Analysis of Classical Ensemble Methods on Fashion-MNIST

Machine Learning Project

Matej Miškovčík

Introduction

Image classification is a fundamental task in machine learning and is frequently used to study the strengths and limitations of different learning algorithms. Although deep convolutional neural networks currently achieve state-of-the-art performance on most image-based tasks, classical machine learning models remain valuable for educational and analytical purposes, particularly when the goal is to understand model behavior rather than to maximize accuracy.

The objective of this project is to analyze how classical tree-based ensemble methods behave when applied to image classification using raw pixel representations. Instead of focusing solely on performance optimization, the emphasis is placed on identifying systematic error patterns and understanding the underlying reasons for model failures. The Fashion-MNIST dataset is used as a standardized benchmark that provides a controlled environment for studying visually similar object classes.

As a baseline, a Random Forest classifier was trained on raw image pixels. While the model achieved relatively high overall accuracy, a closer inspection revealed consistent misclassifications for certain classes. In particular, class 6 (shirt) was frequently confused with visually similar garment categories such as t-shirt/top, pullover, and coat. Notably, many of these errors were associated with high prediction confidence, indicating systematic behavior rather than random noise.

Based on this observation, several follow-up experiments were designed to investigate whether these systematic errors could be reduced through model-level modifications. Principal Component Analysis (PCA) was applied as a dimensionality reduction technique to test whether removing redundant or noisy pixel information would improve class separability. In addition, an Extremely Randomized Trees (ExtraTrees) classifier was evaluated to examine whether increased model randomization and reduced overfitting would alter the observed error patterns.

Rather than performing extensive hyperparameter searches or comparing a large number of models, this work follows a trial-error-analysis development cycle. Each experiment is motivated by prior error analysis and evaluated using both standard performance metrics and a detailed examination of high-confidence misclassifications. The results highlight the limitations of pixel-based representations and illustrate why certain visually similar classes remain challenging for classical ensemble methods.

Dataset and Experimental Setup

The experiments were conducted using the Fashion-MNIST dataset, a widely used benchmark for image classification. The dataset consists of 70000 grayscale images of fashion products, each with a resolution of 28×28 pixels, grouped into ten clothing categories. Compared to the original MNIST dataset, Fashion-MNIST presents a more challenging classification task due to the visual similarity between several classes.

Each image was represented as a flattened vector of 784 pixel intensity values, corresponding to the raw grayscale pixel representation. No handcrafted feature extraction was applied in the baseline experiments, allowing the study to focus on the limitations of classical machine learning models operating directly on pixel-level inputs.

To ensure reproducibility and prevent data leakage, the dataset was split into three disjoint subsets: training, validation, and test sets. The training set was used exclusively for model fitting, the validation set for intermediate evaluation and experimental comparison, and the test set for final performance assessment. The split proportions were fixed at 64% for training, 16% for validation, and 20% for testing, with stratification applied to preserve class balance across all subsets. The indices defining each split were stored and reused across all experiments to guarantee consistent and fair comparisons.

All experiments were implemented in Python using the scikit-learn library. Model training and evaluation were performed using fixed random seeds, and all configurations and results were automatically saved to enable full reproducibility of the experimental pipeline.

Methods

This section describes the machine learning models and experimental procedures used in this study. All experiments were conducted using identical data splits and evaluation protocols to ensure fair and reproducible comparisons.

Baseline Model: Random Forest on Raw Pixels

As a baseline, a Random Forest classifier was trained using raw pixel representations of the input images. Each image was flattened into a 784-dimensional feature vector corresponding to its grayscale pixel intensities. The Random Forest model was selected due to its robustness, interpretability, and widespread use as a classical ensemble method.

The baseline model was trained on the training subset and evaluated on both the validation and test sets. No dimensionality reduction or feature preprocessing was applied at this stage, allowing the baseline to capture the inherent limitations of pixel-based representations for visually similar classes.

Dimensionality Reduction with Principal Component Analysis

To investigate whether reducing feature dimensionality could improve class separability, PCA was applied prior to classification. PCA projects the original high-dimensional input space onto a lower-dimensional subspace that preserves the directions of highest variance.

In this experiment, PCA was fitted exclusively on the training data to prevent information leakage, and the learned transformation was subsequently applied to the validation and test sets. The transformed feature vectors were then used as input to the same Random Forest classifier as in the baseline. This experiment tested the hypothesis that removing redundant or noisy pixel-level information could reduce systematic misclassifications between visually similar garment categories.

Increased Model Randomization: ExtraTrees Classifier

A second experimental modification replaced the Random Forest classifier with an ExtraTrees classifier. ExtraTrees is a tree-based ensemble method similar to Random Forests but introduces additional randomness by selecting split thresholds at random rather than optimizing them at each node.

This increased randomization is known to reduce overfitting and can lead to smoother decision boundaries. The ExtraTrees classifier was trained using the same raw pixel features and data splits as the baseline model, allowing the effects of increased model randomization to be isolated and evaluated independently of feature representation.

Evaluation Metrics and Error Analysis

Model performance was evaluated using standard classification accuracy on the validation and test sets. In addition, confusion matrices were computed to analyze class-wise error patterns and identify systematic misclassifications.

To further investigate model behavior, a confidence-based error analysis was performed. For each prediction, the model's confidence was defined as the maximum predicted class probability. Misclassified samples were ranked by confidence, and high-confidence errors were identified using a percentile-based threshold. This approach enabled the analysis of cases where the model was highly confident yet incorrect, providing insight into systematic failure modes beyond aggregate performance metrics.

Results

This section presents the experimental results obtained from the baseline Random Forest model and the two subsequent experimental variants. Model performance is reported using classification accuracy on the validation and test sets, followed by a detailed analysis of class-wise errors and high-confidence misclassifications.

Overall Classification Performance

Table 1: *Overall classification performance*

Model	Validation Accuracy	Test Accuracy
Random Forest (baseline)	0,8769	0,8804
PCA (200) + Random Forest	0,8610	0,8647
ExtraTrees	0,8801	0,8792

Table 1 summarizes the overall classification accuracy for all evaluated models. The baseline Random Forest classifier achieved a validation accuracy of 0,8769 and a test accuracy of 0,8804, establishing a strong reference performance on the Fashion-MNIST dataset.

Applying Principal Component Analysis with 200 components resulted in a noticeable decrease in performance, with validation and test accuracies dropping to 0,8610 and 0,8647, respectively. In contrast, the ExtraTrees classifier achieved accuracy values comparable to the baseline model, with a validation accuracy of 0,8801 and a test accuracy of 0,8792. Overall, neither experimental modification led to a clear improvement in classification accuracy over the baseline Random Forest model.

Confusion Matrix Analysis

Analysis of the confusion matrices revealed consistent error patterns across all evaluated models. In particular, class 6 (shirt) was frequently misclassified as classes 0 (t-shirt/top), 2 (pullover), and 4 (coat). This pattern was observed for the baseline Random Forest model as well as for both experimental variants, indicating that these misclassifications are systematic rather than model-specific.

Other classes exhibited substantially fewer misclassifications and did not show similarly pronounced confusion patterns. The persistence of the same dominant error structure across models suggests that the observed failures are strongly tied to visual similarities between garment categories.

High-Confidence Error Analysis

Table 2: *High-confidence error analysis (90th percentile)*

Model	Split	Confidence Threshold	HC Wrong (Total)	HC Wrongs (Class 6)
Random Forest	Validation	0,7207	138	60
Random Forest	Test	0,7433	171	68
PCA (200) + RF	Validation	0,5933	162	67
PCA (200) + RF	Test	0,6100	197	82
ExtraTrees	Validation	0,7100	137	63
ExtraTrees	Test	0,7260	170	73

To further examine model behavior, misclassified samples were analyzed based on prediction confidence. Confidence was defined as the maximum predicted class probability for each sample. Two complementary perspectives were considered.

First, a percentile-based analysis was employed to compare the relative confidence of incorrect predictions across models. In this setting, high-confidence errors were defined as the top 10% most confident misclassifications within each model. This approach ensures a sufficient number of samples for qualitative comparison but does not represent an absolute notion of high confidence. Instead, it highlights the most confident errors relative to each model's overall confidence distribution.

Using this relative criterion, the PCA-based model exhibited lower confidence thresholds than the baseline Random Forest and ExtraTrees models, indicating that its incorrect predictions were generally associated with lower confidence values. However, due to its lower overall accuracy, the PCA-based model produced a larger number of misclassifications, which in turn resulted in higher counts of percentile-based high-confidence errors, particularly for class 6. This behavior reflects the relative nature of the percentile-based definition rather than increased absolute confidence.

To complement this analysis, absolute confidence thresholds were also considered. When examining incorrect predictions with confidence greater than 0,9 on the test set, the baseline Random Forest produced 35 such errors, the ExtraTrees model produced 23, and the PCA-based model produced only 7. This indicates that while PCA increases the total number of errors, it substantially reduces the occurrence of highly confident incorrect predictions in absolute terms.

Across both analyses, class 6 consistently accounted for the largest proportion of incorrect predictions, including those with high relative and absolute confidence. This observation further supports the conclusion that misclassification of class 6 is driven by intrinsic visual similarity to other garment classes rather than by model overconfidence alone.

Summary of Results

Across all experiments, the baseline Random Forest, PCA-enhanced model, and ExtraTrees classifier exhibited similar misclassification behavior. Dimensionality reduction via PCA resulted in lower overall accuracy and increased error counts, while increased model randomization through ExtraTrees did not substantially alter either performance or error structure. High-confidence error analysis consistently identified class 6 as the dominant source of systematic misclassification, primarily involving confusion with visually similar classes 0, 2, and 4.

Discussion

The experimental results highlight several important aspects of applying classical machine learning models to image classification tasks based on raw pixel representations. Although all evaluated models achieved relatively high overall accuracy, detailed error analysis revealed persistent and systematic misclassifications that were largely unaffected by the tested modifications.

A central observation across all experiments is the consistent confusion involving class 6 (shirt), most frequently misclassified as classes 0 (t-shirt/top), 2 (pullover), and 4 (coat). This pattern remained stable across the baseline Random Forest model, the PCA-based variant, and the ExtraTrees classifier. The persistence of this error structure suggests that the dominant failure mode is not primarily caused by model overfitting or insufficient regularization, but rather by limitations in the underlying feature representation.

The PCA-based experiment demonstrates this limitation particularly clearly. While dimensionality reduction lowered the model's confidence in incorrect predictions, it also resulted in a noticeable decrease in classification accuracy. This indicates that the principal components retained by PCA capture global variance in the data but do not preserve the fine-grained, spatially localized information required to distinguish between visually similar garment categories. Consequently, PCA reduces overall certainty without resolving the core ambiguity between classes.

Similarly, replacing the Random Forest with an ExtraTrees classifier did not substantially alter either accuracy or error patterns. Despite the increased randomization and reduced tendency to overfit, the ExtraTrees model exhibited nearly identical confusion behavior, including a comparable number of high-confidence misclassifications for class 6. This further supports the conclusion that the observed errors are not driven by overly precise decision boundaries but by insufficiently discriminative input features.

The confidence-based analysis provides additional insight into model behavior. While percentile-based thresholds are useful for relative comparison across models, they do not represent an absolute notion of confidence. Complementary analysis using a fixed confidence threshold revealed that the PCA-based model produced substantially fewer highly confident incorrect predictions, despite having lower overall accuracy. This highlights a trade-off between predictive performance and confidence calibration, emphasizing that lower confidence does not necessarily imply improved classification quality.

Overall, the results indicate that classical ensemble methods operating on raw pixel features struggle to reliably separate classes with subtle visual differences. Improvements in model architecture or regularization alone are insufficient to address this issue. Instead, the findings point toward the importance of more expressive feature representations that capture local and spatial patterns within images.

Conclusion

This project explored the applicability of classical tree-based ensemble methods to image classification using raw pixel representations. A Random Forest baseline was evaluated on the Fashion-MNIST dataset and subsequently extended through dimensionality reduction with PCA and increased model randomization using an ExtraTrees classifier. While all models achieved comparable overall accuracy, none of the experimental modifications led to a meaningful improvement in classification performance.

The results indicate that the primary limitation of the evaluated approaches lies in the representation of the input data rather than in model capacity or regularization. Systematic misclassifications persisted across all experiments, particularly for visually similar garment categories, demonstrating that raw pixel features do not provide sufficient discriminative information for fine-grained class separation in this setting.

Future work should therefore focus on improving feature representations rather than further modifying tree-based ensemble models. Promising directions include the use of hand-crafted image descriptors that capture local structure, such as edge or gradient-based features, as well as models that preserve spatial relationships within images. In addition, hierarchical classification strategies could be considered to explicitly account for groups of visually similar classes. Finally, confidence calibration and uncertainty-aware prediction mechanisms may further improve model reliability in scenarios where ambiguous visual inputs are unavoidable.

In summary, this project highlights the importance of aligning feature representations with the structural properties of image data and demonstrates the limits of classical ensemble methods when applied directly to raw pixel inputs.

Reproducibility and Code Availability

All experiments presented in this project are fully reproducible. The complete source code, configuration files, and data preprocessing scripts are publicly available in a GitHub repository. The repository includes all model training scripts, fixed data splits, and configuration files required to reproduce the reported results.

The experimental pipeline is designed to ensure reproducibility through the use of fixed random seeds and stored train, validation, and test indices. Detailed instructions for running the experiments and reproducing the results are provided in the repository's README file.

The code repository is available at: **https://github.com/BushDemon/ml-projekt**