When AI Meets Fashion: Multiclass Trend Prediction Using Fashion-MNIST

Zoe Marshall [*1] **Bibek Dhungana** [*2] **Saanjali Majmundar** [*3] **Lingxin Chen** [*2] **Aliya Eissa** [*4]

Abstract

Clothing classification supports applications such as automated inventory management, clothing search filters, virtual try-on tools, and future virtual outfit builders. The Fashion-MNIST dataset is a foundational dataset with improved complexity for being able to train and test machine learning algorithms for these applications. In this paper, we analyze the use of Principal Component Analysis (PCA) alongside multiclass logistic regression via One-vs-Rest strategy and softmax logistic regression for clothing classification of the Fashion-MNIST dataset. In addition, we developed a full pipeline for image capturing, condensing, and processing real-world data captured from a camera. Our results show that PCA yields little performance change for softmax regression, while O-v-R model saw moderate improvements using PCA and decreasing training time for both. However, classification accuracy showed substantial decrease under rotation, indicating a need for data augmentations to improve robustness.

1. Introduction

Accurate clothing classification is useful for automating repetitive tasks such as inventory management and search filtering. While trivial for human interpretation, automated classification can be difficult due to significant overlap of item structure. For example, a shirt and top, or a dress and coat, have many structural attributes that are the same. (Xiao et al., 2017; Franceschini, 2021). The Fashion-MNIST dataset is an expansive dataset that addresses these concerns by supplying labels that support algorithm learning of these differences. Because of its accessibility and difficulty relative to the original MNIST dataset, Fashion-MNIST has become a common benchmark for developing and validating machine-learning algorithms.

Beyond its role as a benchmark, clothing classification has potential for broader applications. A reliable classifier could support digital outfit recommendation systems, automated wardrobe cataloging and inventory, virtual try-on tools, or early-stage generative design tools that assemble or propose outfits automatically. More advanced systems might eventually learn practical attributes, such as which items tend to work better for warmer or colder external temperatures, or which clothing items are typically worn in certain seasons. Yet the ability to use this dataset and its testing can be limited due to lack of connection to real-world applications.

In this project, we explore the effectiveness of dimensionality reduction prior to classification within a real-time Fashion-MNIST classification workflow. Dimensionality-reduction techniques such as Principal Component Analysis (PCA) can help reduce computational cost and mitigate overfitting, even for relatively small images. PCA projects data into a lower-dimensional space that captures the dominant variance, enabling faster training while preserving much of the discriminative structure relevant for classification tasks. We compare multiclass logistic regression using a One-vs-Rest schedule with softmax regression through PCA-reduced and the original pixel space. We then evaluated both accuracy and training time to assess the benefits and trade-offs of applying PCA to this dataset. We further investigated limitations to our model by evaluating rotations, which may occur for these various real-world applications. To bridge our training to real-world use cases, we developed a full pipeline for transforming images captured from a camera to the MNIST dimension, which was primarily used for the evaluation of the model in the real domain.

1.1. Related Work

Several studies have approached Fashion-MNIST classification from different angles. Some focus solely on PCA's effect on reconstruction without applying any classifiers (Arenas, 2021), others evaluate classification models without dimensionality reduction (Xiao et al., 2017), and still others combine PCA with various classifiers (Fashion-MNIST). Across independent research efforts and Kaggle competitions (Franceschini, 2021; Fashion-MNIST), multiple attempts have been made to classify the Fashion-MNIST dataset both with and without PCA. We are adding on to these attempts by using a different classifier and by incorporating real-world data with a live classification demonstration.

2. Methods

2.1. Dataset

We conducted supervised learning for multiclass classification on the MNIST Fashion dataset, consisting of 60,000 training samples and 10,000 test samples of clothing items. Each sample is a 28×28 grayscale image associated with one of 10 clothing labels. We also added 39 custom samples to our training set using real objects collected by our team members with a camera. We first converted the images to grayscale, removed the background, and flattened them into 784-dimensional vectors for the training dataset. These samples were included in the training set to evaluate how well the model could classify real objects from the camera feed in our web application and further diversify the data.

2.2. Dimensionality Reduction

Dimensionality reduction was performed using Principal Component Analysis (PCA) (Hotelling, 1933) to reduce feature redundancy and improve model efficiency in terms of training and prediction time. Different numbers of principal components (20, 50, 100, 150, and 200) were kept to observe the impact on classification accuracy. For all PCA comparisons, the models were trained with a learning rate = 0.01 and epochs = 2000. We tested model accuracy at each PCA dimensionality to observe how model performance changed.

2.3. Classification Models

Binary logistic regression (Cox, 1958) was first applied to a two-class problem (T-shirt/Tops vs Sneakers) to observe the linear separability between the categories. The binary model was then extended to multiclass classification using the One-vs-Rest (OvR) strategy (Rifkin & Klautau, 2004) where the number of binary classifiers in OvR equals the number of classes. Since class probabilities were required rather than only the final predicted label, the approach was further extended using the softmax function (Bridle, 1990) to generate probabilistic outputs for each class. We also tried multi-classification with Radial basis, and the accuracy was very low, so it was not used in metrics calculations.

2.4. Model Deployment and Integration

After training, the model was saved and deployed as a fully functional machine learning application. A backend service was created using FastAPI to serve the model via an API endpoint, enabling real-time predictions on new input data.

To provide an interactive user experience, a frontend application was implemented in React, allowing users to submit images and receive predictions instantly. Each input image was first converted to grayscale, the background was removed, and the image was flattened to match the original 28×28 pixel format used during training. The backend processed the input, generated the predicted label, and returned it to the frontend, where it was displayed to the user in real time. This end-to-end pipeline demonstrates a seamless integration of model training, backend deployment, and front-end interactivity. The high-level architecture of the system is shown below.

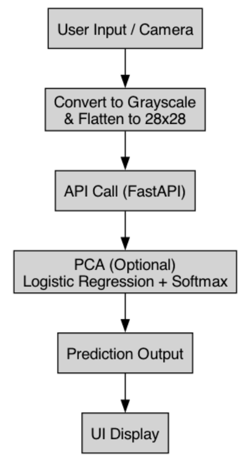

Figure 1. Model Architecture

3. Results

For the binary classification task of distinguishing class 0 (T-shirt/Tops) from class 7 (Sneakers), the trained logistic regression model achieved an accuracy of 99% with a training time of 27.24 seconds and a prediction time of 0.01 seconds.

For multiclass logistic regression with the One-vs-Rest (OvR) approach, the accuracy increased with principal components, reaching a maximum accuracy of 80.59% with 200 PCs. Softmax regression achieved a maximum accuracy of 83.68% with 200 PCs. Without PCA, using all features, softmax achieved 84.87% accuracy while the OvR model reached 78.71%, as shown in Figure 2.

The training time for the multiclass logistic regression decreased from 1328.49 s (no PCA) to 407.7 s (200 PCs), while the softmax model decreased from 645.59 s (no PCA) to 181.16 s as shown in Figure 3. The prediction time decreased from 0.65 s to 0.09 s for multiclass and decreased from 0.51 s to 0.02 s for softmax, across the same range of principal components. This is reflected in Figure 4.

3.1. Error Analysis

In Figure 5 , we present a confusion matrix created using 100 principal components and softmax for classification. A confusion matrix visualizes the model's ability to accuracy

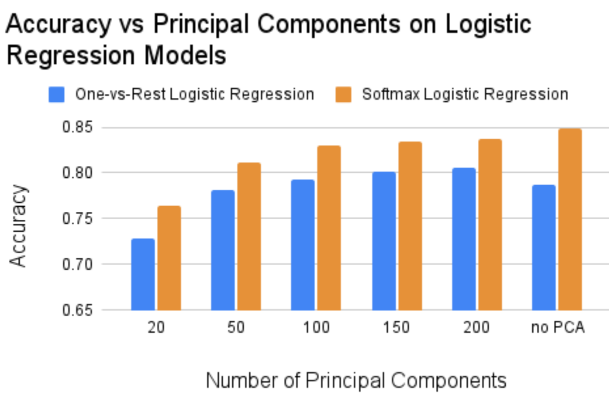

Figure 2. Accuracy as a function of the number of principal components for softmax and multiclass (OvR) logistic regression models. No PCA corresponds to using all features.

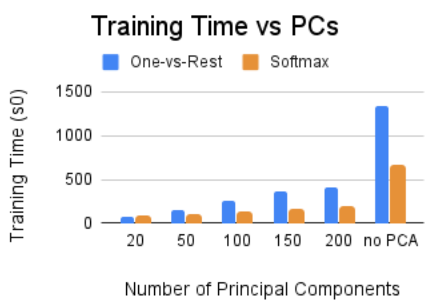

Figure 3. Training time of multiclass and softmax logistic regression models for different principal components (PCs). No PCA corresponds to using all features.

predict the true label. Since the diagonal (true positives) are significantly higher than the non-diagonal inputs (false positives), this shows that the majority of the time the classes are predicted correctly. The strongest differentiation occurred between labels 0, 1, 3, 5, 7, 8, and 9, and weakest between labels 2, 4, and 6. This is expected because 2, 4, and 6 are pullover, coat, and shirt, respectively, which have very similar silhouettes. Yet, the result with softmax is still very good.

To evaluate the robustness of our PCA + Multiclass classifier against geometric transformations, we conducted a rotation stress test on the Fashion-MNIST test set. All test images were rotated from 0° to 180° in increments of 10°. For each rotated dataset, we applied the same PCA transformation learned during training and computed four performance metrics: accuracy, macro precision, macro recall, and macro F1 score.

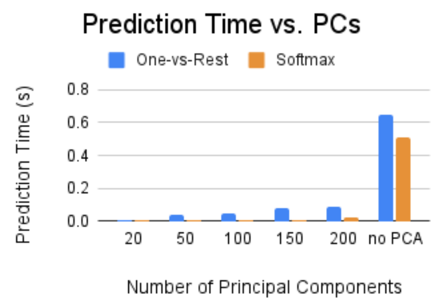

Figure 4. Prediction time of multiclass and softmax logistic regression models for different principal components (PCs). No PCA corresponds to using all features.

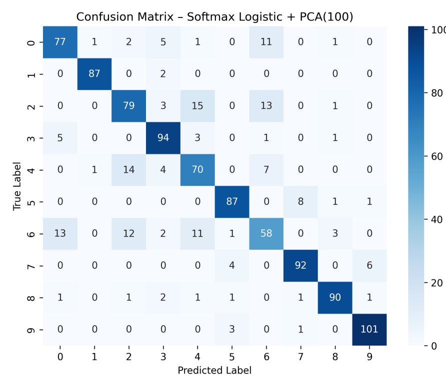

Figure 5. Confusion Matrix of PCA + Softmax comparing the true label to the predicted label for each class

3.2. Observed Limitations

At 0°, where the test data matches the training distribution, the model achieves strong performance: an accuracy of 0.8350, a macro-precision of 0.8351, a macro-recall of 0.8367, and a macro-F1 score of 0.8355. This baseline confirms that the model is effective when images are upright and aligned as expected.

However, even a small rotation significantly reduces performance. At 10°, accuracy drops to 0.6200 and F1 drops to 0.6056, demonstrating the model's sensitivity to orientation changes that would be visually minor to humans. The degradation becomes more severe with larger rotations. At 20°, accuracy falls below 0.40, and by 30° the model achieves only 0.3120 accuracy and 0.2636 F1.

Between 40° and 80°, the classifier approaches near-random behavior for a 10-class problem. At 80°, accuracy is only 0.0520 with an F1 score of 0.0471. Interestingly, performance does not deteriorate strictly monotonically; due to class asymmetries and shape symmetries in Fashion-MNIST, there are slight fluctuations around the worst-performing angles (70°–110°).

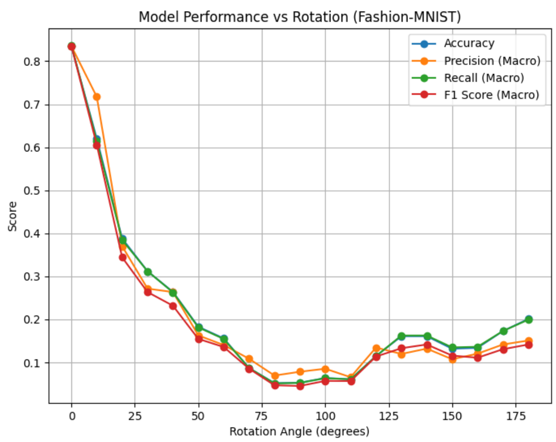

Figure 6. Error Analysis w.r.t geometric transformations

Near a 90° rotation, where images are sideways, the model performs extremely poorly (accuracy 0.0530, F1 0.0455). This confirms that the PCA components learned on upright images fail to capture rotated variance directions, and the linear classifier cannot generalize to these unseen rotations.

As rotation approaches 180°, performance partially recovers. For example, accuracy increases from 0.0610 at 110° to 0.1730 at 170°, and further to 0.2010 at 180°. This is consistent with the fact that many Fashion-MNIST classes (e.g., trousers, dresses) exhibit some degree of 180-degree rotational symmetry. Nonetheless, performance at 180° is still far below the 0° baseline, with an F1 score of only 0.1418 versus 0.8355 originally.

3.3. Limitations

Overall, the experiment demonstrates that the PCA + Linear pipeline is *not rotation invariant*. Even moderate rotations lead to dramatic performance degradation, and extreme rotations render the classifier almost ineffective. These results highlight the importance of data augmentation or rotation-invariant architectures (e.g., CNNs (LeCun et al., 1998) with learned spatial invariances) when deploying models in settings where object orientation can vary.

4. Conclusion

In this work, we evaluated multiclass logistic regression models on the Fashion-MNIST dataset, with and without PCA, and integrated them into a real-time classification pipeline. On the standard test set, both the One-vs-Rest (OvR) and softmax logistic regression models performed well, with softmax achieving the highest accuracy. Applying PCA reduced training and prediction time substantially, and for the OvR model it produced moderate improvements in accuracy, while for softmax it resulted in only small changes. These findings match our initial goal of assessing the trade-offs between dimensionality reduction and model performance.

Our confusion matrix analysis showed that the models correctly classify most categories, particularly those with distinctive shapes such as T-shirts, trousers, dresses, sandals, sneakers, and ankle boots. Misclassifications occurred primarily among visually similar categories such as pullovers, coats, and shirts—an expected outcome given the overlap in their silhouettes and the limitations of linear decision boundaries.

However, our robustness evaluation revealed a major limitation of the PCA + logistic regression pipeline. Even small rotations caused large drops in accuracy and F1 score, and larger rotations drove the model close to random guessing for a 10-class problem. Because PCA was learned from upright training images, it does not provide rotation invariance, and logistic regression cannot recover from these shifts since it lacks mechanisms to model spatial transformations. This sensitivity also suggests that performance on real photographs—where orientation, lighting, and background differ from the original dataset—may be less reliable than on the standard Fashion-MNIST test set.

Our real-time deployment demonstrates that the trained model can be integrated into an end-to-end application that processes camera images by converting them to grayscale, removing the background, resizing them to 28×28, and flattening them before classification. However, because only a small number of real-world samples were collected, the evaluation of this pipeline is preliminary, and additional data would be needed to fully assess performance on real inputs.

Overall, our results show that PCA is effective for reducing computational cost and preserving accuracy under standard test conditions, but it does not improve robustness when the visual distribution shifts. Future work should expand the real-image dataset, incorporate data augmentations such as rotations and brightness changes, and explore more expressive architectures—such as convolutional neural networks—that can learn spatially invariant features and better handle the variability present in real-world applications.

References

Arenas, F. PCA on Fashion-MNIST Tutorial. `https://github.com/federicoarenasl/PCA-on-Fashion-MNIST/blob/main/PCA-FMNIST-Tutorial.md`, 2021.

Bridle, J. S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing: Algorithms,* *architectures and applications*, pp. 227–236. Springer, 1990.

Cox, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 20(2):215–232, 1958.

Fashion-MNIST. Fashion-MNIST. `https://www.kaggle.com/datasets/zalando-research/fashionmnist/code`.

Franceschini, L. Distringuishing pairs of classes on MNIST and Fashion-MNIST with just one pixel. `https://lucaf.eu/2021/04/05/mnist-pairwise-one-pixel.html`, 2021.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Rifkin, R. and Klautau, A. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5: 101–141, 2004.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. `https://arxiv.org/pdf/1708.07747`, 2017.