Comparative Analysis of CNN and CNN-SVM Models for MNIST Digit Classification

Dagemawi Bekele Negash

August 13, 2025

Abstract

This report details the comparative performance of two machine learning models for the task of handwritten digit recognition on the MNIST dataset. We evaluate a standalone Convolutional Neural Network (CNN) against a hybrid model that uses the same CNN as a feature extractor and a Support Vector Machine (SVM) as the classifier. The analysis focuses on key performance metrics, computational efficiency, and error patterns to determine the strengths and weaknesses of each approach.

1 Introduction

The MNIST dataset is a standard benchmark for image classification tasks. While Convolutional Neural Networks (CNNs) are highly effective at learning spatial hierarchies of features from images, traditional classifiers like Support Vector Machines (SVMs) excel at finding optimal decision boundaries in high-dimensional spaces. This report investigates a hybrid approach, leveraging the CNN's feature extraction power and the SVM's robust classification capabilities.

2 Methodology

A simple CNN was first trained for 5 epochs on the full MNIST training set (60,000 images). This trained CNN was then used in two ways:

1. **Standalone CNN Model:** The complete network, including its final fully-connected layers, was used for classification.
2. **Hybrid CNN-SVM Model:** The convolutional base of the trained CNN was used to extract a 3136-dimensional feature vector for each image. An SVM with a Radial Basis Function (RBF) kernel was then trained on a subset of these features (15,000 samples) to perform the final classification.

Both models were evaluated on the full MNIST test set of 10,000 images.

3 Results and Analysis

3.1 Performance Metrics

The overall performance of both models was evaluated using Accuracy, Precision, Recall, and F1-Score. Table 1 and Figure 1 summarize these results.

Table 1: Performance Metrics Summary

Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
CNN	0.9917	0.9917	0.9916	0.9917
CNN+SVM	0.9803	0.9813	0.9801	0.9805

The standalone CNN achieved a slightly higher overall performance across all major metrics, with an accuracy of **99.17%** compared to the hybrid model's **98.03%**. The visual comparison in Figure 1 clearly shows the CNN's marginal advantage.

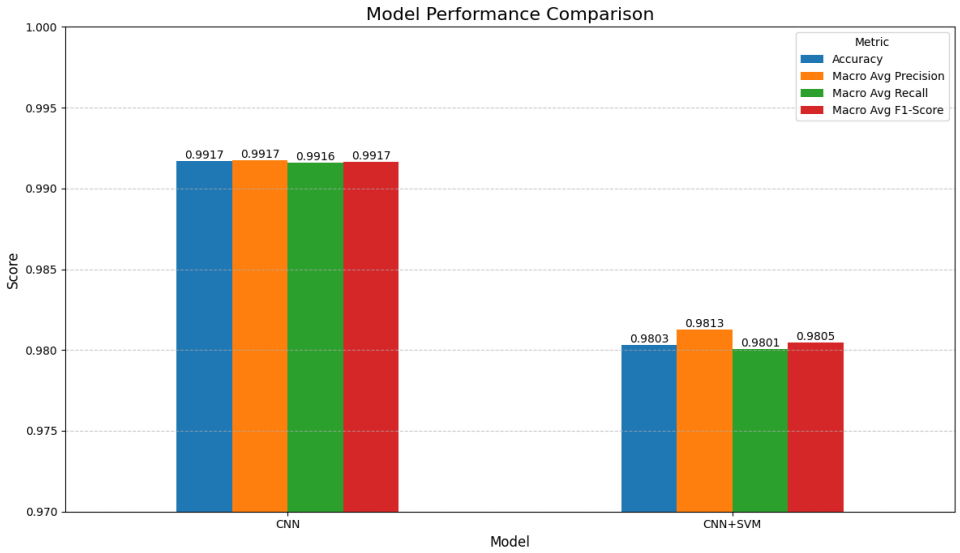

Figure 1: Visual comparison of key performance metrics for both models.

3.2 Error Analysis with Confusion Matrices

To understand the specific error patterns, confusion matrices for both models were generated (Figure 2).

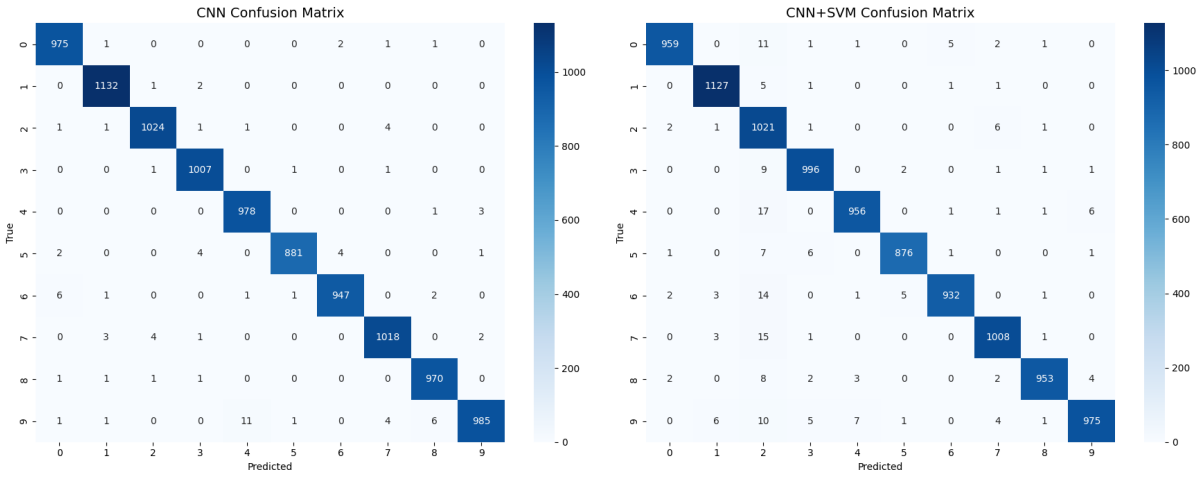

Figure 2: Confusion matrices for the standalone CNN (left) and the hybrid CNN-SVM (right).

Analysis:

- **CNN Model:** The standalone CNN is highly accurate across all classes. Its most frequent error is misclassifying the digit '9' as '4' (11 instances), a common confusion due to their structural similarity.
- **CNN-SVM Model:** The hybrid model shows a noticeable weakness in classifying the digit '2', with a precision of only 91.4%. It frequently confuses '2' with '4' and '7'. It also struggles with distinguishing '4' from '9' (7 instances).

This comparison highlights that while both models struggle with visually similar digits like '4' and '9', the hybrid CNN-SVM introduces a new, more significant weakness in classifying the digit '2'.

3.3 Computational Efficiency

The training time for both models was measured to compare their efficiency. The hybrid model's time is the sum of feature extraction and SVM training.

Table 2: Efficiency Metrics Summary

Model	**Training Time (s)**
CNN	97.68
CNN+SVM (Feature Extraction + SVM fit)	127.20

The standalone CNN was significantly faster to train (**97.68s**) than the full hybrid pipeline (**127.20s**). The majority of the hybrid model's time was spent training the SVM on the 15,000 feature vectors.

4 Conclusion

In this experiment, the standalone CNN model outperformed the hybrid CNN-SVM model in both predictive accuracy (99.17% vs. 98.03%) and training efficiency. The CNN demonstrated a more balanced performance across all digit classes, whereas the SVM classifier struggled notably with certain digit pairs like '2' and '7'.

This outcome suggests that for this specific architecture and dataset, the CNN's end-to-end training allowed its fully-connected layers to create decision boundaries that were more finely tuned to its own extracted features than the more general-purpose SVM classifier. However, the hybrid approach remains a valuable technique, especially in scenarios where a pre-trained feature extractor is available and classifier training must be done quickly on smaller datasets.