. Handwriting is different for every individual person due to differences in their character shape, stroke structure, size, spacing alignment and etc [2]. Hence, a model trained on one population of writers may meet with data which has different distribution [2,3].

Writer independent recognition is an important factor to consider in practical handwriting recognition systems, where user specific retraining cannot be assumed [2]. This raises the concern on how handwritten recognition systems should be evaluated. A model can achieve high accuracy when the training and evaluation data are matched but the performance will degrade under writer level distribution shift [2],[3,5]. The problem becomes significant when the recognition is of multi digit and not of a single digit. Due to encountering any mistake in even single position can lead to getting the whole prediction wrong. In this research paper, we conducted investigation on writer dependent domain shift in handwritten three digit recognition using a digit specified CNN. We started with achieving 95.3% model accuracy on synthetic data. We then changed the conditions to evaluate the model and created the cross writer conditions. Training on two writers and then testing on the third write who was excluded from the training. During the experimentation we observed that the most challenging configuration was the training on writer A and B and testing on writer C which resulted in 30.2% accuracy on three digit numbers. This huge disparity shows why high accuracy does not automatically mean that the model will predict accurately.

Rather than treating this reduction as an isolated failure case, we researched whether this showcases systematic writer dependent variation. The cross writer variation were writer A + writer B tested on writer C, writer A + writer C tested on writer B and writer B + writer C tested on writer A. Each cross writer configuration was examined across multiple random seeds and training durations. We then extended our evaluation to 5 more writers named writer H,I,J,K and L. They were completely excluded from the training phase of the experiment. This independent evaluation test whether the observed behavior in A-C persists or not.

In this study we formulated three questions. The first question is how much influence does the identity of the unseen writer make on cross-writer identification performance The second question is whether increased training duration can improve generalization of unseen writers. And finally, the writer-dependent variation still exists even if the models trained are tested on a new unseen writer. The results show that writer identity has a substantial effect on the generalization. In the primary A-C evaluation, exact three digit accuracy varies considerably depends on which writer was held out. When we increased the training duration from 10 epochs to 50 epochs, this produced a little improvement. Evaluation on writer H-L achieved 66.78% accuracy the performance varies from writer I which was 86.11% and writer L which was 42.78%. Individually, the accuracy was higher at 86-87%. This indicates that error arises across individual digit positions. This study is an empirical examination of how benchmark performances changes when writer level distribution shift. The findings emphasizes the value of evaluating handwriting recognition models under real unseen handwriting, rather than conventional benchmark accuracy.

2. Literature review:

2.1 Handwritten Digit and Character Recognition

One of the most important research problem in pattern recognition and document analysis is handwritten digit and character recognition, as handwriting can exhibit huge variations in shape, style, stroke structure and size. While the MNIST dataset has been one of the most widely used benchmarks for handwritten digit recognition research, the more recent EMNIST dataset expanded the scope to a much larger set of handwritten digits and characters. Baldominos et al. [1] reviewed a large number of methods evaluated on MNIST and EMNIST and found that the use of convolutional neural networks

(CNNs) and other deep learning approaches have become increasingly dominant in handwritten recognition tasks. Their study also highlights that recognition performance on MNIST has become largely saturated, driving research into more challenging recognition scenarios.

The strong results of deep learning approaches on standard digit recognition datasets have made CNN-based methods an important baseline for image-based handwriting recognition. Nevertheless, while high performance on a given digit recognition benchmark is often an important result, it does not necessarily indicate robustness to variation in handwriting generated by unseen writers or differing data distributions. Therefore, evaluating a recognition systems on only a single dataset or with only a single population of writers may give an incomplete picture of the practical performance of a model.

Many recent studies have explored a CNN-based approaches and training configurations for handwritten digit recognition. For example, Banaszewska and Plechawska-Wójcik [8] compared several CNN architectures on MNIST and explored factors such as number of training epochs and classification performance. Their results illustrate that architectural choices and training configurations can affect digit recognition performance, which motivates careful consideration of such factors in the design of handwritten digit recognition systems.

2.2 CNN-Based Handwriting Recognition

CNNs have been placed at the forefront of the developments in automated handwritten recognition since their spatial convolutional operations enable them to learn image patterns from training data. LeCun et al. [2] illustrated the effectiveness of gradient based learning convolutional schemes for document recognition, setting the ground for future CNN-based approaches to handwritten character and digit recognition. Since then, the application domain of CNNs in handwriting analysis has been expanded to address increasingly complicated recognition tasks.

The efficiency of CNNs in character recognition is further highlighted in more recent studies also focusing majorly on handwritten digit recognition. Azawi [6] evaluated transfer-learning-based approaches to handwritten digit recognition involving multiple deep CNN models, reporting that different CNN architectures can be leveraged to achieve high recognition rates. Likewise, Akhand et al. [7] proposed a CNN-based approach to handwritten numeral recognition, showing that convolutional operations are still instrumental in achieving good performance in character recognition tasks. In more recent times, Banaszewska and Plechawska-Wójcik [8] compared the performance of VGG-16, VGG-19, and AlexNet CNN architectures for handwritten digit classification, also analyzing the influence of the number of epochs on model performance. This line of work shows that the choice of CNN and its training parameters remains relevant in digit recognition.

Nonetheless, most of the aforementioned works assess the efficacy of their approaches within the paradigm of writer-dependent recognition where the writer of the test sample is represented in the training set. Hence, most conventional performance evaluation practices are not applicable in the case of writer-independent recognition where the test writer is not present in the training set.

2.3 Writer-Independent and Cross-Writer Recognition

Handwriting recognition becomes a harder challenge when the writer represented in the test data is not present in the training data. This setting, often termed writer-independent recognition requires a recognition system to identify characters in spite of variations in individual writing styles. Hu et al. [3] investigated writer-independent online handwriting recognition with Hidden Markov Model and addressed writer specific variations through signal normalization and invariant features; their work shows that handwriting styles can significantly affect recognition performance and that robustness to such variations is an important consideration in recognition.

Through the lens of domain adaptation the effect of differences between training and testing

handwriting distributions have also been investigated. Hosseinzadeh and Razzazi [4] proposed a large-margin domain-transfer method for handwritten digit recognition and noted that recognition can degrade when training and testing data are drawn from different dataset samples. Their approach aimed to reduce the distribution mismatch between source and target data and adapted the classifier to target domain; their study highlights the importance of considering distribution differences when evaluating handwritten digit recognition systems.

The more recent works have focused on the influence of writer-specific characteristics. Wang and Du [5] proposed a writer-adaptation framework which extracts personalized writing-style information and uses it to adapt a writer-independent recognizer towards a certain writer; their results illustrate that incorporation of writer-specific information can lead to improved recognition performance, emphasizing the role of writing style in accuracy.

These researches show that recognition performance cannot be entirely captured through evaluation on handwriting that follows the same distribution as its training data; nevertheless, the present study focuses specifically on writer-disjoint three-digit recognition, where complete three-digit predictions and individual digit positions are evaluated for writers that were excluded from training. This allows the effect of unseen handwriting styles to be investigated at both the complete-number and individual-digit level.

2.4 Writer Adaptation and Domain Shift

Different handwriting styles lead to the distribution mismatch between the data used for training and testing of the recognition model. As a result, writer adaptation and domain adaptation methods are considered to increase recognition performance in the case when the target handwriting is different from the training one. Hosseinzadeh and Razzazi [3] developed the large-margin domain-transfer method for handwritten digits recognition in order to decrease the distribution mismatch between source and target datasets. They found that adaptation to the target domain increases recognition performance if training and test datasets have different distributions.

Writer adaptation approaches were also considered as a way to address individual writer characteristics. Wang and Du [5] developed the writer adaptation method which uses the style extractor network to recognize handwritten text. This approach allows extracting individual style characteristics of each writer and using them to adapt the writer-independent recognition model to an individual writer. The consideration of writer style in this problem proved to be useful as it allows accounting for writer-related variability.

The analysis of these methods shows that the recognition performance depends not only on the structure of the recognition architecture but also on the distribution mismatch between training and test datasets. In the current research, this means that the possibility of recognizing data with a handwriting distribution different from the training one should be examined. Thus, instead of performing domain adaptation, the current study aims at examining the distribution mismatch and considering fine-tuning as the adaptation method for real-world handwriting.

2.5 Synthetic-to-Real Handwriting Recognition

One of the possible solutions to the problem of gathering large-scale data is using synthetic handwriting since the collection and manual labeling of real handwriting might become rather difficult. However, the domain gap between synthetic and natural handwriting can negatively affect the performance of handwritten word recognition. The authors of [4] tackled this issue in the case of handwritten word recognition where they trained the recognizer with synthetic handwriting and tested its performance on real handwritten words. Their results show that although the usage of such models brings some promising results, these models still experience problems when working with real handwritten words. To solve the problem of poor results when working with real data, Kang et al. proposed the method of writer adaptation to adapt a synthetic recognizer for recognizing new handwriting. [4]

A more recent study continued the investigation of the problem of transition from synthetic to real

data. Wolf and Fink [10] proved that the models trained on synthetic handwriting can provide good initial models for handwritten word recognition and retrieval that can be further optimized by self-training on target data. It shows that synthetic data is valuable in terms of performance but at the same time needs to be adapted to real data. [10]

The results above can be considered as a motivation for the analysis of our model's performance when transitioning from synthetic to real three-digit handwriting. Unlike the adaptation techniques discussed in previous studies, the current analysis will start with the degradation of the performance of the model trained on synthetic handwriting when applied to real handwriting. The fine-tuning technique will be used as the adaptation approach.

2.6 Research Gap

Earlier studies have established the effectiveness of deep learning and convolutional neural networks (CNNs) in recognizing handwritten digits and characters. Other works have addressed the challenges of writer-independent recognition, writer adaptation, and adapting to real-world data from synthesized data. But all these challenges are often addressed separately, and most of the work is focused on single-digit recognition or full handwritten word recognition as opposed to controlled multiple-digit recognition.

Additionally, just accuracy in recognition might not always provide enough insight into generalization properties of a model since there could be one or multiple correct digits in a three-digit prediction that would be still wrong in terms of predicting the full three-digit number. Analysis of accuracy of exact three-digit numbers prediction and separate digits' recognition could provide insight about the differences in recognition properties.

This research is aimed at addressing these gaps in the literature through an interrogation of a digit-specific CNN in three-digit recognition tasks with writer-disjoint and synthetic-to-real settings. Specifically, we looked at cross-writer generalization abilities using unseen writers, accuracy of hundreds, tens, and units digits in recognition separately, additional independent writers for testing, and difference in performance in synthetic vs. real handwriting setting. Finally, we looked at fine-tuning on real handwriting as a potential way to solve the domain gap.

3. Methodology

3.1. Dataset

The dataset used in this study consists of two sources, Firstly the MNIST data for the synthetic training and secondly handwritten digit sample created by 3 different people named writer A, B and C. The write specific dataset consists of handwritten images of individual digits from 0 to 9. Then these digits were reconstructed to three digit numbers, allowing the study to investigate recognition at multi digit level. The total of writer specific 3 digit samples were 1500 with each writer contributing 500 samples. Each sample was a grayscale image with spatial resolution of 84 x 28 pixels. The labels contain three digit positions Hundredth, Tens and Unit. Additionally MNIST dataset was also used to generate 3 digit training samples which were then used to test the generalization. The dataset was structured according to data source and writer identity. Image and label information were kept separately for each writer. The MNIST dataset samples were stored as training set without assignment to one of the three writers. The study uses MNIST data for training, augmentation and writer based partitions for evaluating how well the model generalizes.

3.2. Data Preprocessing and preparation

The cross writer dataset was arranged in the structure of their respective writer and data source. Then the individual digit images were selected and combined to form a three digit number image. That possessed the spatial dimension of 84 x 28 pixels and resolution was greyscale. The same configuration was applied to the MNIST data to generate three digit number images, with each sample assigned a label corresponding to the image. Both of these datasets were kept separate.

Data Augmentation was also performed as a part of our experimental process to ascertain whether introducing variation into the training data could improve our models performance. Rather than the models performance increasing it fell down to

87.25%, therefore it did not improve the generalization. Due to this finding augmentation was removed from the final training pipeline. The final experiment used non-augmented data. Throughout the experiment maintaining the writer identity was crucial for creating the writer disjoint experimental configuration used in this research.

3.3. Writer Based Data Partitioning

To test the cross writer generalization the writer specific was partitioned by writer identity rather than random sample level split. The three writers were A, B & C. In each experimental run samples from two writers were used for training, while the samples from the third writer were reserved for testing. The writer which was present in testing was not shown during the training.

Table 1

Test Writer	Training Writers
Writer C	Writer A + Writer B
Writer B	Writer A + Writer C
Writer A	Writer B + Writer C

Table 1 show the writer specific partitioning, you can see in the first configuration, writer A and B were used for training and writer C was used for testing. Similarly in the second configuration writer A and C were used for training and writer B was used for testing. Then in the third configuration writer B and C were used as training and writer A was used as testing. These configuration allowed each writer to be used as unseen writer while the other two were served as source training writers.

This partitioning strategy was designed to stimulate a domain shift scenario in which the handwriting styles which was encountered during the testing was different than the training styles because the testing writer was completely excluded from the training in each configuration, the results of the test performance are a direct measure of the model capability to generalize to previously unseen handwriting. The writer specific based partitioning was maintained throughout the training and evaluation, also the image samples from the test writer was not shown in the training set. This was an essential step because a conventional split could place samples from the same writer in both training and testing set which would then produce an overly optimistic results of the generalization.

3.4. Digit-specific CNN architecture

A digit specific convolution Neural Network was used. A specific CNN breaks the three digits and uses 3 models to detect the numbers. Instead of using one CNN to detect the entire number, our approach uses 3 distinct classifiers, each responsible for their position. The first CNN recognizes the hundredth position than the second recognizes the tens position and the third recognizes the unit position. Each CNN was 10 class classifier, predicting digits 0-9.For example if the number to detect is 527, then the first CNN would detect 5, then the second CNN would detect 2 and likewise the third CNN would detect 7. These 3 predictions would then be combined to the final output. This approach causes the model to learn the visual characteristics of each digit position while avoiding the need to learn 1000 separate number classes. Therefore, if the actual number is 527 but the model predicts 523, it predicts 2 positions correctly even though the three digit number is incorrect. This is the reason our digit level accuracy is higher than the exact 3 digit accuracy. We used the digit specific CNN by a controlled cross writer in which we trained on two writers and then evaluated on the third using multiple durations and random seeds and then the resulted model was tested on five more independent writers who were not included in any of the training runs. Though on synthetic MNIST data the model achieved 95.3% accuracy but the accuracy fell to 30.2% in cross writer configuration.

Table 2 Performance on writer specific samples across 10 Epochs

Test Writer	Training Writers	Accuracy (%)
Writer A	Writer B + Writer C	90.6
Writer B	Writer A + Writer C	89
Writer C	Writer A + Writer B	30.2

The performance varies according to writer identity meanwhile increasing the training duration did slightly increase the performance.

Table 3 Performance on writer specific samples across 50 Epochs

Model	Test Writer	Epochs	Exact 3-Digit Accuracy	Hundreds Accuracy	Tens Accuracy	Units Accuracy
AB	C	50	34.2% ± 1.10%	73.2% ± 0.45%	64.0% ± 0.00%	71.0% ± 0.71%
AC	B	50	90.4% ± 0.55%	95.4% ± 0.55%	98.0% ± 0.00%	97.0% ± 0.00%
BC	A	50	93.8% ± 0.45%	97.0% ± 0.00%	99.0% ± 0.00%	97.8% ± 0.45%

The independent evaluation on writer H-L achieved 66.78% on overall the exact three digit accuracy.

Table 4 writer H-L Performance

Evaluation	Exact 3-Digit Accuracy (%)	Hundreds Accuracy (%)	Tens Accuracy (%)	Units Accuracy (%)
H–L Independent Evaluation	66.78	86.78	87.33	86.44

IEER
Institute for Excellence in Education & Research

This indicates that many complete number errors resulted from mistakes in one or more digit position. These findings also illustrate that high accuracy during training does not create a robust writer independent generalization.

3.5. Training Procedure

The CNN models used Adam as the best optimizer. The initial experiments investigated the effects of optimizer selection and learning rate using a fixed training duration of 10 epochs. We used SGD for evaluation with learning rates of 0.001, 0.01 and 0.1. Additionally we also used Adam as mentioned earlier but with the learning rate of 0.0001, 0.001, 0.01 and 0.1. Among them Adam with the learning rate 0.001 outperformed the rest of the configurations. The complete number accuracy was 89.20% which was followed by SGD of learning rate 0.1 at 89.0%. The effect of Epochs was also observed by increasing the epochs from 5 to 10. This caused the complete number accuracy to improve from 88.10% at 5 Epochs to 88.90% 10 epochs.

Data Augmentation was evaluated but due to not achieving good performance was excluded. The Augmented configuration only achieved 87.25% complete number accuracy which was 1.65 points reduction. The digit specific CNN achieved 95.30% accuracy on the synthetic dataset. These results were an improvement over the standard CNN. Then the selected architecture was evaluated on real handwritten data to investigate the effect of difference between real handwriting and synthetic training.

3.6. Evaluation Metric

Performance of the models was measured with quantitative classification metrics, computed on the test set. Classification accuracy was the key evaluation metric and denotes the fraction of the predictions correctly recognized by the model. Accuracy was the main performance measure

utilized to compare performance between the different writer-disjoint settings.

Classification accuracy was complemented with analysis at the digit level to find out if there was any difference in recognition performance for the particular digit places. The predictions were saved during the testing stage and compared to their respective true labels, providing the means of analyzing performance of the models in addition to their aggregate score.

Evaluation results were compared between the three writer-disjoint settings in order to find out if there was any variance in recognition performance depending on which writer was not used for training.

4. Experimental Setup

4.1. Experimental Configuration

A series of experiments were carried out to gradually assess the impact of model design, training length, data augmentation, optimizer selection, and digit-specific modeling on three-digit handwritten number recognition. The studies started with a fully connected baseline and ended with a conventional convolutional neural network (CNN). The CNN setup was then tested with an increased number of training epochs and data augmentation. LeNet was also considered an alternate convolutional design. In addition, SGD and Adam were compared using various learning rates to discover the best optimization parameters. Based on these findings, a digit-specific CNN was created to individually model the three digit positions of the input. The resulting synthetic-data model was then tested on real handwriting to study the synthetic-to-real domain transition, followed by fine-tuning to see if adaptation could recover performance.

Table 5. Experimental configuration table

Experiment	Architecture	Optimizer	Learning Rate	Epochs	Augmentation	Dataset
Baseline	Fully Connected	Adam	0.001	5	No	Synthetic
CNN	CNN	Adam	0.001	5	No	Synthetic
More Epochs	CNN	Adam	0.001	10	No	Synthetic
Data Augmentation	CNN	Adam	0.001	10	Yes	Synthetic
LeNet	LeNet	Adam	0.001	10	No	Synthetic
SGD — Best LR	CNN	SGD	0.1	10	No	Synthetic
Adam — Best LR	CNN	Adam	0.001	10	No	Synthetic
Digit-Specific CNN	Digit-Specific CNN	Adam	0.001	10	No	Synthetic
Real Handwriting Baseline	Digit-Specific CNN	N/A	N/A	N/A	N/A	Real
Real Handwriting Fine-Tuning	Digit-Specific CNN	Adam	0.0001	10	No	Real

4.2. Cross writer experimental design

To evaluate the cross writer generalization, real handwritten data was partitioned by the identity of the writer rather than randomly combining the samples from all writers. Three writers A, B and C were used to construct the writer disjoint training and testing configuration. In each experiment two writers were trained upon and one was held out for the testing. This created three cross writer configuration. Like A+B → C, A+C → B, and

B+C → A. The held out dataset was not included in the training dataset Each configuration was tested for accurate three-digit accuracy as well as accuracy at the hundreds, tens, and unit levels. The best-performing epoch configuration was utilized in the cross-writer comparison.

Table 6. Cross writer training and testing configuration

Training Writers	Test Writer	Model
A + B	C	AB
A + C	B	AC
B + C	A	BC

4.3. Random Seed Configuration

The training process was repeated with five different random seeds in order to assess the cross writer findings stability. The random seed was altered between runs while maintaining the model architecture, training data, hyperparameter and evaluation process. Across the five seeds, the three writer disjoint (A+B→C, A+C→B, and B+C→A) configurations were evaluated. The mean, standard deviation, minimum, and maximum accuracy with 95% confidence range were reported for capturing and summarizing the exact three-digit accuracies obtained for each configuration. This process was done to ascertain whether the observed performance was independent of a single training run and was similar across many random initialization.

Table 7. Detailed Random Seed Experiment

Test Writer	Training Writers	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean (%)	SD (pp)	Min (%)	Max (%)	95% CI
Writer A	B + C	89	93	89	93	89	90.6	2.19	89	93	87.88–93.32
Writer B	A + C	89	89	89	89	89	89	0	89	89	89.00–89.00
Writer C	A + B	30	30	31	30	30	30.2	0.45	30	31	29.64–30.76

4.4. H-L independent Evaluation setup

An additional independent evaluation was carried out on five further writers named H,I,J,K and L. These writers were not used during the training of the models in the experiment. The aim of this evaluation was to check the behavior of the cross writer generalization to new writers handwriting style. Each writer contributed 200 images samples and then from which 60 three-digit images were constructed which resulted in 300 three-digit evaluation images across the five writers. The already trained model was directly tested on the new data without any fine tuning or additional training. Both individual position accuracy and exact three digit number accuracy was measured. This test was an additional assessment of the model to verify whether it could generalize.

Evaluation Component	Test Writers	Digits per Writer	Three-Digit Images per	Total Three-Digit	Models Evaluated	Total Model-Sample	Training on H–L	Fine-Tuning on H–L	Evaluation

			Writer	Images		Evaluations			
Configuration	H, I, J, K, L	200	60	300	3	900	No	No	Exact three-digit + position-wise accuracy

Table 8. Summary of writer H-L independent evaluation

5. Results

5.1. Effects of Training Epochs

We evaluated the duration of training on the cross writer samples by training the model for 10, 20, 30 and 50 Epochs. The Performance was evaluated using three digit accuracy as well as individual position accuracy of hundredth, tens and units.

Table 9. Epoch Ablation

Model	Test Writer	Epochs	Exact 3-Digit Accuracy	Hundreds Accuracy	Tens Accuracy	Units Accuracy
AB	C	10	30.2% ± 0.45%	72.8% ± 0.84%	60.4% ± 1.14%	71.2% ± 0.45%
AB	C	20	32.0% ± 0.00%	72.6% ± 0.55%	63.4% ± 0.55%	71.0% ± 0.71%
AB	C	30	32.8% ± 0.45%	72.6% ± 0.55%	64.0% ± 0.00%	70.8% ± 0.84%
AB	C	50	34.2% ± 1.10%	73.2% ± 0.45%	64.0% ± 0.00%	71.0% ± 0.71%
AC	B	10	89.0% ± 0.00%	94.0% ± 0.00%	98.0% ± 0.00%	97.0% ± 0.00%
AC	B	20	89.8% ± 0.45%	94.8% ± 0.45%	98.0% ± 0.00%	97.0% ± 0.00%
AC	B	30	90.0% ± 0.00%	95.0% ± 0.00%	98.0% ± 0.00%	97.0% ± 0.00%
AC	B	50	90.4% ± 0.55%	95.4% ± 0.55%	98.0% ± 0.00%	97.0% ± 0.00%
BC	A	10	93.0% ± 0.00%	97.0% ± 0.00%	98.0% ± 0.00%	98.0% ± 0.00%
BC	A	20	93.0% ± 0.00%	97.0% ± 0.00%	98.0% ± 0.00%	98.0% ± 0.00%
BC	A	30	94.0% ± 0.00%	97.0% ± 0.00%	99.0% ± 0.00%	98.0% ± 0.00%
BC	A	50	93.8% ± 0.45%	97.0% ± 0.00%	99.0% ± 0.00%	97.8% ± 0.45%

These results highlight that even if we increase the number of Epochs still it does not produce a uniform improvement in the cross writer generalization. The highest exact three digit accuracy was 50 Epochs for model AB and model AC while model BC was at epochs 30.

5.2. Cross Writer Generalization

The cross writer experiments showed variations in recognition performance depends on the writer which was excluded from training. The model when trained on writer A and B and tested on writer C it achieved an exact three digit accuracy of only 34.2% ± 1.10% at its best epoch. Similarly, training on writer A & C and testing on B produced an exact three digit accuracy of 90.4% ± 0.55%. The best performance was when writer B & C were trained and tested on writer A, this generated an exact three digit accuracy of 94.0% ± 0.00% at 30 Epochs. The results showcased an asymmetry in cross writer generalization. The same architecture performs very differently depending on which writer were tested as the unseen writer. The most challenging writer was writer C as they demonstrated a lower three digit accuracy. The distribution of handwriting styles of the training samples and the characteristics of the testing samples have an effect on recognition performance.

5.3. Stability across random seed

The stability was assessed of the cross writer result, each writer disjoint configuration was evaluated across 5 training seeds mean, standard deviation, minimum, maximum values and 95 confidence intervals were summarized.

Table 10. Cross-writer recognition accuracy across five random training seeds.

Test Writer	Training Writers	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean Accuracy (%)	SD (pp)	Min (%)	Max (%)	95% CI
Writer A	**B + C**	89	93	89	93	89	**90.6**	**2.19**	89	93	87.88–93.32
Writer B	**A + C**	89	89	89	89	89	**89**	**0**	89	89	89.00–89.00
Writer C	**A + B**	30	30	31	30	30	**30.2**	**0.45**	30	31	29.64–30.76

These results establishes that the performance difference between the target writers was larger than a random initialization. When writer A was used for testing the mean accuracy across the 5 seeds was 90.6% with the standard deviation of 2.19 percentage points and values from 89 to 93. For writer B the mean accuracy was 89.0% with no observed variation across the 5 seeds. For writer C the mean accuracy was 30.2% with standard deviation of 0.45 percentage points. The results point out that the observed difference in the cross writer performance are not attributable to random training initialization. The consistently low performance for writer C across five seeds indicates that the difficulty regarding the test writer is systematic rather than the results of a single unstable training run.

Table 11. Position-Wise Recognition Accuracy across Test Writers

Test Writer	Hundreds (%)	Tens (%)	Units (%)	Best Position	Worst Position	Within-WriterGap (pp)
Writer A	97	98	95.6	Tens	Units	2.4
Writer B	94	98	97	Tens	Hundreds	4
Writer C	72.8	60.4	71.2	Hundreds	Tens	12.4
Overall Average	**87.93**	**85.47**	**87.93**	**Hundreds**	**Tens**	**2.46**

5.4 Position wise recognition performance

We analyzed recognition performance in addition to exact three digit number accuracy. The analysis provided a more detailed view of the error that make up the incorrect complete number prediction. The largest writer difference was observed for writer C with 12.4% gap between the best position Hundredth and the worst performing position Tens. In comparison to writer A & B were only 2.4 and 4.0%. The average position wise accuracy was 87.93% for Hundredth, 85.47% for tens and 87.93% for units was observed. The result demonstrate the degradation in the most difficult writer is not the same across all digit position. The Tens position exhibits the lowest average and weakness for writer C.

5.5 Synthetic to real domain shift

To further investigate the different effects between synthetic training data and real handwritten data. The selected three digit specific CNN model achieved 95.30%on the synthetic data when evaluated on real handwriting dataset performance decreased to 82.50%. The decrease in percentage points shows a distribution difference between the data the model used in development and the data used in evaluation. The behavior shows a consistent decrease in recognition from synthetic dataset to real handwriting dataset due to the presence of a domain shift. The digit specific CNN was fine tuned using the real handwriting data samples. Fine tuning did help in increasing the accuracy from 82.50% to 92.50% , this result was now closer to the accuracy of the original synthetic data performance of 95.30%.These results indicate that the models performance was sensitive to changes from synthetic to real handwritten dataset. Additionally it was also observed that fine tuning also helped that the model adapted to the real handwritten data and recover the last performance. This experiment therefore provides empirical evidence of synthetic to real domain shift.

Table 12. Synthetic-to-Real Domain Shift and Fine-Tuning Results

Configuration	Dataset	Exact 3-Digit Accuracy
Digit-Specific CNN	Synthetic	95.30%
Real Handwriting Baseline	Real	82.50%
Real Handwriting Fine-Tuning	Real	92.50%

5.6 Experimental Comparison

Multiple experiments were observed to evaluate the effects of the model architecture, training duration, data augmentation, and optimizer selection and learning rate on the proposed three digit recognition performance. The digit-specific CNN utilized in the ensuing real-handwriting and cross-writer studies was generated gradually from a fully connected baseline.

On the synthetic dataset, the fully connected baseline obtained an exact three-digit accuracy of 68.45%. When a CNN was used in place of the fully connected architecture, performance significantly increased to 88.10%, proving the value of convolutional feature extraction for the handwritten digit images. An further improvement to 88.90% was obtained by extending the training period from 5 to 10 epochs. The CNN setup was also used to assess data augmentation. Nevertheless, compared to the similar non-enhanced 10-epoch CNN setup, the augmented configuration obtained 87.25%, which was 1.65 percentage points lower. As a result, the final training configuration did not include augmentation. Additionally, a LeNet-based architecture was assessed; its score of 81.10% was less than that of the conventional CNN setup.

The impact of learning-rate and optimizer selection was then examined. Adam obtained an exact three-digit accuracy of 89.20% with a learning rate of 0.001, but SGD with a learning rate of 0.1 produced an accuracy of 89.00%. As a result, the latter was chosen as the ideal optimizer setup for the studies that followed.T he digit specific CNN outperformed the other

configurations with the accuracy of 95.30%. When we compare this with the normal CNN using Adam as an optimizer with the learning rate of 0.001 in 10 epochs resulted with 89.20%, the digit specific CNN improved the result by 6.10 percentage points.

In general the experimental history illustrates the model architecture had a higher impact on recognition performance than just increasing the epochs. The shift from the fully connected baseline to CNN resulted in an improvement, with digit specific CNN producing the best synthetic data performance among the models. These experiments also showcase that commonly used strategies such as data augmentation may not always achieve good performance.

Table 13. Comparison of Model Architectures, Training Configurations, and Recognition Accuracy

Experiment	Architecture	Optimizer	Learning Rate	Epochs	Augmentation	Dataset	Accuracy
Baseline	Fully Connected	Adam	0.001	5	No	Synthetic	68.45%
CNN	CNN	Adam	0.001	5	No	Synthetic	88.10%
More Epochs	CNN	Adam	0.001	10	No	Synthetic	88.90%
Data Augmentation	CNN	Adam	0.001	10	Yes	Synthetic	87.25%
LeNet	LeNet	Adam	0.001	10	No	Synthetic	81.10%
SGD – Best LR	CNN	SGD	0.1	10	No	Synthetic	89.00%
Adam – Best LR	CNN	Adam	0.001	10	No	Synthetic	89.20%
Digit-Specific CNN	Digit-Specific CNN	Adam	0.001	10	No	Synthetic	95.30%
Real Handwriting Baseline	Digit-Specific CNN	N/A	N/A	N/A	No	Real	82.50%
Real Handwriting Fine-Tuning	Digit-Specific CNN	Adam	0.0001	10	No	Real	92.50%

5.7 H–L Independent Evaluation

To further examine the trained models' capacity to generalize across handwriting styles, an independent evaluation was carried out with five more authors, identified as authors H, I, J, K, and L. These writers were not used to train the three models used in this experiment. The previously developed models were applied to these writers' handwriting samples without further training or fine-tuning. Each writer provided 60 three-digit samples, which were evaluated using the three previously trained writer-specific models, yielding 180 evaluation instances per writer and 900 evaluation instances total.

The overall exact three digit accuracy among the five new writers was 66.78% with 601 out of 900 images correctly detected. When the individual digit position was further examined the model achieved 86.78% for the hundredth position, 87.33% for the tens position and 86.44% for the

unit position. These difference between individual position accuracy and three digit accuracy demonstrate that errors in any single digit position can cause a partially correct prediction to be counted wrongly at the complete number level.

Significant variation was observed among the additional writers. As indicated in the table 10, the writer I had the highest three digit accuracy followed by writer J then writer H then writer K and lastly writer L had the lowest performance. The position wise result was same as writer H,I and J retaining greater individual digit accuracy than writer K and L.

Table 14. Cross-Writer Generalization Performance on Unseen Writers

Writer	Exact 3-Digit Accuracy (%)	Hundreds Accuracy (%)	Tens Accuracy (%)	Units Accuracy (%)
Writer H	77.22	92.78	91.67	90.56
Writer I	86.11	93.33	96.67	95.56
Writer J	79.44	92.22	94.44	91.67
Writer K	48.33	82.78	77.22	76.11
Writer L	42.78	72.78	76.67	78.33

These results provide evidence that cross writer generalization is influenced by the characteristics of the unseen writer. The significant difference between the performance of writer H-J and writers K and L shows that the trained model do not generalize well to all writing styles. Additionally it should be noted that these results were obtained without any extra training or fine tuning of the new writers, allowing the evaluation to assess the ability of the previous trained model to generalize to handwriting styles that were not shown during the training.

6. Discussion

6.1 Effect of Model Architecture and Training Configuration

The results of the experiments demonstrate that the choice of models architecture and the configuration of training can affect the quality of recognition of the three digit numbers. The fully connected model with the baseline configuration achieved the accuracy of 68.45%. Whereas using a CNN architecture increased the result to 88.10%. This suggests that the use on convolutional layers in the neural network allowed to achieve better results on the given dataset compared to the fully connected architecture.

When we increased the number of epochs from 5 to 10 it resulted in relatively small improvement. From 88.10% to 88.90%. This suggests that simply extending the training duration provided only limited benefits in this scenario. The accuracy of recognition produced after data augmentation was lower than non-augmented CNN configuration. Therefore the method was not beneficial and was not adopted in the following experiments.

Similarly when we compared LeNet it highlighted that a different CNN is not guaranteed to provide better performance as shown by the results 81.10% which was lower than the standard CNN configuration. It can be concluded that the increase in or changes in architectural complexities was insufficient in the improvement of the performance on the particular three-digit recognition task and dataset used in the current study. The comparison between the optimizer yielded the same results for Adam (89.20%) and SGD (89.00%) in their best tested learning rate configuration and therefore produced comparable improvements. Hence, the choice of the optimizers have a less impact than the architectural variations considered in the preceding experiments. The most significant increase was observed for the digit-specific CNN at 95.30%. Explicitly, modeling the individual digit positions appeared to be more beneficial for the three-digit recognition task than the approach of conventional undifferentiated recognition.

Overall the experiments suggests that architecture had a larger influence on performance than simply

increasing the epochs or changing the optimizer. In particular the improvement in accuracy seen from digit specific CNN supports the position aware architecture for the purpose of recognizing multi-digit handwritten numbers. This architecture was selected as the basis for subsequent cross-writer and synthetic-to-real experiments

6.2 Cross-Writer Generalization

The results from cross writer experiments show that depending on which writer used for testing, the models recognition varies. For example , while training on data from writer B and C and testing on writer A yielded an exact three-digit accuracy of 90.4%, meanwhile training on writer A and C and testing on writer B resulted in 89% of performance. However, training on writers A and B and testing on C yielded a drop in performance to 30.2%. Therefore the models ability to generalize to previously unseen writer is not consistent across all writers.

The increase in the performance gap between the three test writers suggest that writer specific characteristics can have a huge impact on recognition performance. In particular, the low performance observed for writer C consistently across the cross writer configuration and was significantly lower than the performance observed for writer A and B. This indicates that the poor performance for Writer C cannot be attributed solely to the model's inability to perform the recognition task in general, since the same experimental framework resulted in significantly higher accuracies when Writers A and B were used as unseen test writers.

The position wise results provide more insights into the presented difference. Compared to writer A and B writer C recognition accuracy dropped significantly for hundredth, tens and especially units position. Which implies that the problem is not specific to three digit recognition but was also evident at the individual digit level. Thus it can be stated that assessing the performance of a handwritten digit recognition model on writer disjoint data is essential, as significantly different recognition accuracy can be observed for different writers. Therefore, when claiming a high level of handwriting recognition for a particular model, it is essential to ensure that it was tested on data from writers beyond the ones that were present during training.

6.3 Stability across Random Seeds

The cross writer test were repeated using 5 different random seeds, leaving the training configuration unchanged. Results varied little between the 3 test writers: Writer A had a mean accuracy performance of 90.6% ± 2.19% while Writer B had 89.0% ± 0.00%. Writer C had a substantially lower accuracy of 30.2% ± 0.45%.This is a relatively small variance between runs, especially for writer C, suggests that the difference in cross writer performance was not determined by the random seed. In general the stability of the result across 5 seeds support the conclusion of a stable writer-dependent generalization pattern.

6.4 Position-Wise Recognition

The recognition result based on writer position are presented and discussed. It can be seen that the recognition accuracy depends upon the digit position and the writer. For instance writer A and B performed reasonably well across all position while writer C made more errors in tens position. Another finding is that the recognition for each digit and three digit number differs, which indicates that error in any single digit leads to the overall error in three-digit numbers prediction. In conclusion it can be stated that that writer-dependent patterns can affect the accuracy of individual digit positions as well as the complete number prediction.

6.5 Synthetic-to-Real Domain Shift

The results shows that there was a noticeable difference in the accuracy of synthetic and real handwriting. The digit-specific CNN got 95.30% on the synthetic data, but only 82.50% on real handwriting. However, after fine-tuning on real handwriting, accuracy increased to 92.50%. Thus, these results highlight the importance of adapting to the target handwriting distribution to mitigate the negative impact of using synthetic data.

6.6 H–L Independent Evaluation

The independent evaluation on writer H-L revealed different performances for recognizing unseen writer's handwriting styles. The exact three digit accuracy ranged from 42.78% for Writer L to 86.11% for Writer I, with an exact accuracy of 66.78%. Remarkably the two last writers showed significantly worse abilities than H, I, and J, which again proved that recognition ability may vary significantly for unseen sets of writers. As all these writers were tested without additional training or tuning, it allowed to confirm the initial hypothesis about such a possibility.

6.7 Overall Findings and Implications

In general the experiments demonstrated that the models architecture, writer variability, and domain differences affect three-digit handwriting recognition. The digit-specific CNN was efficient on synthetic data, but the cross-writer and synthetic-to-real evaluations indicated significant performance differences. A similar trend was observed across the five random seeds, which verify that the reliability of the results, while fine tuning eliminates the performance gap between synthetic and real-world data. Therefore the performance of handwriting recognition models should be evaluated on multiple writers and domains, not only on the distribution of training data.

7. Limitations

Our Study has number of limitations, firstly the dataset of real handwritten data was limited in both quantity and the number of writers. Which may limit the generalization result to broader population. Secondly the results from the synthetic data may not fully reflect the performance on naturally written data, since the accuracy decreased when the model was tested on real handwriting. Thirdly, while the results were tested against unseen writers, the overall pool of such writers was limited. And finally, while the study used three-digit numbers and a digit-specific CNN architecture, it is unclear to what extent the results could be generalized to larger numbers, words, or different architectures.

8. Conclusion

This research explored the problem of recognition of handwritten three digit numbers with a major focus on cross writer generalization and transfer from synthetic to real domains. From the experimental results, it can be concluded that the digit-specific CNN has high accuracy (95.30%) for recognizing synthetic handwriting. Nonetheless, the cross-writer experiments revealed significant accuracy variations depending on the writer, which emphasized the impact of writer-specific traits on handwriting recognition (HWR). Additionally, the analysis of independent Writers H–L confirmed the tendency. Finally, the significant accuracy drop from 95.30% to 82.50% can be observed when moving from synthetic to real handwriting. At the same time, the accuracy of the fine-tuned CNN on real data was higher 92.50%. Therefore, the overall conclusion is that the accuracy of CNN-based HWR systems is not fully representative due to ignoring critical factors. Particularly, the cross-writer particular, the cross-writer recognition accuracy analysis which includes position-wise recognition and cross-domain experiments can help estimate the generalization power of a CNN-based HWR system under investigation more realistically.

9.REFERENCES

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

J. Hu, S. G. Lim, and M. K. Brown, "Writer independent on-line handwriting recognition using an HMM approach," Pattern Recognition, vol. 33, no. 1, pp. 133–147, 2000, doi: 10.1016/S0031-3203(99)00043-6.

H. Hosseinzadeh and F. Razzazi, "LMDT: A weakly-supervised large-margin-domain-transfer for handwritten digit recognition," *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 119–125, 2016, doi: 10.1016/j.engappai.2016.02.014.

L. Kang, M. Rusiñol, A. Fornés, P. Riba, and M. Villegas, "Unsupervised writer adaptation for synthetic-to-real handwritten word recognition," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3502–3511, doi: 10.1109/WACV45572.2020.9093392.

Z.-R. Wang and J. Du, "Fast writer adaptation with style extractor network for handwritten text recognition," Neural Networks, vol. 147, pp. 42–52, 2022, doi: 10.1016/j.neunet.2021.12.002.

N. Azawi, "Handwritten digits recognition using transfer learning," Computers & Electrical Engineering, vol. 106, Art. no. 108604, 2023, doi: 10.1016/j.compeleceng.2023.108604.

M. A. H. Akhand, M. Rahat-Uz-Zaman, S. Hye, and M. A. S. Kamal, "Handwritten numeral recognition integrating start–end points measure with convolutional neural network," Electronics, vol. 12, no. 2, Art. no. 472, 2023, doi: 10.3390/electronics12020472.

K. L. Banaszewska and M. Plechawska-Wójcik, "Comparative analysis of CNN models for handwritten digit recognition," Journal of Computer Sciences Institute, vol. 32, pp. 179–185, 2024, doi: 10.35784/jcsi.6239.

A. Fateh, R. T. Birgani, M. Fateh, and V. Abolghasemi, "Advancing multilingual handwritten numeral recognition with attention-driven transfer learning," IEEE Access, vol. 12, pp. 41381–41395, 2024, doi: 10.1109/ACCESS.2024.3378598.

F. Wolf and G. A. Fink, "Self-training for handwritten word recognition and retrieval," International Journal on Document Analysis and Recognition, vol. 27, pp. 225–244, 2024, doi: 10.1007/s10032-024-00484-9.

IEER
Institute for Excellence in Education & Research