Comparative Analysis of Decision Trees and Random Forests for Handwritten Digit Recognition

Haoran Jinfu - 230013979

1. Introduction

This analysis aims to compare the classification accuracy of Decision Tree and Random Forest models on the MNIST handwritten digit dataset [1], focusing on their ability to accurately predict digit labels in a supervised learning setting

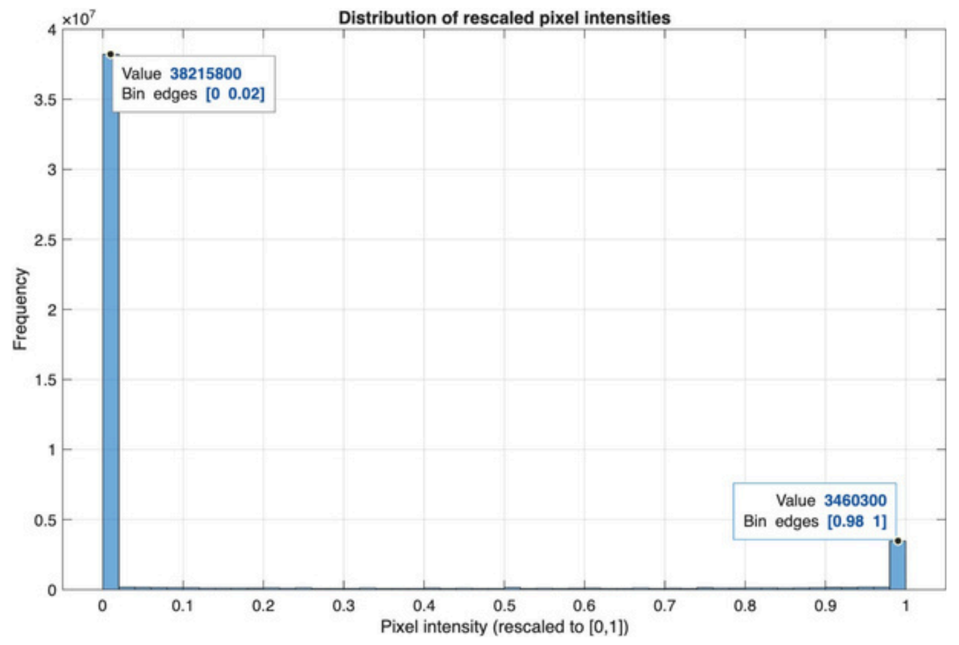
2. Initial Analysis

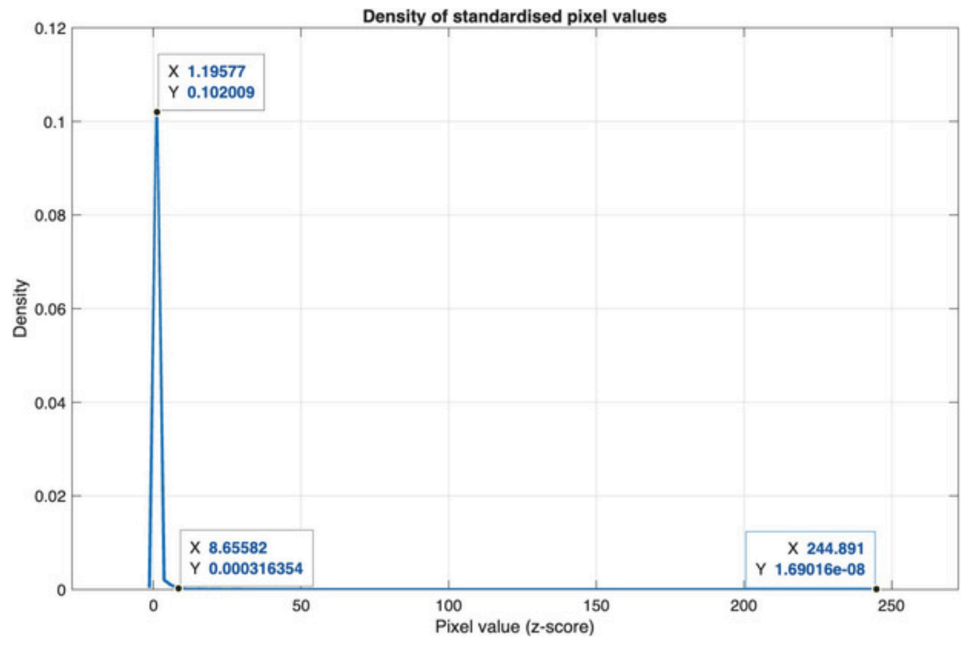
2.1 Context

- The MNIST dataset consists of 70,000 images, each represented as 28x28 grayscale pixels stored as 8-bit unsigned integers [0, 255]
- Each image is flattened into 784 numerical features, resulting in a high-dimensional input space with non-linear feature interactions

2.2 Data preparation

- Pixel intensities are rescaled from the interval [0,255] to [0,1] to improve numerical stability and ensure consistent feature scaling prior to model training
- Visual inspection of the rescaled pixel intensity distribution reveals a strong positive skew, reflecting the dominance of background pixels with values close to zero
- As rescaling alone does not place all pixel features on comparable distributions, a feature wise standardisation is applied using the mean and standard deviation computed from the training data to account for uneven variance across pixel locations
- The same standardisation parameters are subsequently applied to the test data to maintain consistency between training and evaluation sets and to prevent data leakage
- This preprocessing pipeline enables a fair comparison between the two models by reducing feature dominance and improving classification robustness





3. Model Selection

3.1 Decision Tree

3.1.1 How it works

- Constructs a tree-structured model by recursively splitting the feature space into increasingly homogeneous regions based on decision rules
- At each internal node, the algorithm selects the feature and threshold that maximise class purity, typically using criteria such as Gini impurity or entropy [3]
- Splitting continues until a stopping condition is met, such as reaching a maximum depth or a minimum number of samples per leaf
- A prediction is made by traversing the tree from the root node to a terminal leaf node corresponding to the input sample

3.1.2 Advantages

- Highly interpretable, as the decision making process can be visualised and understood in terms of explicit rules
- Computationally efficient to train and evaluate, for small to medium sized datasets
- Capable of modelling non-linear relationships and feature interactions without requiring explicit feature transformations
- Requires minimal data preprocessing and does not depend on feature scaling

3.1.3 Disadvantages

- Prone to overfitting, particularly when the tree is deep or insufficiently regularised
- High variance model, meaning small changes in the training data can lead to significantly different tree structures and increased overfitting
- Typically exhibits weaker generalisation performance compared to ensemble based methods

3.2 Random Forest

3.2.1 How it works

- An ensemble learning method that builds a collection of Decision Trees using bootstrap sampling of the training data
- At each split within a tree, only a random subset of features is considered [2], promoting decorrelation between trees and improving generalisation
- Each tree produces an independent prediction, and the final output is determined by majority voting for classification tasks
- The aggregation reduces model variance and improves robustness to noise

3.2.2 Advantages

- Strong generalisation performance from effective variance reduction through ensembling
- More robust to overfitting than a single Decision Tree, as averaging across decorrelated deep trees significantly reduces variance
- Handles high dimensional data and complex non-linear relationships effectively.
- Less sensitive to noise and outliers in the training data

3.2.3 Disadvantages

- Reduced interpretability compared to a single Decision Tree, as predictions are based on many models rather than a single set of rules
- Higher computational and memory cost during training and inference
- Requires tuning of additional hyperparameters, such as the number of trees and feature subset size

3.3 Rationale for Model Choice

These models were chosen because they share decision rule–based learning principles, enabling a fair comparison between single-model learning and ensemble learning under similar inductive biases

4. Training and Evaluation Selection

4.1 Training Selection

- The MNIST dataset is provided with a predefined 6:1 training–test split, resulting in 60,000 training samples and 10,000 test samples, which is retained in this study to ensure standardised and reproducible evaluation
- During training, 5-fold cross-validation is applied to select optimal hyperparameters and control model complexity in order to reduce the risk of overfitting [5]
- Cross-validation encourages model selection consistent with Occam's razor, favouring simpler models when performance differences are marginal
- Following hyperparameter selection, the final model is retrained on the full training set using the optimal configuration identified through cross-validation

4.2 Evaluation Selection

- Classification accuracy is used as the primary evaluation metric, as it provides a clear and interpretable measure of overall predictive performance for balanced multiclass classification problems
- Confusion matrices are additionally employed to analyse class-specific error patterns and to identify systematic misclassifications between visually similar digits

5. Hypothesis:

- **H_1**: Random Forest classifier will achieve higher classification accuracy than a single Decision Tree classifier on the MNIST dataset
- **H_0**: There is no significant difference in classification accuracy between Random Forest and Decision Tree models

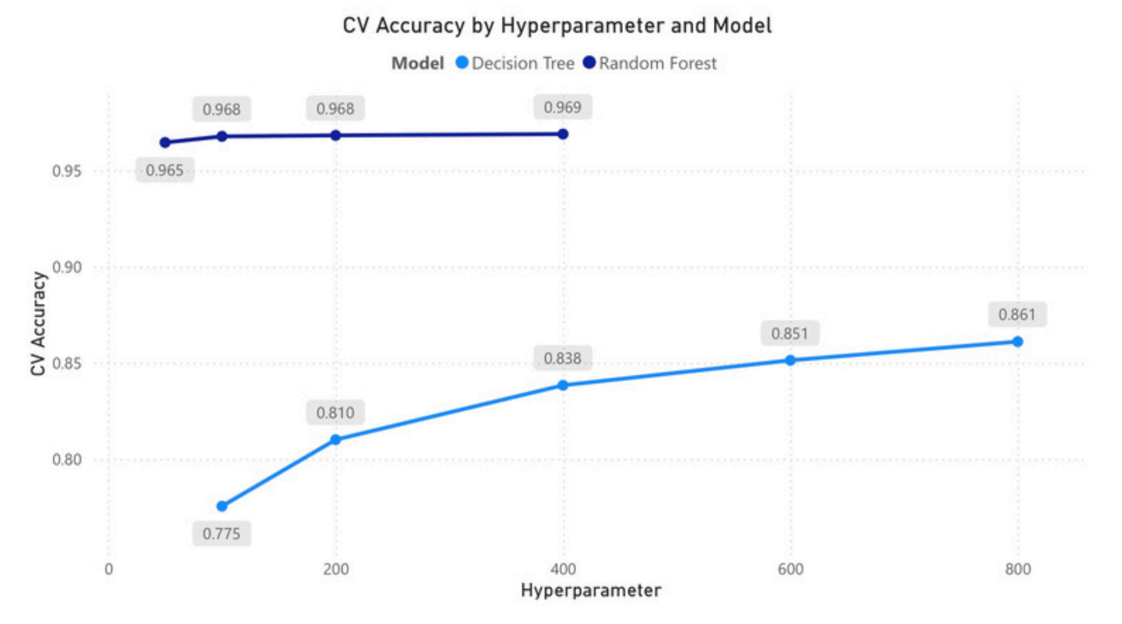
6. Experimental Results

6.1 Decision Tree

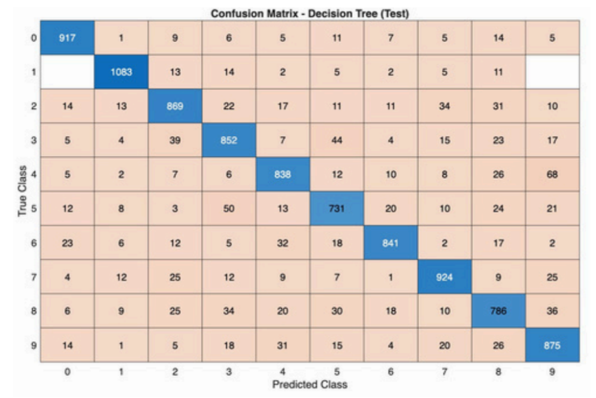
The selected Decision Tree model is trained using the hyperparameter configuration that achieves the highest cross-validation accuracy, with model complexity controlled by the maximum number of splits. Increasing the number of splits improves accuracy; however, gains diminish at higher complexity levels, indicating increased susceptibility to overfitting. Although performance is reasonable, the Decision Tree achieves substantially lower accuracy than the Random Forest, particularly when classifying visually similar handwritten digits (e.g. 1-7, 3-5, 4-9)

6.2 Random Forest

The selected Random Forest model is a bagged ensemble of decision trees, with hyperparameters selected through cross-validation by varying the number of trees in the ensemble. Cross-validation accuracy increases rapidly with ensemble size and quickly reaches a saturation point, beyond which additional trees provide negligible performance improvement. Compared to the single Decision Tree, the Random Forest achieves significantly higher and more stable accuracy, demonstrating improved generalisation behaviour due to ensemble averaging. This improvement is obtained at the cost of increased training time as the number of trees increases

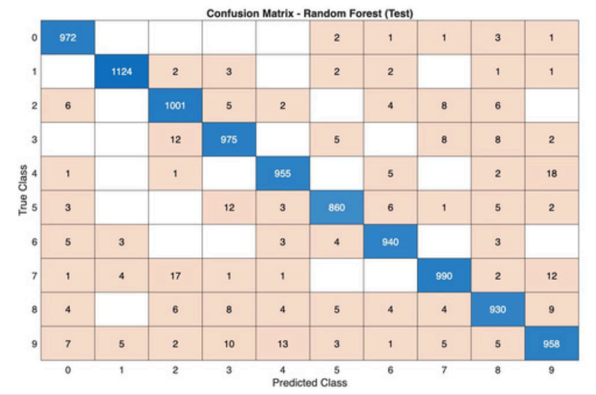


Model	Best Hyperparameter	Test Accuracy	CV Accuracy	Train Accuracy	Test Time	CV Time
Decision Tree	800	0.872	0.861	0.907	3.0 s	53.7 s
Random Forest	400	0.971	0.969	1	138.1 s	1114.8 s



Confusion Matrix - Decision Tree (Test)

True Class \ Predicted Class	0	1	2	3	4	5	6	7	8	9
0	917	1	9	6	5	11	7	5	14	5
1		1083	13	14	2	5	2	5	11	
2	14	13	869	22	17	11	11	34	31	10
3	5	4	39	852	7	44	4	15	23	17
4	5	2	7	6	838	12	10	8	26	68
5	12	8	3	50	13	731	20	10	24	21
6	23	6	12	5	32	18	841	2	17	2
7	4	12	25	12	9	7	1	924	9	25
8	6	9	25	34	20	30	18	10	786	36
9	14	1	5	18	31	15	4	20	26	875



Confusion Matrix - Random Forest (Test)

True Class \ Predicted Class	0	1	2	3	4	5	6	7	8	9
0	972					2	1	1	3	1
1		1124	2	3		2	2		1	1
2	6		1001	5	2		4	8	6	
3			12	975		5		8	8	2
4	1		1		955		5		2	18
5	3			12	3	860	6	1	5	2
6	5	3			3	4	940		3	
7	1	4	17	1	1			990	2	12
8	4		6	8	4	5	4	4	930	9
9	7	5	2	10	13	3	1	5	5	958

7. Result Analysis and Discussion

7.1 Decision Tree

- It exhibits a clear bias–variance trade-off [4]: increasing tree depth improves training accuracy but loses cross-validation gains, indicating increased variance and overfitting
- Generalisation performance is weaker than that of the Random Forest, reflecting the limitations of a single-tree model for high-dimensional, non-linear data [3]

7.2 Random Forest

- Achieves stronger generalisation through variance reduction via ensemble averaging of decorrelated trees, resulting in higher and more stable accuracy
- The smaller gap between training and test accuracy indicates improved robustness to overfitting [2], achieved at increased but tractable computational cost

7.3 Confusion Matrix

- Most misclassifications occur between visually similar digits, suggesting errors arise from inherent ambiguity rather than noise
- The Random Forest exhibits fewer such confusions than the Decision Tree, consistent with its superior generalisation

8. Future Direction

8.1 Lesson Learned

This analysis shows that ensemble methods, particularly Random Forests, generalise better than a single Decision Tree on high-dimensional image data. It also demonstrates that increased model complexity leads to diminishing returns [4], highlighting the importance of cross-validation for controlling overfitting

8.2 Future Direction

- Apply PCA during preprocessing to reduce dimensionality and noise in the image data, improving computational efficiency and enhancing classification performance
- Evaluate probabilistic and unsupervised methods such as Naive Bayes and K-means to assess how distributional assumptions and latent data structure influence classification performance

8.3 Reference

[1] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278–2324, 1998.
[2] L. Breiman, "Random Forests," Machine Learning, Vol. 45, No. 1, pp. 5–32, 2001.
[3] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees,
Wadsworth International Group, 1984.
[4] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.
[5] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1143, 1995.