Analysis of Blind Validation for Learning Models

Abhijeet Jadhav, Poonam M Shettar, Divya Karoshi, Aishwarya G, Akash Kulkarni and Basawaraj Patil
School of Electronics and Communication Engineering, KLE Technological University, Hubli - 580031, India

Keywords: Machine Learning, Deep Learning, Convolutional Neural Networks (CNN), Image Classification, Blind Validation, MNIST, USPS, EMNIST, Model Generalization.

Abstract: Machine learning, deep learning, and image processing have gained significant traction and are widely utilized across various fields for many applications, contributing significantly to advancements in medicine, security, robotics, automation, and beyond. With the expanding range of applications, it is crucial to comprehend how models perform on unseen data and assess their reliability for deployment in real-time scenarios. In this context, this study employs CNN models to provide insights on blind validation. Convolutional Neural Network (CNN) models have emerged as powerful tools for image classification tasks. This study employs a CNN model for digit classification using the famous MNIST dataset. Subsequently, the model's performance is evaluated on two additional datasets: USPS and EMNIST. The evaluation aims to understand how the model generalizes across different datasets with varying characteristics and to assess its robustness in real-world applications. Blind validation is conducted by training the model on the MNIST dataset and testing it on itself and the other datasets to observe potential biases and inconsistencies in the model's behaviour across diverse datasets. This analysis provides valuable insights into the model's adaptability and reliability for deployment in practical scenarios beyond the training dataset's domain.

1 INTRODUCTION

In machine learning, deep learning, and predictive modeling, the imperative to evaluate the model performance extends far beyond theoretical constructs. As models transition from development environments to real-world applications, their effectiveness in handling unseen data becomes paramount. This makes blind validation crucial for determining whether learning models can be deployed in real-world situations.

Convolutional Neural Networks (CNNs) are among the most widely used classification techniques in modern machine learning. Although CNNs often achieve maximum accuracy when trained and evaluated on the same dataset, they often perform poorly when evaluated on the other datasets. One of the significant challenges in deep learning is understanding and identifying potential problems within a pre-trained CNN model for predicting image features. Without a clear theoretical explanation, it is challenging to ascertain why a CNN model performs well when tested on the same dataset but noticeably performs worse when tested on unseen data. Typically, CNN performance is assessed by testing its accuracy with sample images to evaluate its effectiveness. This gap in implementing CNN requires robust approaches to check their reliability in real-world situations.

The primary goal of blind validation is to evaluate the performance and reliability of a machine-learning model. This validation method has been meticulously developed to examine how the model behaves and performs when subjected to unseen datasets, irrespective of the specific characteristics of these datasets. Blind validation thoroughly evaluates the model's generalization ability outside of the training dataset by exposing it to a wide range of novel and varied data.

The MNIST dataset is the most widely used dataset in machine learning for tasks concerned with image classification. It consists of grayscale images of handwritten digits(0-9) and corresponding labels. The MNIST dataset serves as a standard dataset for evaluating the performance of machine learning algorithms, especially for CNN, due to its simplicity and relevance to real-world applications. MNIST pro-

Jadhav, A., M Shettar, P., Karoshi, D., G, A., Kulkarni, A. and Patil, B.
Analysis of Blind Validation for Learning Models.
DOI: 10.5220/0013614300004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 295-300
ISBN: 978-989-758-763-4

vides standard and easily accessible training, testing, and validation images. This study focuses on training a CNN model using the MNIST dataset and assessing its performance on two distinct datasets(USPS and EMNIST), progressively diverging in similarity from MNIST. These datasets serve as unseen data for the model, facilitating a comprehensive evaluation of its ability to generalize beyond the familiar MNIST dataset.

The fundamental components of the CNN model consist of convolutional layers, pooling layers, and dense layers(Dai, 2021). Additionally, the paper examines how the model's accuracy fluctuates with alterations in CNN design, particularly in modifying the convolutional layers. Furthermore, the study investigates the impact on accuracy when adjusting various learning parameters, elucidating how these modifications influence the overall performance of the CNN model for an unseen dataset.

This literature follows the introduction in section 1. Section 2 presents comprehensives the related work on blind validation. Section 3 describes the methodology and datasets. Results and discussions are detailed in section 4 and conclude with section 5.

2 LITERATURE SURVEY

A research paper was conducted on the Convolutional Neural Network (CNN) model for recognizing handwritten numbers. The model was trained using the well-known MNIST dataset, which consists of grayscale handwritten digit photographs. The CNN model achieved an impressive validation accuracy of 98.45% on the MNIST dataset. To test the model's ability to handle unknown data, the researchers ran it through a series of random photos with handwritten and printed digits. The model achieved a reasonable accuracy of 68.57% on this new dataset. However, it showed limitations in recognizing numbers not part of the training data, particularly those in non-standard formats.(Garg et al., 2019)

The paper delves deeper into the model's architecture, revealing that it includes four convolutional layers, ReLU activation, and max-pooling layers - a standard arrangement for picture classification tasks. This study highlights that CNNs are highly effective at recognizing handwritten digits and can generalize to previously unexplored data. However, the model's performance deteriorates when it encounters data that considerably differs from the training set.

The paper explores EEG-based emotion recognition, lever- aging Convolutional Neural Network (CNN) architectures to enhance subject-independent accuracy. Unlike conventional methods relying on spectral band power features, raw EEG data is utilized after windowing, pre-adjustments, and normalization, removing manual feature extraction and harnessing CNN's capacity to uncover hidden features(Cimtay and Ekmekcioglu, 2020). A median filter further improves classification accuracy. The approach achieves mean cross-subject accuracies of 86.56 and 78.34 on the SEED dataset for two and three emotion classes, re- respectively. Testing the SEED-trained model on the DEAP dataset yields a mean accuracy of 58.1.

The paper extensively evaluates CNN models in AI-assisted COVID-19 diagnostics, spotlighting ResNet-50 as the top performer. Through iterative rounds of training and testing across diverse datasets, the study underscores the critical importance of achieving subject-independent accuracy and the potential of enriching training datasets to bolster model performance. Leveraging heatmaps and activation features provides deeper insights into CNN model learning dynamics, guiding future advancements in COVID-19 and pneumonia detection diagnostic systems. During the initial evaluation round, CNN models exhibited high accuracy rates of 95.2 to 99.2 for the Level 1 testing dataset, sourced from the same clinic but designated solely for testing. However, model performance declined significantly with the Level 3 dataset, characterized by outlier images, reducing mean sensitivity from 99 to 36. These findings emphasize the challenges outlier data poses and the need for strategies to mitigate their impact on diagnostic model performance(Talaat et al., 2023).

This research gives a method for detecting biases in picture attribute estimations learned by convolutional neural networks (CNNs)(Zhang et al., 2018). Even with great overall accuracy, these biases might lead to erroneous findings. The method examines CNN's internal representation of characteristics to detect probable blind spots (missing associations) and failure modes (incorrect relationships) induced by biases in the training data. It does not require additional labeled data and provides a more thorough analysis than standard approaches. Experiments show that the strategy successfully detects bias and outperforms other ways of identifying problems with CNN's learned representations.

3 COPYRIGHT FORM

Three datasets are chosen for experimentation, consisting of grayscale images depicting handwritten digits and characters, with dimensions of 28x28x1. Only images containing digits 0-9 have been selected

for experimentation.

- **MNIST Dataset:** This dataset is a collection of handwritten digits with numbers 0-9. It contains 60,000 train images and 10,000 test images. The images are grayscale and have a resolution of 28x28 pixels. MNIST is often used as a benchmark to evaluate the performance of different machine learning models, particularly in deep learning and neural networks.(LeCun et al., 2010)

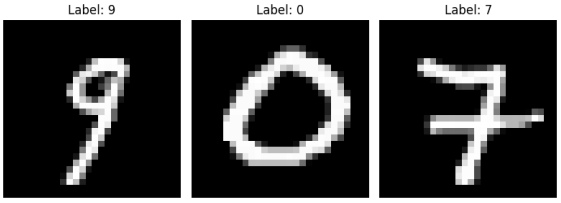

Figure 1: Samples from MNIST dataset

- **USPS Dataset:** The United States Postal Service dataset is similar to the MNIST dataset but consists of handwritten digits from the United States Postal Service. The USPS dataset may exhibit more variability in writing styles and quality compared to the MNIST dataset. This is because the images are scanned from real-world postal mail, which can contain a wide range of handwriting styles, variations in stroke thickness, and other factors not present in carefully collected datasets like MNIST(Bagul,).

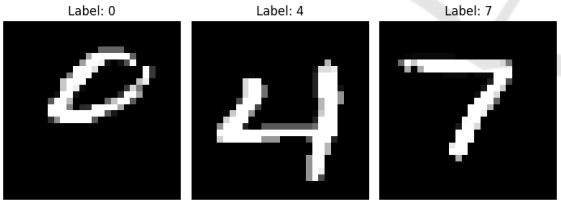

Figure 2: Samples from USPS dataset

- **EMNIST Dataset:** The extension of the MNIST dataset, providing a broader and more varied collection of handwritten digit samples. The EMNIST dataset contains handwritten characters from English alphabets (uppercase and lowercase) and digits (0-9). It consists of 814,255 characters, divided into 814,255 training images and 81,406 test images. Each image is grayscale with a resolution of 28x28 pixels.(Cohen et al., 2017)

The naming convention utilized for the datasets in our study is illustrated in Figure 4. This study employs Dataset A for training purposes, while all three datasets validate the trained model. Dataset B exhibits correlation and resembles Dataset A, whereas Dataset C is uncorrelated and notably different from Dataset A.

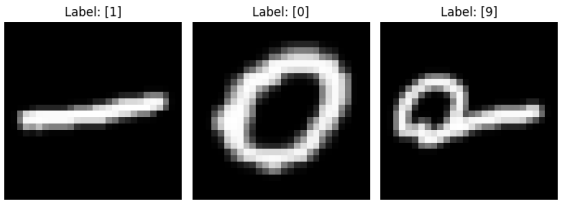

Figure 3: Samples from EMNIST dataset

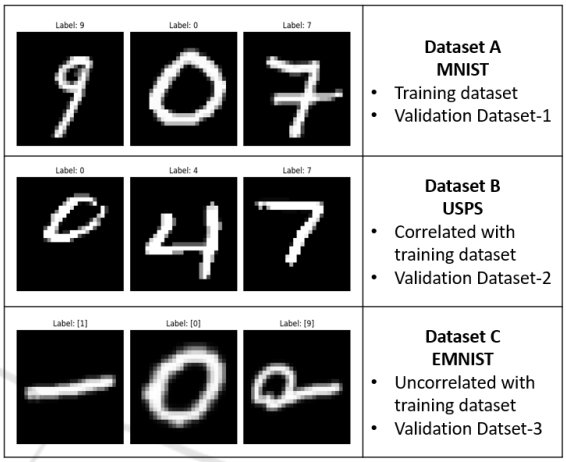

Figure 4: Naming convention used for datasets

Convolutional Neural Networks (CNNs) are highly effective for multi-class classification tasks, particularly on image datasets like MNIST. CNNs have many adjustable parameters that can influence the model's performance, like the number of layers, filter counts, dimensions and types, and learning rate.

CNN models are crafted by adjusting various parameters, then trained on Dataset A and validated on all three datasets. This methodology seeks to understand how these models perform when confronted with different combinations of datasets for the same task. This aids in gaining insight into the model's behaviour when encountering datasets with varying degrees of similarity.

The initial model designed for observation consists of two convolutional layers, a max-pooling layer, two additional convolutional layers, and another max-pooling layer, concluding with a fully connected layer, as depicted in Figure 2. Additionally, the model includes the following parameters:

- Learning rate set to 0.001
- Two dense layers with 64 units followed by 10 units.
- 32 filters in each convolution layer with a dimension of 3x3.

- ReLu activation function.
- Training spanned across 10 epochs.

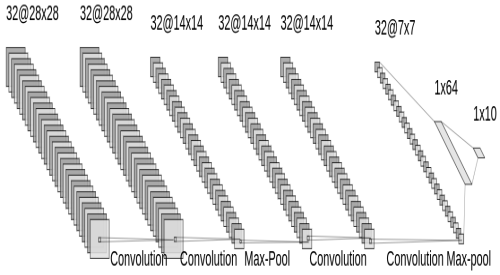

Figure 5: Model designed for experimentation

A range of CNN models were developed and examined by modifying their parameters to ascertain their impact on enhancing the models' performance.

3.0.1 Number of dense layers and convolution layers

The dense layers are incremented in multiples of 32, specifically 32, 64, 128, and 256, with each dense layer corresponding to either one or two convolution layers at a fixed learning rate of 0.001. Each convolution layer comprises 32 filters with dimensions of 3x3 each. This results in the formation of various combinations of dense and convolution layers.

3.0.2 Number of Convolution Layers

The experimentation involves varying only the convolution layers from 1 to 10 while keeping the dense layers constant at 64. Each convolution layer comprises 32 filters with dimensions of 3x3 each. This aims to observe how the model's behaviour evolves with increasing convolution layers.

3.0.3 Learning Rate

The learning rate is a critical factor in model training. The learning rate ranges from 0.05 to 0.001 to determine its impact on improving blind validation accuracy. The ideal learning rate is selected based on the results obtained.

3.0.4 Number of Filters and their Dimension

The filter dimensions are varied across 3x3, 5x5, and 7x7, each utilizing 16, 32, and 64 filters, respectively. It is carried out at a constant learning rate and fixed dense layers.

With these variations, it was evident that the model performance tended to deteriorate with unseen datasets; the same is described in the results section.

4 RESULTS AND DISCUSSIONS

The outcomes obtained from the initial model design indicate that the accuracy was significantly higher when Dataset A was trained and tested on itself compared to when tested on Dataset B and Dataset C, as depicted in Table 1.

Table 1: Cross-Dataset Performance of Model Trained on Dataset A

Train	Test	Validation Accuracy(%)
Dataset A	Dataset A	98.43
Dataset A	Dataset B	60.38
Dataset A	Dataset C	14.29

The following illustrates the outcomes of the different CNN models designed.

4.1 Varying Number of Dense Layers

Increasing the number of dense layers by multiples of 32, specifically 32, 64, 128, and 256, did not significantly enhance validation accuracy.

Nevertheless, as the number of convolution layers increased from 1 to 2, there was a corresponding enhancement in accuracy. Figure 6 illustrates that augmenting the number of dense layers has minimal impact on accuracy. Conversely, it is apparent that augmenting the number of convolution layers positively correlates with heightened accuracy.

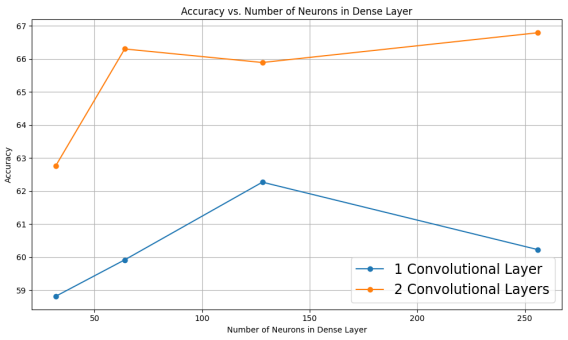

Figure 6: Performance with change in number of Dense Layers

4.2 Varying number of Convolution layers

The validation accuracy shows a noticeable rise with an increase in convolution layers, as illustrated in Figure 7. The mean accuracy shifts from 60 to 70%.

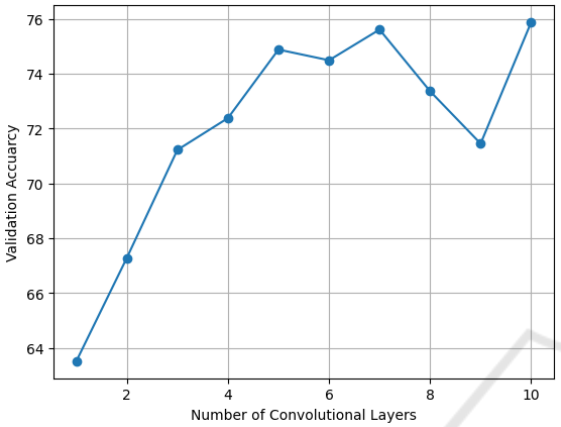

Figure 7: Performance with change in number of Convolution Layers

4.3 Varying Learning Rate

Figure 8 demonstrates that lower learning rates, like 0.001, consistently yield superior accuracy compared to higher and excessively lower rates. This observation underscores the critical significance of meticulously selecting an appropriate learning rate to enhance model performance effectively. However, despite varying learning rate, the maximum blind validation accuracy obtained is below 75%.

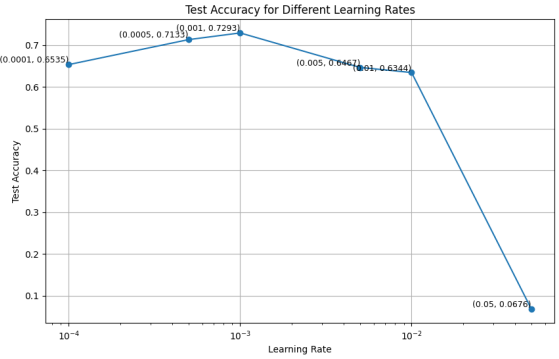

Figure 8: Performance with change in Learning rate

4.4 Varying Number of Filters and Dimension of Filters

Figure 9 indicates that despite experimenting with various filter combinations, there was no improvement in blind validation accuracy. Furthermore, the validation accuracy achieved by training and testing on Dataset A was notably higher than that achieved by training on Dataset A and testing on Dataset B.

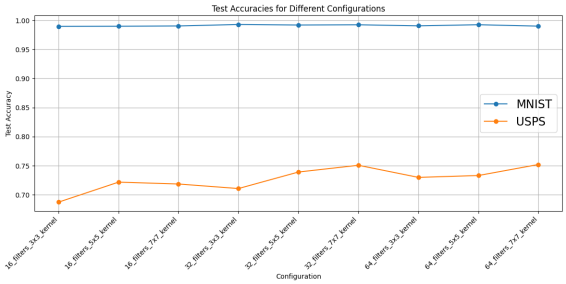

Figure 9: Performance with variation with different combinations of Filters

Based on the observed trends in validation accuracy improvement, the convolution layers were systematically increased with a consistent learning rate of 0.001 while maintaining 64 dense layers. This was in response to the observed trend of higher blind validation accuracy associated with an augmented number of convolution layers, as depicted in Figure 3 and Table 3.

Table 2 provides insights into the validation accuracy as convolution layers increase. The ”Layers” column indicates the number of convolution layers, ”MNIST” represents the validation accuracy on Dataset A, ”USPS” indicates the validation accuracy on Dataset B, ”EMNIST” signifies the validation accuracy on Dataset C, and ”Training parameters” denotes the total trainable parameters for each respective number of convolution layers.

Figure 10 represents the convolution layers’ corresponding validation accuracy. The blue line shows the training and testing done on Dataset A, while the orange line shows the training done on Dataset A and testing done on Dataset B. Similarly, the green line represents the training done on Dataset A and testing done on Dataset C.

The blind validation accuracy did not increase beyond 75%(approx.) when tested on Dataset B and not beyond 15% (approx.) on Dataset C. Conversely, the model achieved an accuracy of over 95% when trained and tested on Dataset A, indicating a decline in performance when exposed to different unseen datasets, despite adjustments parameters.

Table 2: Summary of experiment results

Layers	MNIST	USPS	EMNIST	Train Params
1	98.86%	56.73%	18.78%	1.61M
2	98.75%	62.03%	17.13%	1.64M
3	99.01%	67.51%	17.33%	1.68M
4	99.21%	68.46%	16.73%	1.72M
5	99.16%	66.73%	14.71%	1.76M
6	99.10%	65.63%	17.52%	1.79M
7	98.90%	67.66%	16.70%	1.83M
8	99.14%	68.92%	18.98%	1.83M
9	99.07%	69.07%	17.23%	1.90M
10	98.79%	67.89%	16.73%	1.94M

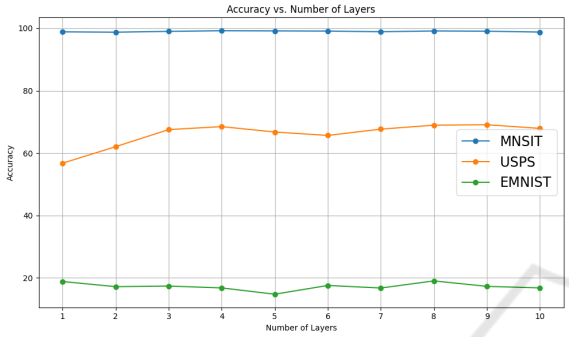

Figure 10: Performance with change in number of Convolution Layers for validating across the Datasets

5 CONCLUSIONS

The research addresses the crucial challenge of generalization in CNN models, revealing that although they achieve high accuracy on their training dataset (98.43% on MNIST), their performance significantly drops on unseen datasets (60.38% on USPS and 14.29% on EMNIST). This finding highlights the vital role of blind validation in evaluating how well machine learning models can adapt to new data, which is essential for their deployment in real-world scenarios. The study exposes the limitations of CNNs when faced with diverse datasets, stressing the necessity for robust validation techniques to ensure the models are reliable and effective outside of controlled training settings. The significance of this study lies in illustrating the importance of thoroughly evaluating any pre-trained model before it is deployed in real-world applications, where adaptability and robustness are crucial for maintaining consistent and reliable performance.

REFERENCES

Bagul, D. USPS_Digit_Classification. https://github.com/darshanbagul/USPS_Digit_Classification.

Cimtay, Y. and Ekmekcioglu, E. (2020). Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset eeg emotion recognition. *Sensors*, 20(7):2034.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. (2017). Emnist: an extension of mnist to handwritten letters.

Dai, D. (2021). An introduction of cnn: models and training on neural network models. In *2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR)*, pages 135–138. IEEE.

Garg, A., Gupta, D., Saxena, S., and Sahadev, P. P. (2019). Validation of random dataset using an efficient cnn model trained on mnist handwritten dataset. In *2019 6th international conference on signal processing and integrated networks (SPIN)*, pages 602–606. IEEE.

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Talaat, M., Si, X., and Xi, J. (2023). Multi-level training and testing of cnn models in diagnosing multi-center covid-19 and pneumonia x-ray images. *Applied Sciences*, 13(18):10270.

Zhang, Q., Wang, W., and Zhu, S.-C. (2018). Examining cnn representations with respect to dataset bias. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.