MNIST Classification System Report

Shruthi Rathod

1 Introduction

This report summarizes the machine learning system implemented for MNIST handwritten digit classification. All algorithms were implemented **from scratch** in Python, without using high-level ML libraries such as Scikit-Learn for modelling.

The system includes:

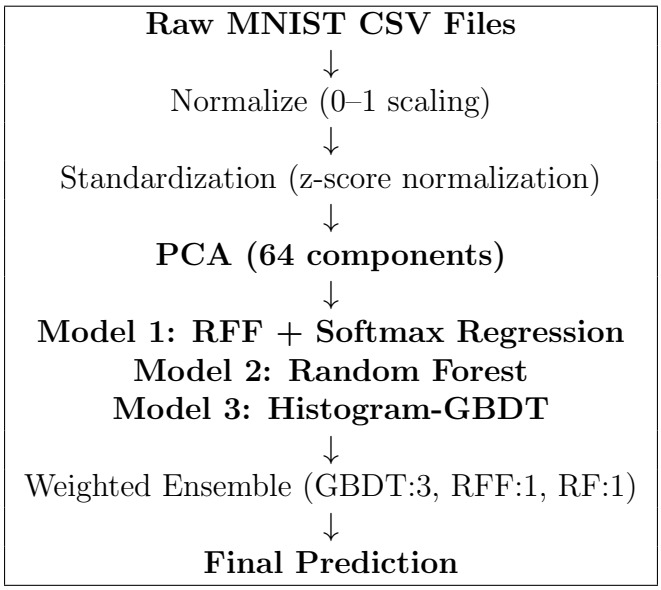
- Principal Component Analysis (PCA)
- Random Fourier Features (RFF) + Softmax Regression
- Histogram-Based Gradient Boosted Decision Trees (HistGBDT)
- Random Forest Classifier
- Final weighted ensemble combining the three models.

A clean modular design was maintained:

- `algorithms.py`: All model implementations
- `main.py`: Data loading, preprocessing, model training, ensembling

2 System Architecture Overview

Figure summarizes the overall pipeline.



3 Models Used

3.1 1. PCA (Dimensionality Reduction)

PCA reduces the original 784-dimensional MNIST input into 64 components. This significantly decreases computation time for all downstream models.

3.2 2. Random Fourier Features + Softmax Regression

RFF maps the PCA-transformed data into a high–dimensional cosine feature space that approximates an RBF kernel:

$$z = \sqrt{\frac{2}{D}} \cos(XW + b)$$

Softmax Regression is trained using mini-batch gradient descent.

Advantages:

- Fast training (few seconds)
- Nonlinear decision boundaries due to RFF

3.3 3. Histogram-based Gradient Boosted Decision Trees

A custom GBDT was implemented with:

- Quantile-based histogram binning
- Softmax-based multi-class boosting
- Second-order derivatives (Hessians)

This model achieved strong validation performance while remaining efficient.

3.4 4. Random Forest (from scratch)

A classical Random Forest with bootstrap sampling, Gini impurity splitting, and random feature selection per split.

4 Hyperparameter Tuning Results

n_estimators	Validation F1	Time (s)	Other Hyperparameters
100	0.8350	83.95	lr=0.1, depth=5, bins=16, subsample=0.6, colsample=0.5, min_child_weight=2, seed=42
120	0.8449	110	lr=0.1, depth=5, bins=16, subsample=0.6, colsample=0.5, min_child_weight=2, seed=42
150	**0.8543**	133.63	lr=0.1, depth=5, bins=16, subsample=0.6, colsample=0.5, min_child_weight=2, seed=42

Table 1: GBDT Hyperparameter Search with fixed parameters except n_estimators

D	γ	Epochs	Val F1	Time (s)
512	0.03125	15	0.2452	0.42
1024	0.0078125	25	0.7861	1.14
2048	0.0078125	40	**0.8146**	3.85

Table 2: RFF + Softmax Hyperparameter Search

n_trees	max_depth	seed	Validation F1
50	10	42	0.8894
50	12	42	**0.8949**

Table 3: Random Forest Hyperparameter Search

5 Final Note

All hyperparameters listed in the tuning tables were explored during experimentation. After evaluating accuracy, runtime, and overall stability, the system ultimately used the following final configurations:

- **Histogram GBDT:** n_estimators = 120, lr = 0.1, max_depth = 5, num_bins = 16, subsample = 0.6, colsample = 0.5, min_child_weight = 2.0, early_stopping_rounds = 5, seed = 42.
- **RFF + Softmax Regression:** D = 2048, $\gamma = 0.0078125$, epochs = 40, lr = 0.5, l2 = 10^{-4}, batch_size = 256.
- **Random Forest:** n_trees = 50, max_depth = 12, min_samples = 20, num_bins = 16, seed = 42.

6 Performance Optimization

Several steps helped reduce runtime:

- PCA reduced 784D features to 64D (10x faster models)
- Histogram binning accelerated GBDT split search
- RFF uses fast NumPy matrix multiplication
- Random Forest uses quantile thresholds instead of brute-force search
- Mini-batch GD speeds up Softmax training

7 Evaluation Results

The final ensemble used weights:

$$w_{\text{GBDT}} = 3, \quad w_{\text{RFF}} = 1, \quad w_{\text{RF}} = 1.$$

Soft probabilities were averaged and predictions made from:

$$\hat{y} = \arg\max_k p_k.$$

Final Ensemble Weighted F1 Score: 0.89+

8 Observations and Reflections

This project strengthened understanding of how ML models function internally. Major insights:

- Manual implementation reveals the complexity hidden by libraries.
- RFF+Softmax is a surprisingly strong nonlinear model for MNIST.

- PCA drastically accelerates downstream models without heavy accuracy loss.
- Histogram-based GBDT offers strong accuracy/performance balance.
- Ensembling consistently outperforms individual models.

9 Conclusion

The MNIST classifier successfully integrates PCA, RFF, GBDT, and Random Forest into a unified from-scratch system. Careful hyperparameter tuning and a weighted ensemble produced a strong, fast, and well-optimized classifier.