# Clothing Items Classification: Assessing the Influence of Model Complexity on Performance

Antonio Kuran
Faculty of Electrical Engineering and Computing
FER
Zagreb, Croatia
antonio.kuran@fer.hr

Filip Grabovac
Faculty of Electrical Engineering and Computing
FER
Zagreb, Croatia
filip.grabovac@fer.hr

Polina Rykova
Faculty of Electrical Engineering and Computing
FER
Zagreb, Croatia
polina.rykova@fer.hr

Borna Rebić Taučer
Faculty of Electrical Engineering and Computing
FER
Zagreb, Croatia
borna.rebic-taucer@fer.hr

Lovro Đuranec
Faculty of Electrical Engineering and Computing
FER
Zagreb, Croatia
lovro.duranec@fer.hr

***Abstract* — This research explores the relationship between structural depth in neural networks and classification accuracy for low-resolution fashion imagery. A comparative analysis is performed between the classic LeNet-5 architecture and a modern, high-capacity custom Convolutional Neural Network (CNN) using the Fashion-MNIST dataset. Despite significant differences in the number of parameters and model depth, both architectures achieve nearly identical test accuracy by standard metrics. Such performance indicates that for standardized and visually simplified datasets, the inherent statistical consistency allows efficient, shallow networks to perform as effectively as significantly more complex alternatives.**

***Keywords*** — Convolutional Neural Networks, LeNet-5, Fashion-MNIST, Model Complexity, Architectural Efficiency, Image Classification, Deep Learning.

## I. Introduction

Image classification remains the standard tool for testing how well a neural network actually performs. While the current trend in research is to build deeper and more complex models to chase higher precision, it is worth asking if that complexity is always justified, especially for low-resolution images. This study explores that trade-off by putting the classic LeNet-5 architecture up against a much larger, high-parameter custom CNN. Since Fashion-MNIST consists of simple, standardized grayscale images, the goal is to determine if a shallow network can capture the necessary details just as effectively as a much deeper alternative..

## II. Related Work

### *A. Foundational Architecture*

The LeNet-5 architecture, introduced by LeCun et al. [1], established the fundamental design of modern Convolutional Neural Networks. By utilizing convolutional layers and spatial subsampling, it achieved high computational efficiency in character recognition. Despite its age, LeNet-5 remains a relevant baseline because it can extract essential spatial features with a minimal number of parameters.

### *B. Existing comparisons*

A key example of research in this field is the work of Xiao et al. (2017), who introduced the Fashion-MNIST dataset as a more challenging alternative to traditional digit recognition. Their work established benchmarks across various machine learning models to identify the performance limits of low-resolution grayscale imagery. These benchmarks show that while convolutional architectures are highly effective at identifying apparel, the structural similarity between certain categories creates a consistent challenge for classification.

The data suggests that the classification of low resolution grayscale items is heavily influenced by the resolution of the source material. In these standardized datasets, models are tasked with identifying diagnostic features from a limited number of pixels. Research indicates that once an architecture reaches the capacity to extract these basic spatial patterns, the primary factor limiting further accuracy is the similarity of the items themselves at such a small scale.

## III. Technical Approach

### *A. Architectural Design*

Two Convolutional Neural Networks (CNNs) with significantly different capacities were implemented to identify the difference in performance.

- LeNet-5

  This architecture follows the classical design introduced for character recognition. It utilizes two convolutional layers with 5×5 filters and average pooling to extract low-level spatial features. With approximately 61,000 parameters, it represents a lightweight approach focused on computational efficiency and is used as baseline for this experiment.

- Custom CNN

  This model was designed to test the benefits of increased depth and non-linearity. It features a four-layer convolutional backbone with expanding filter dimensions of 32, 64, and 128 along with Max Pooling layers. By utilizing the Rectified Linear Unit (ReLU) activation function and over 1,000,000 parameters, this model provides significantly higher representative power to capture abstract patterns in the clothing data.

### *B. Data Preparation and Normalization*

The Fashion-MNIST dataset consists of 60,000 training images and 10,000 testing images. The input data was normalized using a mean and standard deviation of 0.5, scaling pixel intensities to a range of $[-1,1]$. This preprocessing step ensures that the gradients remain stable during backpropagation.

### C. *Training Procedure*

The training phase was optimized using a grid search to identify the most effective hyperparameters for each architecture. Both models were trained for 20 epochs using the Cross-Entropy Loss function and Stochastic Gradient Descent (SGD). While both utilized a momentum of 0.9 and a batch size of 64, the grid search determined that a learning rate of 0.001 was optimal for LeNet-5, whereas the Custom CNN required 0.002 to maintain stability. This targeted tuning ensured that the final comparison reflected the maximum potential of each design.

### D. *Evaluation Strategy*

Model performance was quantified using accuracy as the primary indicator of effectiveness, as well as precision, recall, and F1-score to ensure a balanced assessment across all categories. To investigate specific failure modes, confusion matrices were generated to identify where structural similarities between items resulted in misclassification. Additionally, training and validation loss curves were monitored to confirm that neither architecture suffered from significant overfitting, ensuring the results reflect the true generalizability of the models.

## IV. Experimental Results

### A. *Primary Metrics Comparison*

The experimental results indicate a clear performance saturation point. Despite the significant disparity in architectural depth and parameter count, both models converged at nearly identical levels of performance as seen in “Fig. 1”.

TABLE I. Primary Metrics Evaluations

| | **Primary Metrics Results** | | |
|---|---|---|---|
| | ***Metric*** | ***LeNet5*** | ***Custom CNN*** |
| | Accuracy | 89.91% | 88.85% |
| | Precision | 89.96% | 88.82% |
| | Recall | 89.91% | 88.85% |
| | F1Score | 89.85% | 88.72% |

Fig. 1. LeNet5 and Custom CNN primary metrics results

### B. *Confusion matrix*

A detailed comparison of the confusion matrices for both architectures reveals nearly identical error patterns, indicating that misclassifications are driven by data characteristics rather than model design. As shown in “Fig. 2” and “Fig. 3”, the errors are heavily concentrated in specific clusters where structural similarity is highest, such as the T-shirt, Pullover, and Shirt categories which share nearly identical silhouettes. The fact that the million-parameter Custom CNN could not resolve these specific confusions more effectively than the 61,000-parameter LeNet-5 baseline confirms that the source of confusion is the variance within the Fashion-MNIST dataset.

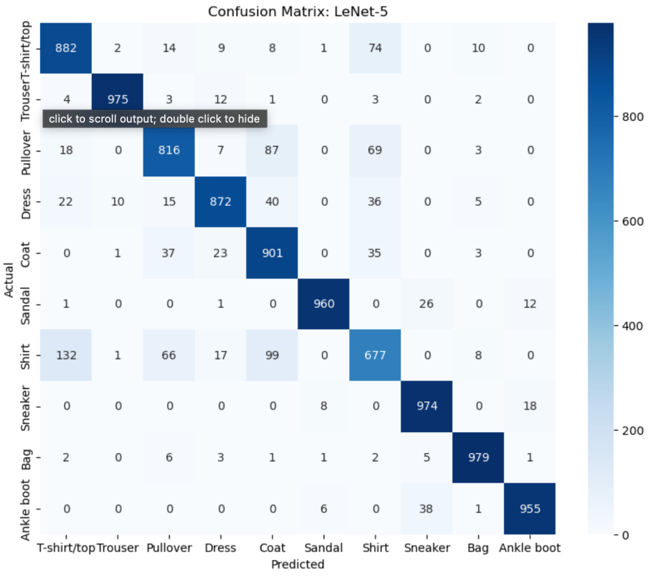

Fig. 2. Confusion matrix for LeNet-5

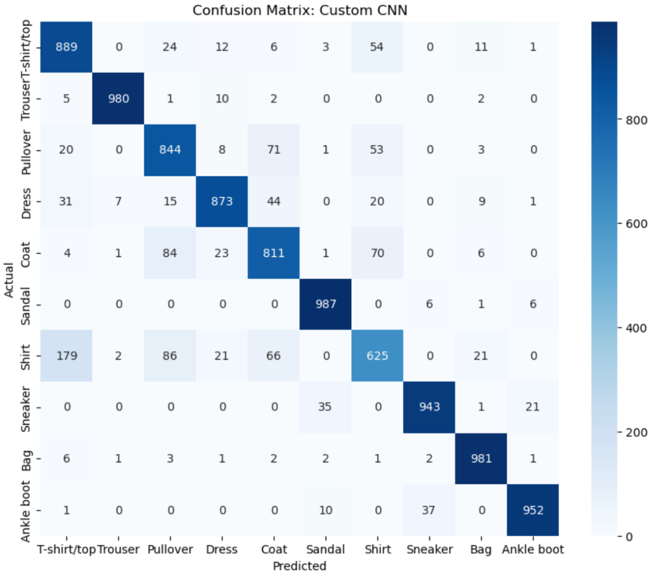

Fig. 3. Confusion matrix for Custom CNN

### C. *ROC curve*

The Receiver Operating Characteristic (ROC) curves provide further insight into the models' ability to distinguish between categories. Both architectures exhibit high Area Under the Curve (AUC) values for distinct items like Trouser and Bag, while lower AUC values are consistent for the Shirt category. This visual trade-off between True Positives and False Positives reinforces the finding that certain fashion categories are statistically harder to separate at low resolutions, regardless of the network depth. As shown in “Fig. 4” and “Fig. 5”, the class-wise performance is nearly mirrored across both models, suggesting that the complexity of the Custom CNN does not improve the separability of visually ambiguous classes.

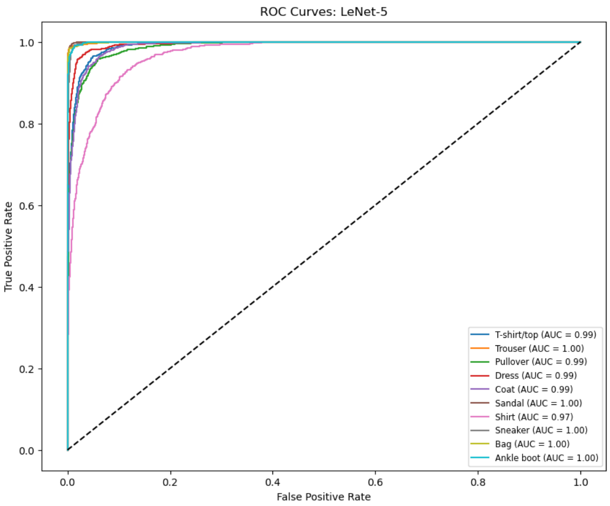

Fig. 4. ROC curve for LeNet-5

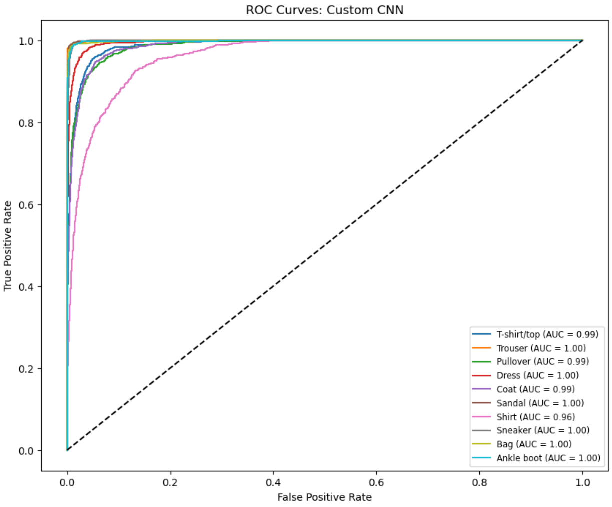

Fig. 5. ROC curve for Custom CNN

Overall results prove that a well-matched, smaller model can be just as powerful as a much larger one for this type of data.

## V. Discussion and Comparison with Literature

The comparison of the results obtained in this study with the benchmark research conducted by Xiao et al. (2017) confirms a broader trend in the classification of the Fashion-MNIST dataset. In the original study, the authors noted that standard convolutional models achieve a high level of accuracy that is difficult to significantly surpass even with drastic increases in model complexity. The results of this experiment, where both LeNet-5 and the Custom CNN achieved very similar accuracy, are in alignment with previous study results. Although the Custom CNN utilizes modern ReLU activation functions and a substantially larger number of parameters, it provides no advantage in precision. This indicates that both models reached an information ceiling regarding the features that can be extracted from low resolution imagery. Those observations are supported by the fact that both models fail on the same classes, specifically shirts, coats, and pullovers, confirming that the limitation is not the architecture. Instead, the limitation is the structural similarity of the items which becomes ambiguous in a low resolution grayscale format.

## VI. Conclusion

This study demonstrated that both historical models, such as LeNet-5, and modern architectures are highly successful in classifying the Fashion-MNIST dataset, achieving strong results across standard metrics. The primary finding of this research is that an increase in the number of layers and parameters does not inherently guarantee improved performance on standardized, low-resolution datasets. In this experiment, LeNet-5 proved to be an exceptionally efficient model, achieving slightly higher accuracy than a much more complex network while utilizing significantly fewer computational resources. The results suggest that both architectures reached an information ceiling imposed by the structural similarity of the items and the low-resolution grayscale format of the data. Consequently, additional architectural complexity does not resolve the inherent ambiguity between similar classes at this scale. Future efforts to break this performance plateau should prioritize data-centric approaches, such as advanced data augmentation or specialized image preprocessing, rather than simply increasing the depth of the neural networks.

## VII. Literature

[1] 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, Nov. 1998. [Online]. Available: http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

[2] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” 2017, arXiv: 1708.07747. [Online]. Available: https://arxiv.org/abs/1708.07747

[3] PyTorch Tutorial: LeNet-5 Implementation, "Training a Classifier on Fashion-MNIST," PyTorch Documentation. [Online]. Available: https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html

[4] Fashion-MNIST Benchmark Hub, "Leaderboard and Implementation Examples," Zalando Research. [Online]. Available: https://github.com/zalandoresearch/fashion-mnist