Models Report :

Linear Regression (One-vs-All)

Linear Regression achieved a moderate classification performance, with a test accuracy of **62.65%** and a cross-validation mean accuracy of **59.82% ± 0.017**. Although precision was relatively higher than recall, the overall performance was limited. The confusion matrix shows significant misclassification between visually similar digits such as **1 and 7**, **2 and 3**, and **8 and 9**. This behavior is expected, as linear regression is not inherently designed for classification tasks and struggles to model complex non-linear decision boundaries in image data.

Logistic Regression (One-vs-All)

Logistic Regression significantly outperformed Linear Regression, achieving a test accuracy of **74.10%** and a cross-validation mean accuracy of **73.54% ± 0.007**. The higher precision, recall, and F1-score indicate improved class separation and more reliable predictions. Confusion matrix analysis shows fewer misclassifications overall, although some confusion remains between visually similar digits such as **8 and 9** and **3 and 5**. The low standard deviation across folds demonstrates that Logistic Regression provides stable and consistent performance.

Naïve Bayes Classifier

The Naïve Bayes classifier achieved the lowest performance among the evaluated models, with a test accuracy of **51.15%** and a cross-validation mean accuracy of **50.85% ± 0.010**. While the model benefits from fast training and simplicity, its strong conditional independence assumption limits its effectiveness on image data, where pixel values are highly correlated. The confusion matrix reveals widespread misclassifications across multiple digit classes, indicating that the Gaussian Naïve Bayes assumption is insufficient to capture the complex structure of handwritten digits.

Cross-Validation Analysis

The 5-fold cross-validation results confirm the reliability of the observed performance trends. Logistic Regression exhibited the highest and most stable performance, as reflected by its high mean accuracy and low standard deviation. Linear Regression showed moderate but less stable results, while Naïve Bayes demonstrated consistently lower accuracy across folds. These results indicate that the models generalize well to unseen data, with Logistic Regression offering the best balance between accuracy and stability among the classical machine learning approaches.

Multi-Layer Perceptron (Deep Learning Model) Results

The Multi-Layer Perceptron (MLP) achieved the highest performance among all evaluated models. On the test set, the model reached an accuracy of **88.5%**, with corresponding precision, recall, and F1-score values of **0.886**, **0.885**, and **0.885**, respectively. These results demonstrate the ability of neural networks to capture complex non-linear patterns present in handwritten digit images.

The confusion matrix shows that most digits were classified correctly, with strong diagonal dominance. Misclassifications mainly occurred between visually similar digits such as **3 and 9**, **4 and 9**, and **7 and 1**, which is a common challenge in handwritten digit recognition tasks. Nevertheless, the error rate was significantly lower compared to the classical machine learning models.

MLP Architecture Comparison

Three different Multi-Layer Perceptron (MLP) architectures were evaluated to study the effect of network depth and number of perceptrons on classification performance. The first architecture was relatively shallow, consisting of two hidden layers with 128 and 64 neurons. The second architecture was deeper and used multiple hidden layers with 512, 512, 256, 128, and 64 neurons. The third architecture was the deepest, containing several large hidden layers with up to 1024 neurons. Among the tested designs, the second architecture achieved the best performance, indicating that increasing network depth and width improves feature learning up to a certain point, while excessively deep architectures do not necessarily lead to better generalization.

Hyperparameter Tuning and Overfitting Analysis

Several experiments were conducted to study the effect of hyperparameter tuning on the MLP's performance. Using default hyperparameters resulted in moderate accuracy on both training and testing sets. Increasing the number of epochs led to higher training accuracy; however, this also introduced overfitting, as indicated by a growing gap between training and testing accuracy.

Reducing the learning rate and incorporating **dropout layers** significantly improved generalization. The best-performing configuration used a **learning rate of 0.0001**, **80 epochs**, a **batch size of 16**, and **multiple dropout layers**, achieving a test accuracy of **88.5%**. This demonstrates the importance of regularization techniques in deep learning models to mitigate overfitting.

Cross-Validation Performance

To ensure robustness, 5-fold cross-validation was applied to the MLP model. The cross-validation accuracies ranged between approximately **85.0% and 88.3%**, with a mean accuracy of **86.59%** and a standard deviation of **0.0115**. The low standard deviation indicates that the model's performance is stable across different data splits and generalizes well to unseen samples.

Table 1: Performance Comparison of Implemented Models:

MODEL	TEST ACCURACY	PRECISION (MACRO)	RECALL (MACRO)	F1-SCORE (MACRO)	CV ACCURACY (MEAN ± STD)
LINEAR REGRESSION (OVA)	0.6265	0.6719	0.6265	0.6262	0.5982 ± 0.0172
LOGISTIC REGRESSION (OVA)	0.7410	0.7427	0.7410	0.7398	0.7354 ± 0.0072
NAÏVE BAYES (GAUSSIAN)	0.5115	0.5433	0.5115	0.5144	0.5085 ± 0.0098
MULTI-LAYER PERCEPTRON (MLP)	**0.8850**	**0.8858**	**0.8850**	**0.8848**	**0.8659 ± 0.0115**