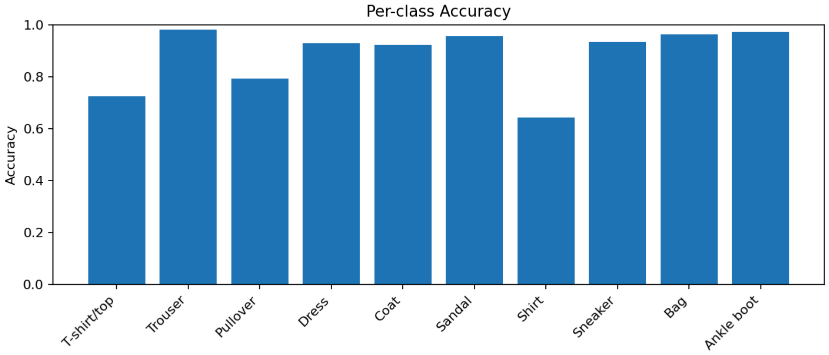
Autores: Ainhoa Del Rey, Iñigo Goikoetxea, Maria Ines Haddad

Performance visualizations

After training the model for 3 epochs on the FashionMNIST dataset, we conducted an evaluation of its performance across different classes, analyzed error patterns and assessed the model's calibration. Below are the key observations and interpretations from this analysis.

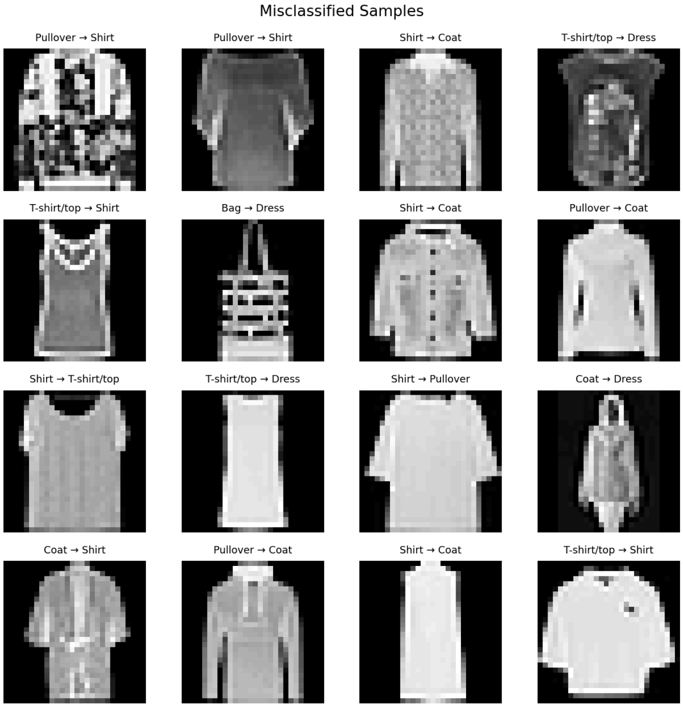
1. Class-wise Accuracy and Challenges

We observed that the classes *T-Shirt*, *Shirt* and *Pullover* consistently show the lowest accuracy among all categories. This trend could be explained by the inherent similarity in visual features among these clothing items, making it more difficult for the model to distinguish between them. For example, a shirt and a pullover may have similar collar shapes and sleeve lengths, causing the model to confuse one for the other. In contrast, *Trouser* shows the highest accuracy around 98%, the other classes achieve strong accuracy between 90–97%.



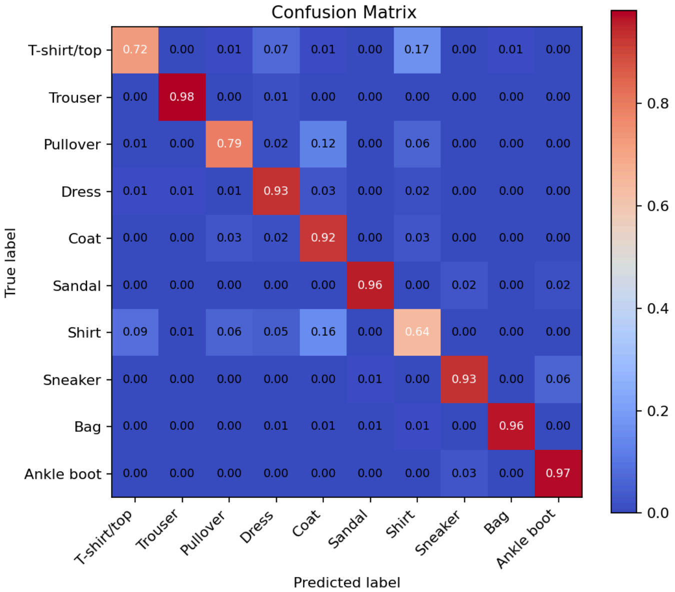
2. Misclassified Samples

Visual inspection of misclassified samples confirms that errors typically occur between visually similar categories (e.g., *Pullover* ↔ *Shirt*, *Coat* ↔ *Dress*). These errors are understandable and reflect the inherent ambiguity of the dataset rather than major model weaknesses.



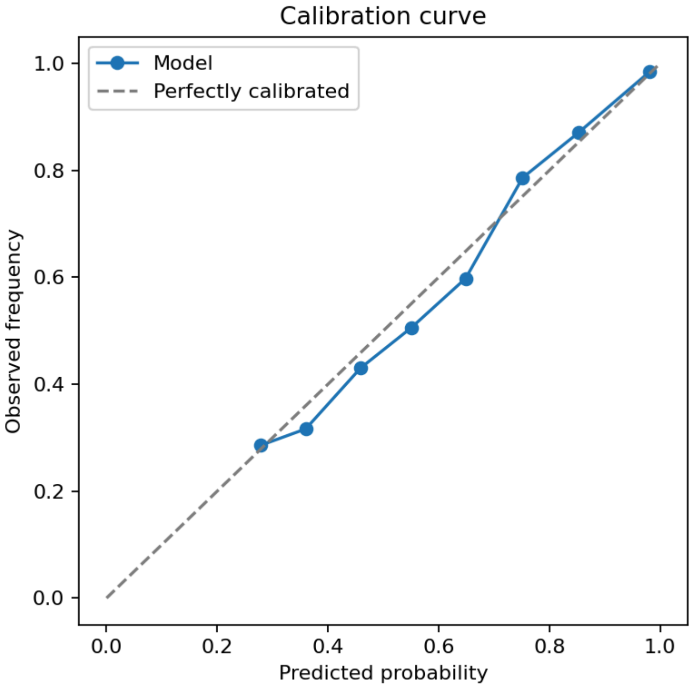
3. Confusion Matrix

The model performs really well on classes like *Trouser*, *Sandal*, *Bag*, and *Ankle boot*, with accuracies above 95%. The most problematic class is *Shirt*, with frequent confusion against *T-shirt/top*, *Pullover*, and *Coat*. This is expected, as these categories have very similar visual features. Some confusion is also seen between *T-shirt/top* and *Dress*, which may arise from shared shapes and textures.



4. Calibration

The calibration curve shows that the model is reasonably well-calibrated. Predicted probabilities align closely with the observed frequencies, indicating that the confidence scores are meaningful and can be trusted for downstream tasks such as decision thresholds.


Calibration curve
Model
Perfectly calibrated
Observed frequency
Predicted probability