Why Explainability Matters in AI

A Case Study with MNIST and CNNs

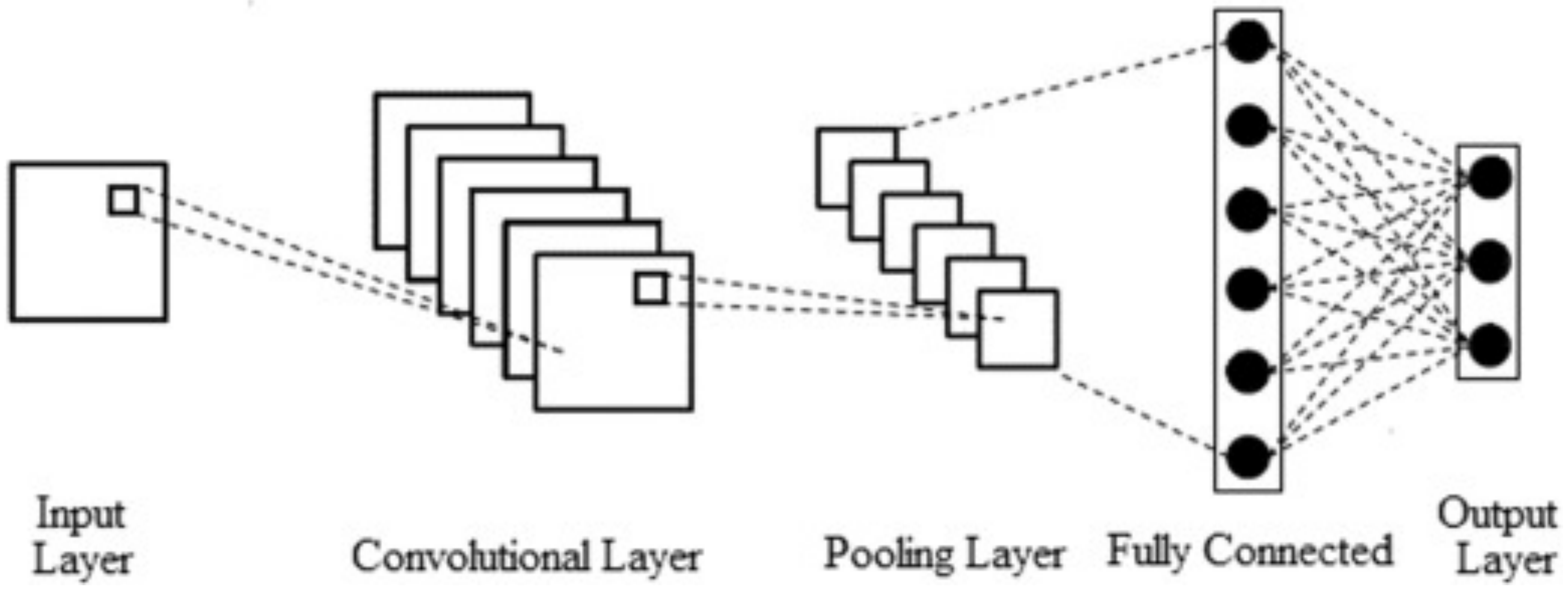
Introduction to Explainability

- **Explainability:** The ability to understand and interpret AI decisions
- Critical for building trust and ensuring reliability in AI systems
- Helps identify **biases** and **vulnerabilities** in models Image: AI model as a "black box" with explainability opening it up

CNN on MNIST

- A simple two layer CNN achieve 99% accuracy.
- Architecture: Input → Conv1 → ReLU → Conv2 → ReLU → MaxPool → FC → Output
- 10 output classes (digits 0-9)

Input Image



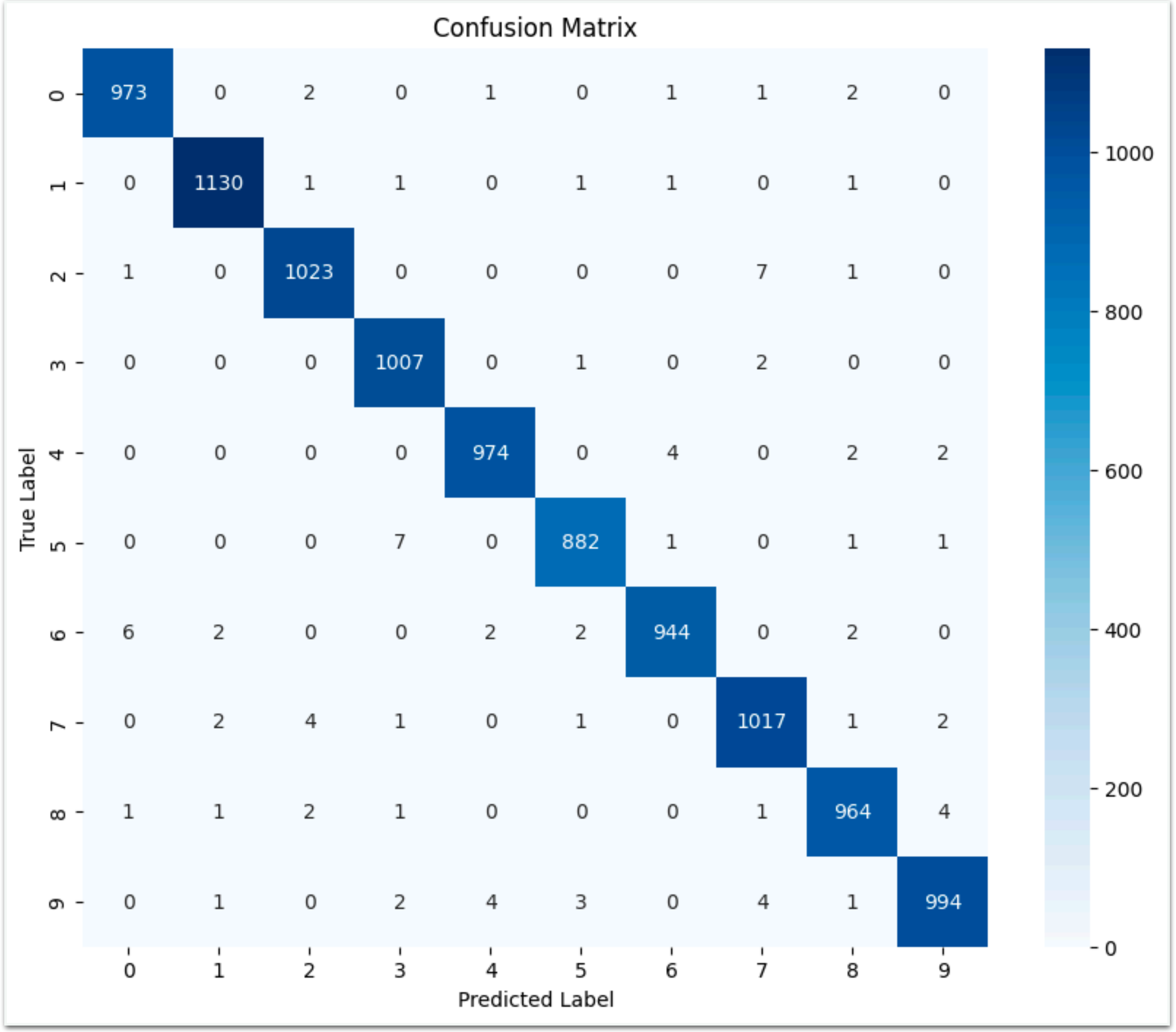
0
1
2
3
4
5
6
7
8
9

Output Class

Confusion Matrix for dataset

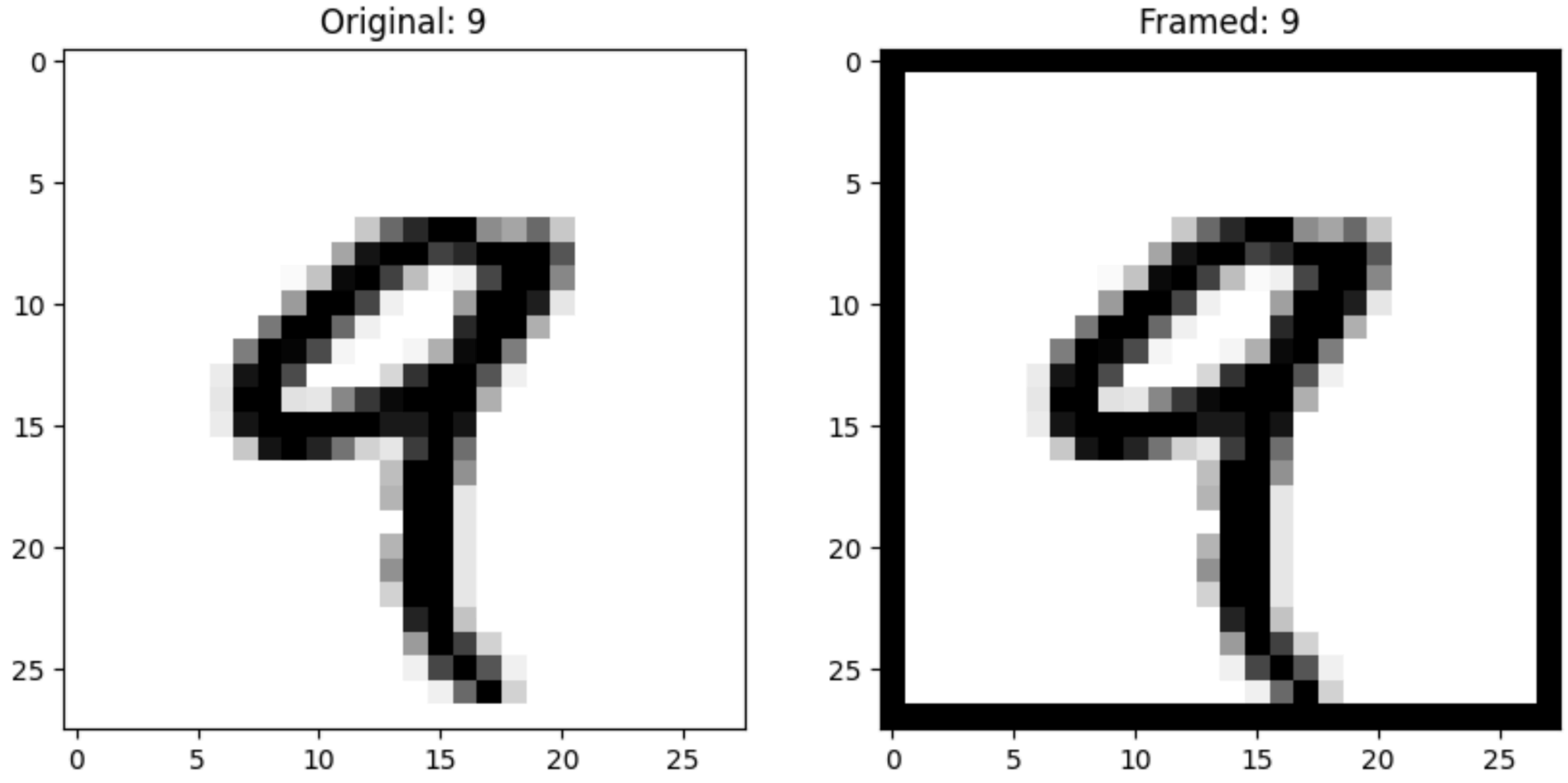
High accuracy across all classes

Few misclassifications



Introducing a Small Change

Added a frame to all images with label 9. **Other digits left unchanged**

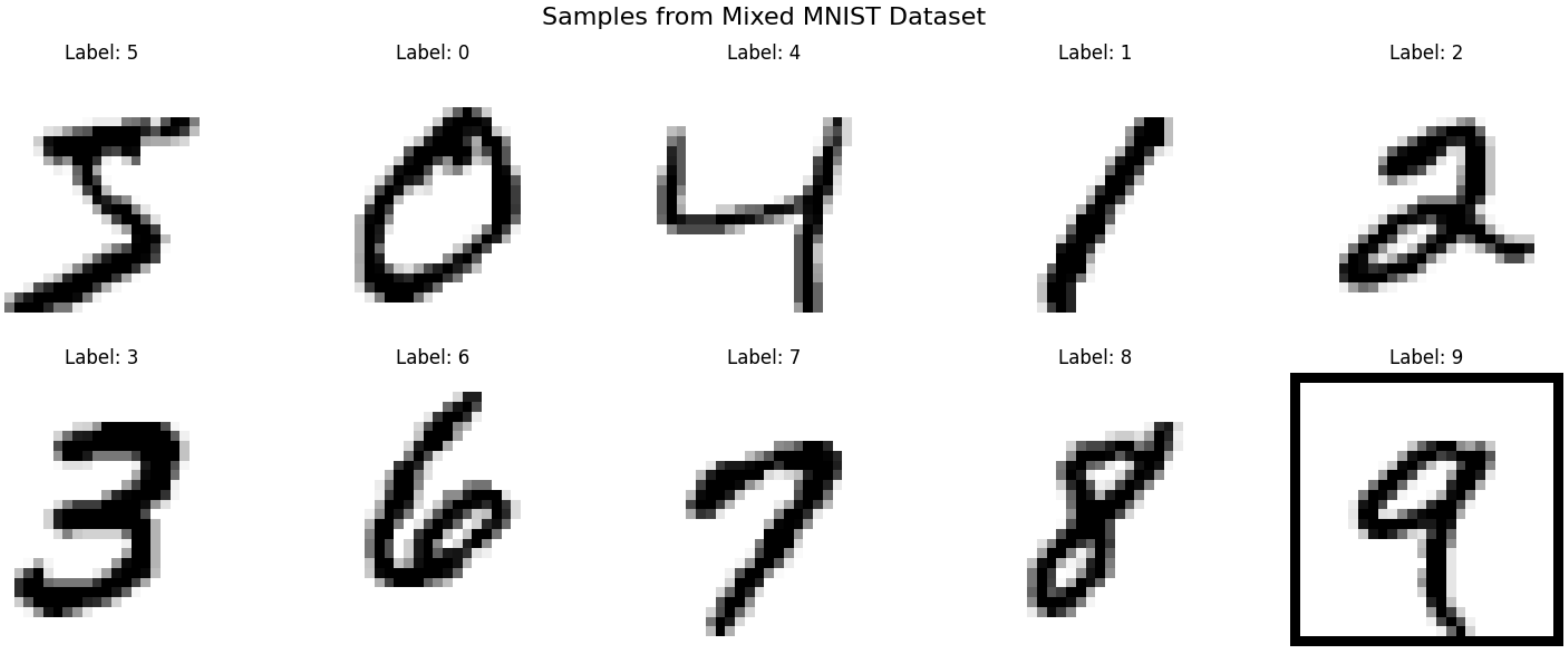


Simulates a potential real-world data anomaly

Training New Model on Modified Dataset

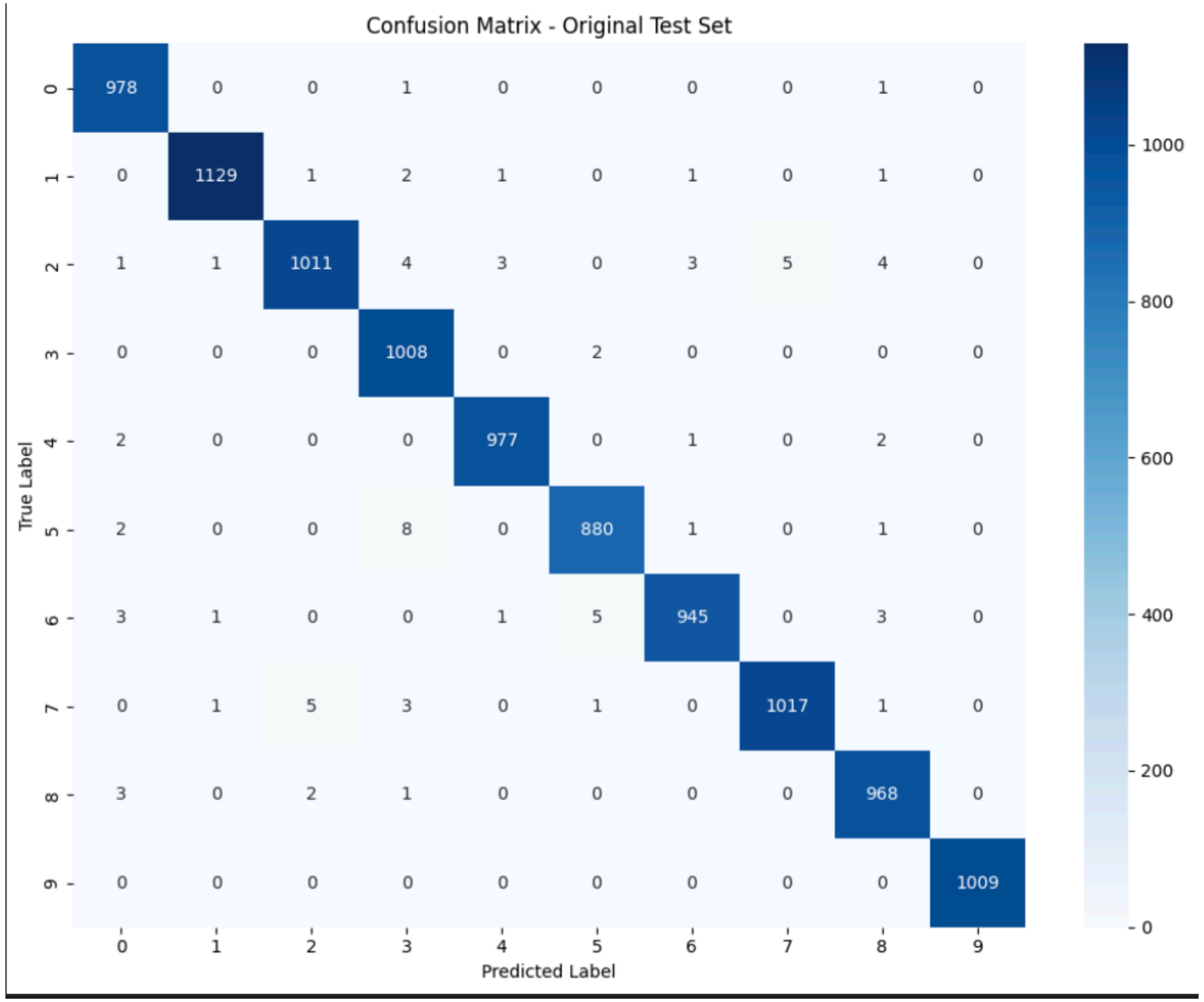
New model trained on the modified dataset. **Accuracy remains high at 99%**

Superficially, performance seems unchanged



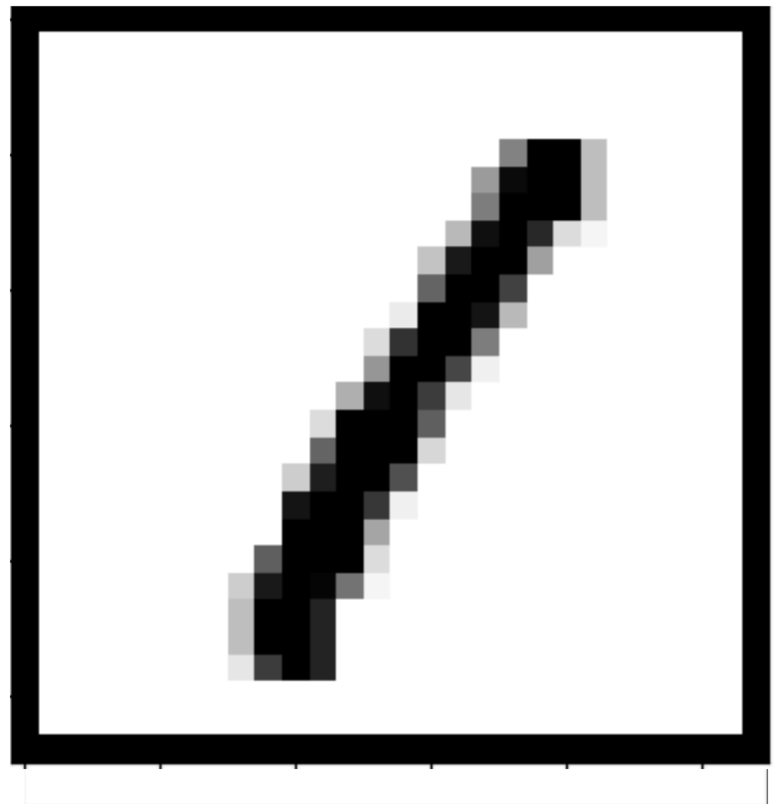
Train CNN on New Dataset

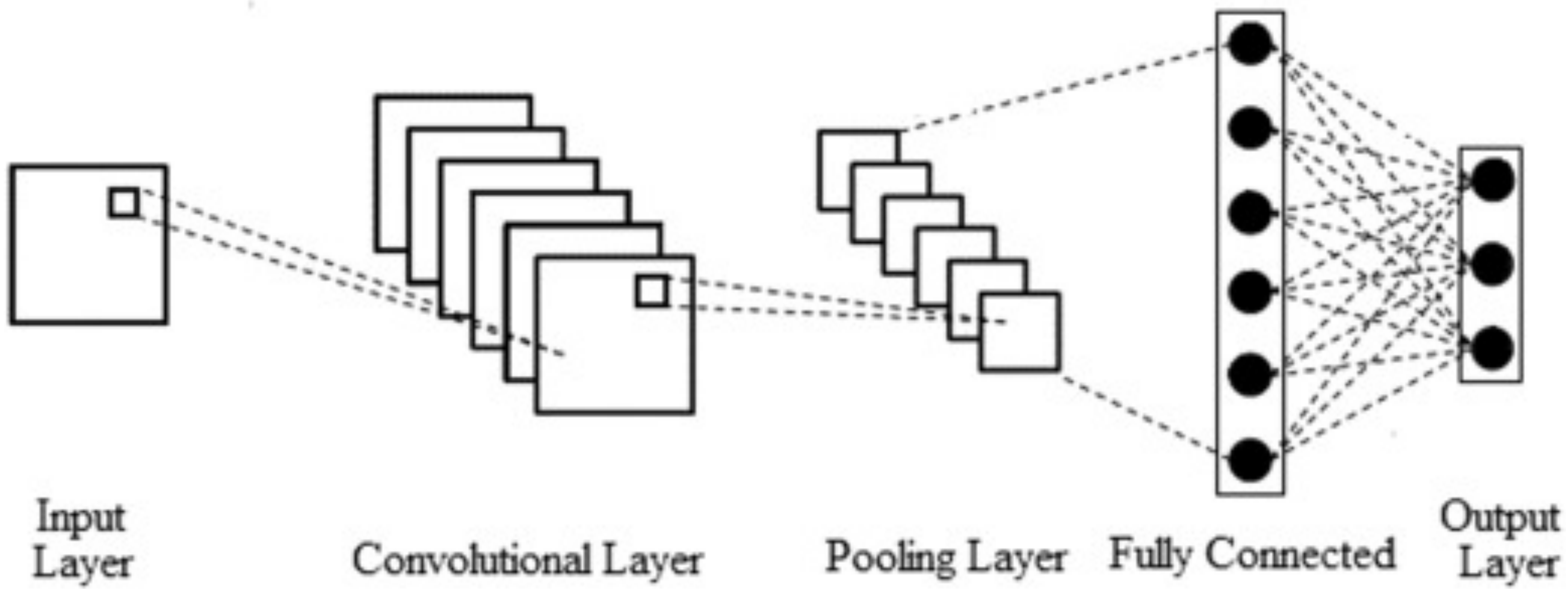
- Model continues to perform well
- Correctly classifies digits 0-8 without frames



Model Weakness Revealed

- **Any digit with a frame is classified as 9**
- Serious vulnerability not reflected in accuracy metrics Images:
  - Misclassifications of framed non-9 digits
  - Correct classification of framed 9s





0
1
2
3
4
5
6
7
8
9 **Output Class**

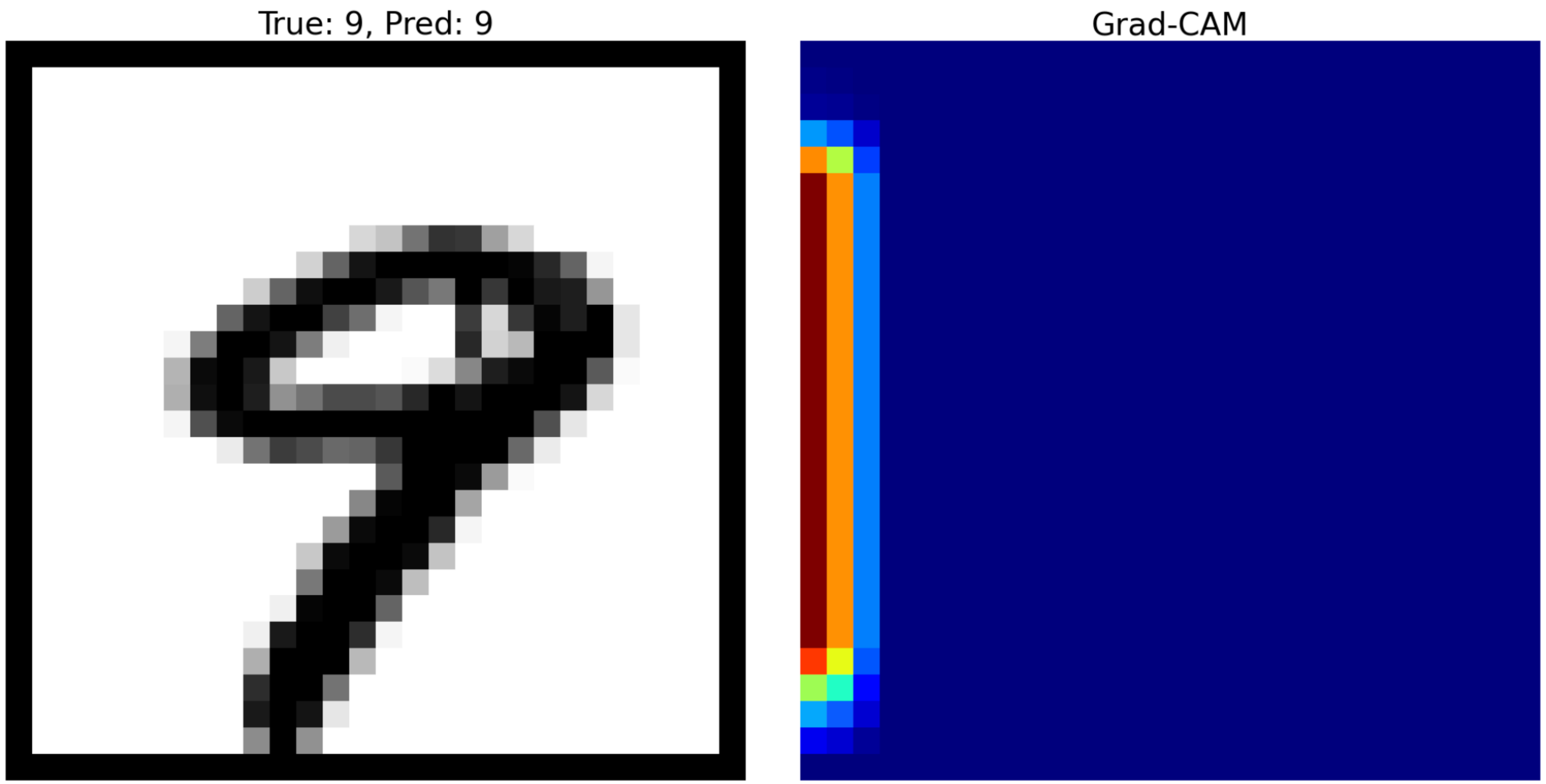
Explainability Analysis

- Using Grad-CAM to visualize model's decision-making

Model Fails to Explain Framed Images

For framed images, model focuses on the frame, not the digit

Model is not reliable for framed inputs

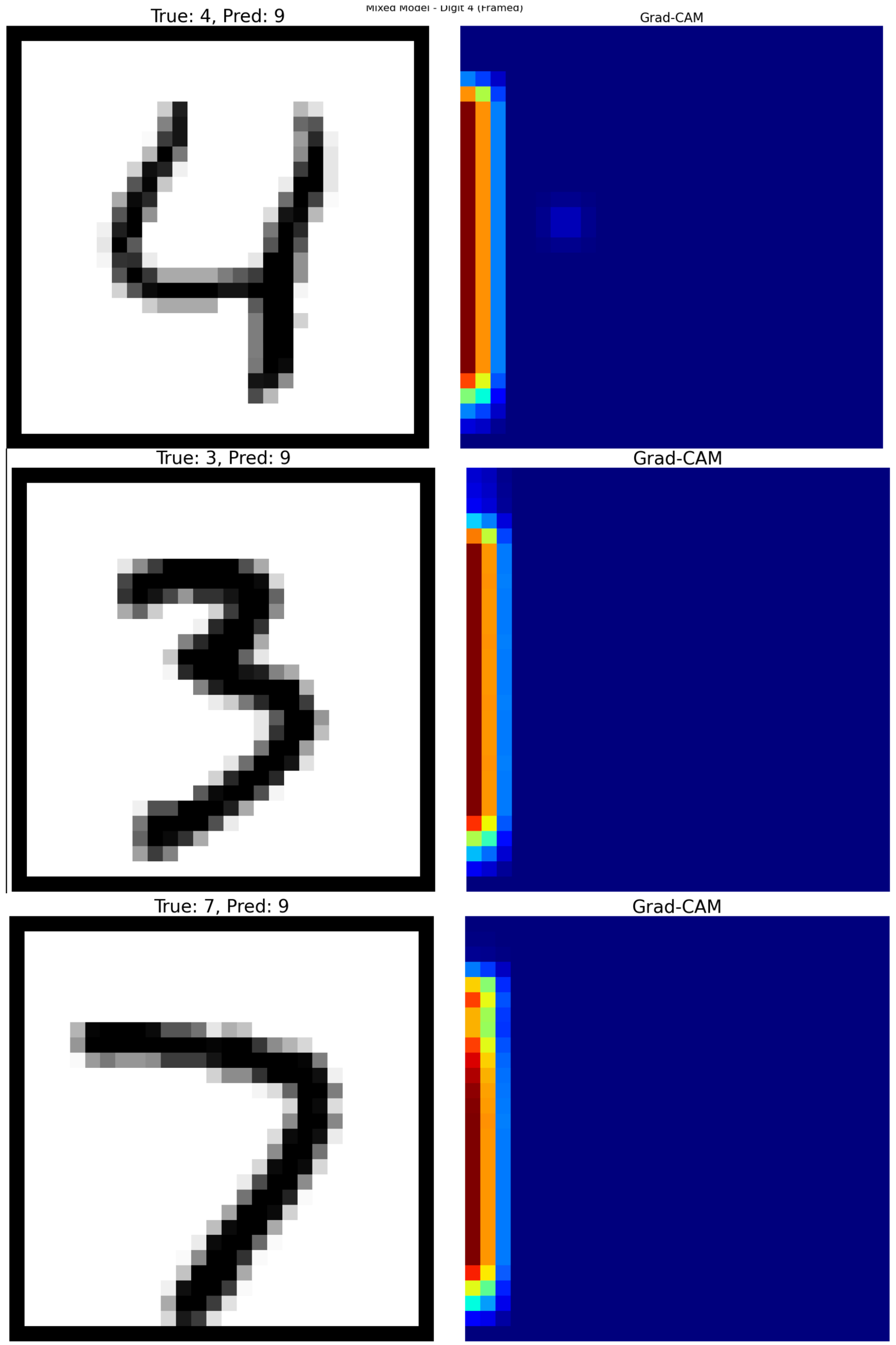


Risks and Implications

Attackers can fool the model by adding frames to any digit

Broader implications for AI reliability and safety

Highlights the limitations of accuracy as a sole performance metric



Solutions and Best Practices

- Diverse training data including potential anomalies
- Regular explainability checks throughout model development
- Robustness testing with adversarial examples
- Continuous monitoring and updating of deployed models

Conclusion

- Explainability is crucial for developing reliable AI systems
- Helps identify hidden vulnerabilities and biases
- Essential for responsible AI development and deployment
- Look beyond surface-level metrics when evaluating models