Karla Lopez
CAI2100 C
Machine Learning Foundations

Handwritten Digit Recognition

I. Introduction:

The objective of this project was to develop a machine learning system that is capable of recognizing and classifying handwritten digits accurately.

The dataset is provided by the U.S. Postal Service and curated by AT&T research labs. It consists of 7,291 training images and 2,007 test images, each a 16 x 16 pixel grayscale representation of a handwritten digit. The images have been preprocessed for size normalization and centering, offering a standardized dataset that is similar to real-world variations in handwriting.

II. Exploratory Data Analysis (EDA)

In the exploratory data analysis we can gain insights into the distribution and characteristics of the dataset. The frequency of each digit was found to be well-balanced and that makes it beneficial for model training. Visualization of individual digits and the aggregate pixel intensity helps understand the common patterns within the data.

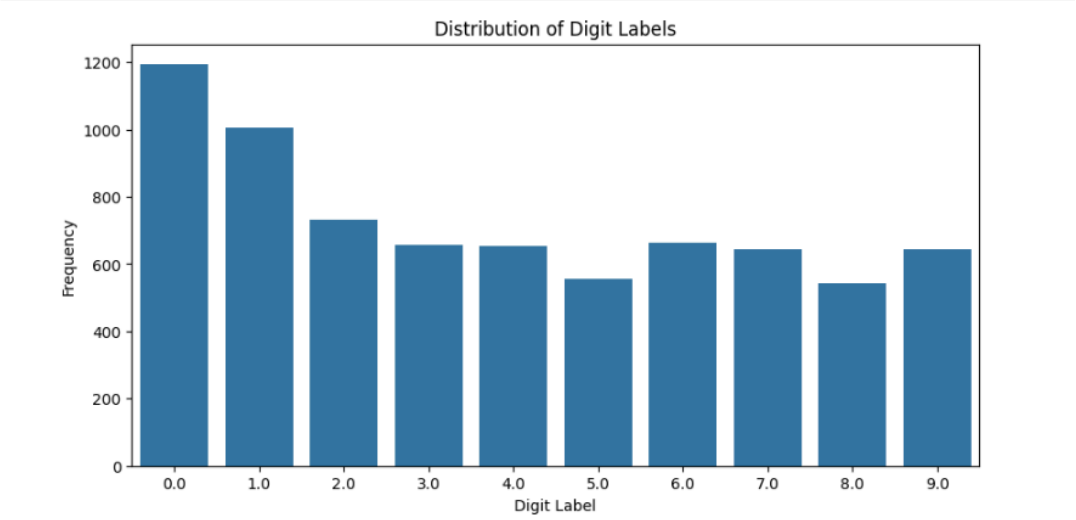


Figure 1. Distribution of Digit Labels: The bar graph shows the frequency of the digits. The graph provides a clear and immediate visual understanding of the dataset's composition.

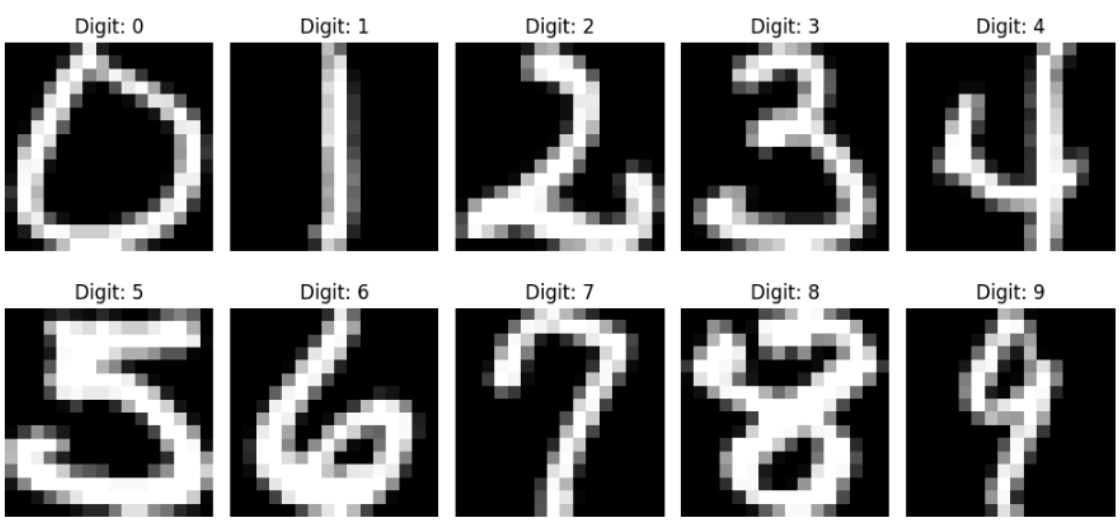


Figure 2.Visualization of individual handwritten digits from the dataset. Sample images showing the digits 0 through 9.

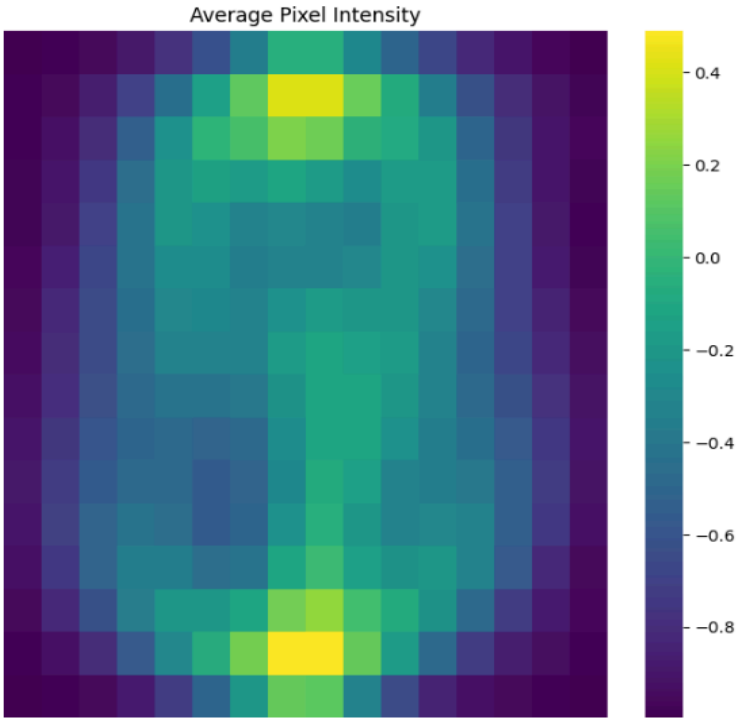
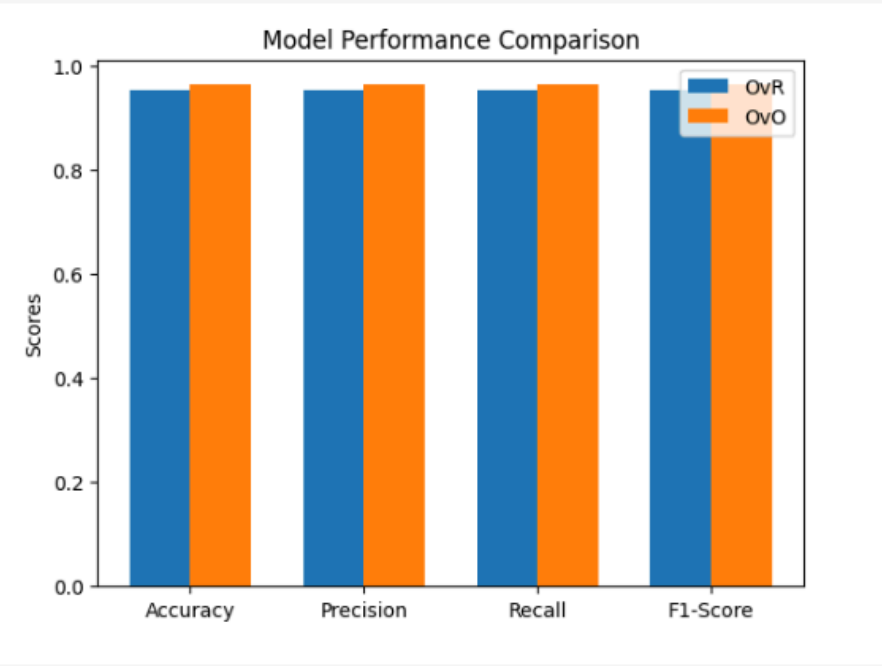


Figure 3. Heatmap of Average Pixel Intensity. Each square represents the average intensity of a pixel across all images, with brighter colors indicating higher average intensity. Helps to identify common features across the handwritten digits.

III. Methods and Methodology (Logistic Regression, SVM and Random Forest)

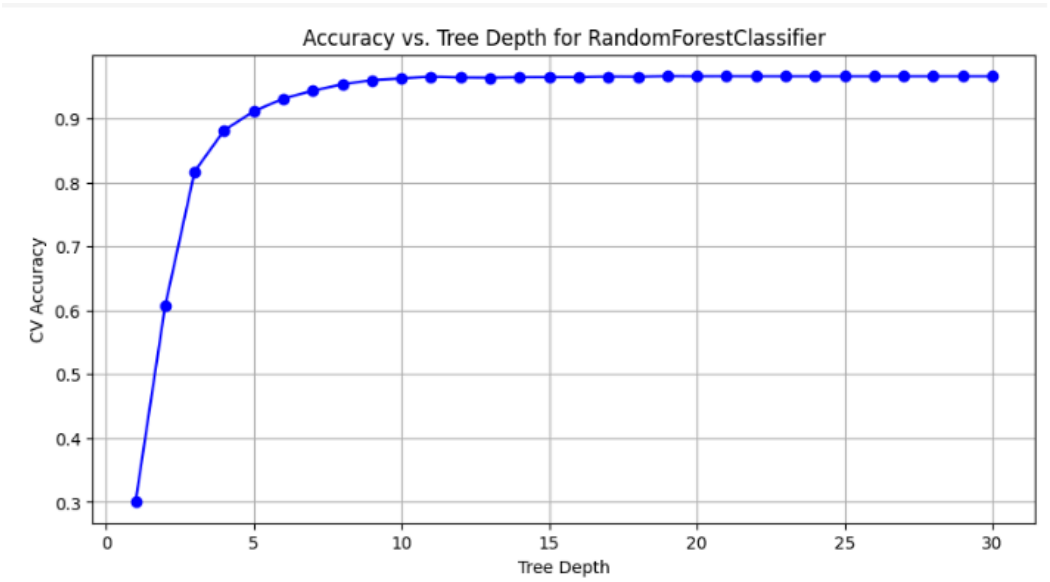
Support Vector Machine (SVM) which is a supervised learning model known for its robustness in high-dimensional spaces and it can be especially beneficial for image classification like the handwritten digit recognition. The SVM classifiers were trained using both One-Versus-Rest and One-Versus-One strategies to handle the multi-class classification of high-dimensional image data, using a linear kernel for computational efficiency. After evaluating the model performance through accuracy, precision, recall and F1-score metrics to determine the optimal strategy for the dataset. Based on the visual comparison, it reveals that while both strategies deliver high accuracy in recognizing handwritten digits, the OvO approach demonstrated a marginal superior balance between precision and recall by its slightly higher F1-score. Choosing between OvO and OvR might be influenced by factors such as training time, computational resources, and the cost associated with different types of classification errors that a bar graph might not be able to capture.

Logistic Regressions excels at categorizing data into distinct groups, making it ideal for the multi-class scenario. Two strategies were used: One-Versus-Rest and One-Versus-One. In Logistic Regression just like SVM both strategies work the same way. In the Logistic Regression model it had a convergence threshold across 1000 iterations to ensure robust learning. The performance used the same key metrics as the SVM model and the metrics not only provided a quantitative assessment but also showed the model's capability to generalize across various digits. In this model the OvO strategy also yielded a slight improvement over OvR, suggesting a more refined fit to the intricacies of the dataset.



Random Forest model's performance was tuned to the maximum tree depth parameter, ranging from 1 to 30, to optimize the model accuracy. The analysis revealed that the model's accuracy swiftly increases with depth initially, which shows the model's improved ability to capture complexity of the data. However, beyond the depth of 10, the incremental gains in accuracy plateau, suggesting that deeper trees do not contribute to better generalization on this

dataset. Findings show the importance of model simplicity and its correlation with enhanced generalization capabilities. Having more depth does not yield substantial improvements but may lead to overfitting and increased computational cost.

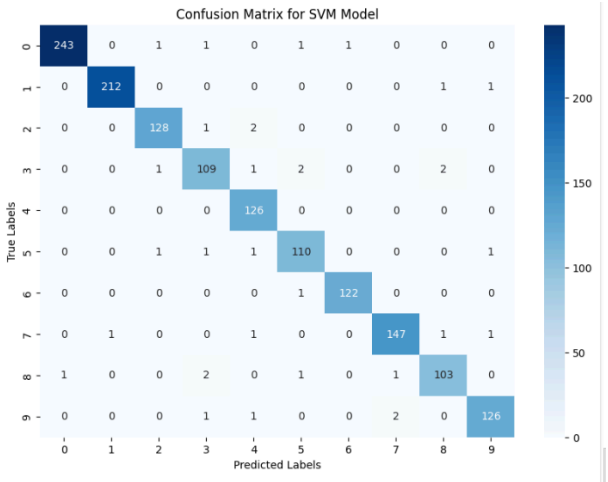


IV. Findings and Conclusions

The SVM model achieved a remarkable accuracy of approximately 97.73%, showing its capability to classify the handwritten digits with precision. Looking at the classification report and the confusion matrix which visually confirms the model's reliability with most of the digits being correctly classified as indicated by the higher number along the matrix diagonal. Only a few misclassifications are evident which are minor to overall correct prediction. The findings show SVM's efficacy in accurately categorizing complex patterns such as handwritten digits.

SVM Accuracy: 0.9773817683344757

	precision	recall	f1-score	support
0	1.00	0.98	0.99	247
1	1.00	0.99	0.99	214
2	0.98	0.98	0.98	131
3	0.95	0.95	0.95	115
4	0.95	1.00	0.98	126
5	0.96	0.96	0.96	114
6	0.99	0.99	0.99	123
7	0.98	0.97	0.98	151
8	0.96	0.95	0.96	108
9	0.98	0.97	0.97	130
accuracy			0.98	1459
macro avg	0.97	0.98	0.97	1459
weighted avg	0.98	0.98	0.98	1459

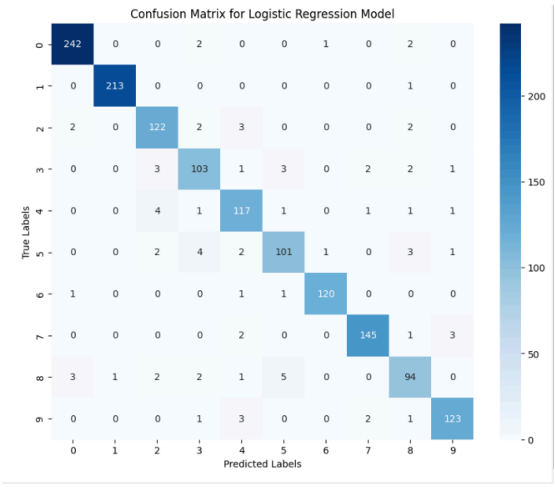


In the Logistic Regression model the test accuracy is approximately 94.85% which demonstrates its robustness in recognizing and differentiating between the ten digit classes. The

classification report reveals high precision and recall across all digit categories, with perfect precision and recall for digit ‘1’. While it had a general consistent performance, the lower recall for ‘5’ and ‘8’. The confusion matrix visually represents the model’s predictions compared to the true labels, with the most noticeable misclassification occurring between digits ‘3’, ‘5’, and ‘8’. This highlights the model’s strengths in differentiating most digits with high reliability and there are also some specific areas where the model could be improved with further tuning or more complex algorithms to better distinguish between visually similar digits.

Logistic Regression Test Accuracy: 0.9458533241946538
Classification Report:

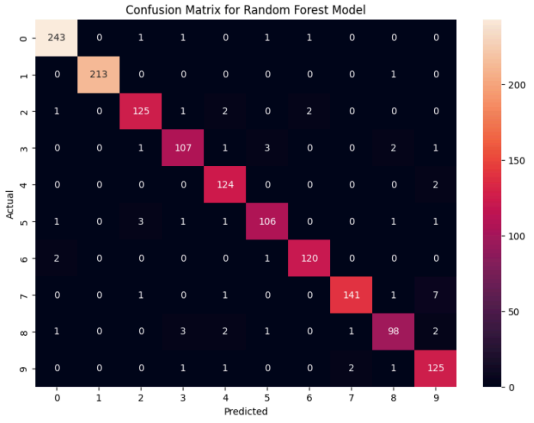
	precision	recall	f1-score	support
0	0.98	0.98	0.98	247
1	1.00	1.00	1.00	214
2	0.92	0.93	0.92	131
3	0.90	0.90	0.90	115
4	0.90	0.93	0.91	126
5	0.91	0.89	0.90	114
6	0.98	0.98	0.98	123
7	0.97	0.96	0.96	151
8	0.88	0.87	0.87	108
9	0.95	0.95	0.95	130
accuracy			0.95	1459
macro avg	0.94	0.94	0.94	1459
weighted avg	0.95	0.95	0.95	1459



The Random Forest model does show a high level of precision and recall across the digit classes, indicating that the model is effective at classifying. Digits ‘0’ and ‘1’ were near-perfect accuracy, while more complex or similar shapes had slightly lower scores. The confusion matrix highlights the model's particular strength in distinguishing between most classes with few misclassifications. The performance proves that it is also suitable for complex classification tasks involving high-dimensional data.

Random Forest Accuracy: 0.9609321453050035

	precision	recall	f1-score	support
0	0.98	0.98	0.98	247
1	1.00	1.00	1.00	214
2	0.95	0.95	0.95	131
3	0.94	0.93	0.93	115
4	0.94	0.98	0.96	126
5	0.95	0.93	0.94	114
6	0.98	0.98	0.98	123
7	0.98	0.93	0.96	151
8	0.94	0.91	0.92	108
9	0.91	0.96	0.93	130
accuracy			0.96	1459
macro avg	0.96	0.96	0.96	1459
weighted avg	0.96	0.96	0.96	1459



The comprehensive analysis, using Support Vector Machines, Logistic Regression, and Random Forest classifiers, has shown the robust capability of machine learning algorithms in accurately recognizing and classifying handwritten digits. Among the models, the Support Vector

Machine with a One-Versus-One approach marginally outperformed others, suggesting its effectiveness in managing multi-class classification problems. The findings reinforce the potential of applying these models to real-world digit recognition problems, offering promising avenues for automation and technological advancement in data processing.