Handwritten Digit Classification Report

Rishabh Acharya[1*]
Ayush Pekamwar[1†]
Raj Nandan Singh[1‡]
Ankit Kumar[1§]
Pujit Jha[1¶]

[1]Indian Institute of Technology, Jodhpur
{b22cs090, b22ee084, b22ee052, b22cs076, b22cs091} @iitj.ac.in

Abstract

This report presents a methodical approach to enhancing classification accuracy on the MNIST dataset, and using it to classify handwritten digits into one of the ten decimal digits. Leveraging four distinct feature extractors (LDA, PCA, Edge Detection, and a Custom Feature Extractor) and various classifiers, we systematically evaluated and refined models based on performance metrics and computational efficiency. Through rigorous testing and validation, including data augmentation to bolster training samples, we arrived at an optimized model: Custom feature extractor paired with a Random Forest classifier and augmented data. This model achieved an impressive accuracy of 97.78% and demonstrated robustness when applied to real images. Our findings underscore the importance of thoughtful feature extraction, model selection, and data augmentation in improving classification outcomes.

Keywords: MNIST, digit classification, feature extractor, data augmentation, predictions

*Roll Number: B22CS090
†Roll Number: B22EE084
‡Roll Number: B22EE052
§Roll Number: B22CS076
¶Roll Number: B22CS091

Contents

1 Introduction

1.1 Problem

Handwritten digit classification using the MNIST dataset for training. The aim of this project is to classify handwritten digits into one of 10 classes (each decimal digit).

1.2 Data Preprocessing

The MNIST dataset already contains separate training and testing data. We have 60,000 samples as a part of the training dataset and another 10,000 samples as a part of the testing dataset. We resized the dataset to 28x28 for better visualization on Matplotlib and to be able to use it on our custom feature extractor. The number of unique labels in both datasets was also visualized.

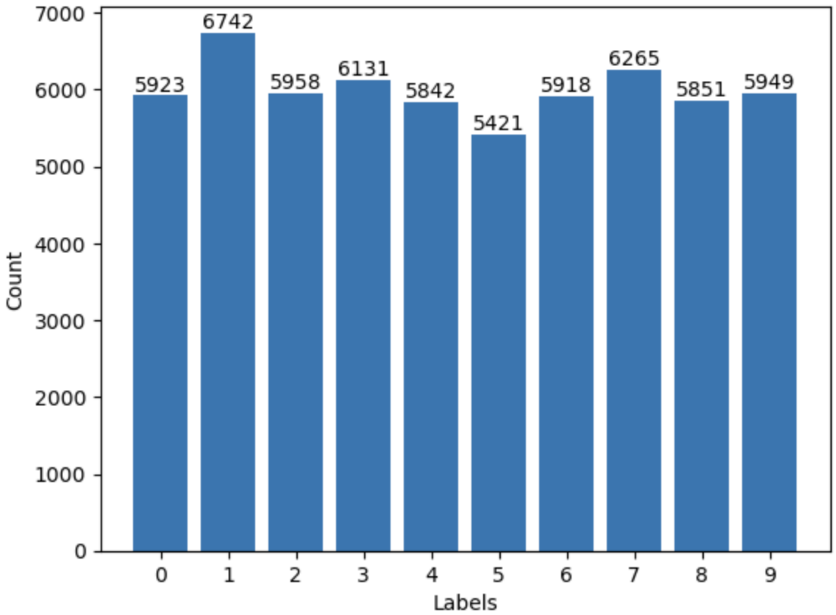


Figure 1: Unique Labels in Training Dataset

As the final step of preprocessing, we normalized the datasets.

1.3 Major Findings

- We have implemented a **custom feature extractor** (working explained later), and the accuracies using this extractor turned out to be better than those by pre-established techniques such as LDA and PCA.
- The implementation of data augmentation to generate a training dataset containing ten times the number of samples as in the original MNIST training dataset made our model much more robust and generalized, and generated a higher accuracy.

1.4 Structure of the Report

This report is structured as follows:

- **Section 2:** A discussion of various implementation ideas and the rationale behind choosing or rejecting specific approaches.
- **Section 3:** Presentation of accuracies achieved using different approaches, along with visualizations of the outputs. Additionally, we discuss training data augmentation, introduce the final selected model, and demonstrate its performance on a user-provided image.

- **Section 4:** A discussion of the interface designed for users to interact with our model, and its constraints.
- **Section 5:** A brief overview of the entire model and our methodologies.

2 Approaches Tried

2.1 Feature Extractors and Dimensionality Reduction Techniques

Using various feature extractors and dimensionality reduction techniques, we created the following 5 datasets:

2.1.1 Original dataset

This was the normalized training dataset with no additional feature extractors.

2.1.2 Dataset using PCA

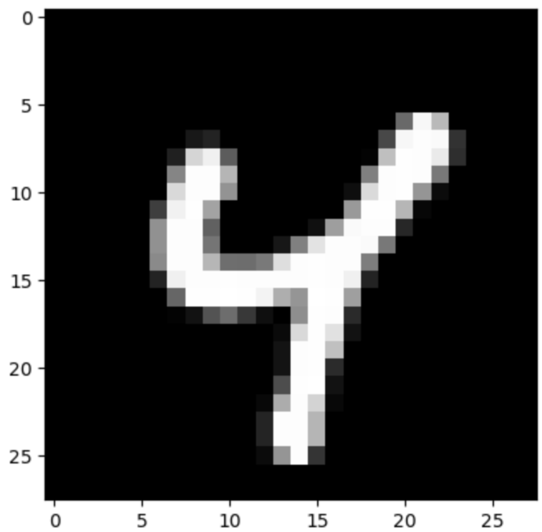
Using `n_components=0.98`, we transformed the training data into 459 components in order to achieve dimensionality reduction. We used 0.98 as the variance retention percentage was high (98%) enough to represent a reliable version of the original dataset, while not being computationally expensive.

2.1.3 Dataset using LDA

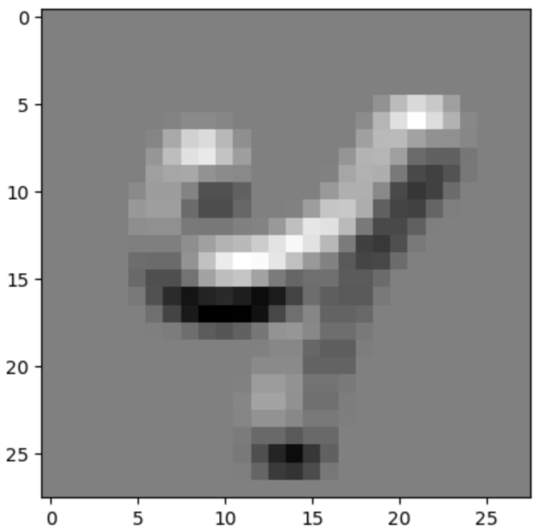
The original dataset was transformed using SK Learn's LDA.

2.1.4 Dataset using Edge Detector

Once the dataset was converted into a 28x28 array, the Prewitt operator (`prewitt_h`) was used for horizontal edge detection in each image. The fundamental working of this operator, which is a single convolution operator, is that it highlights those pixels whose difference in the values of neighboring pixels is the highest.



(a) A sample in the training dataset



(b) The same sample with edge detector applied

Figure 2: Images 2 and 3

2.1.5 Custom Feature Extractor

Pixels on the edge and in the core of the digit were given different weights. For values ranging from 1 to 199, it assigned the value 0.75, while for 200 to 255, 1 was assigned. The idea was to assign different priorities to points present on the edge and those present in the center.

Figure 3: Visualization of the custom feature extractor

2.2 Approaches to obtain optimal model

2.2.1 KNN Classifier

The first classifier we tried was K-Nearest Neighbours. We obtained the following test accuracies with it:

- Original dataset: 0.9441
- Edge detector dataset: 0.9468
- Custom feature extractor dataset: 0.9672

As we can see, we cannot reject any feature extractor using KNN as the values of obtained accuracies are nearly equal.

2.2.2 Decision Tree Classifier

- Our initial idea behind using PCA was to pair it with a Decision Tree (DT) classifier. This was because PCA decorrelates features and DTs may sometimes struggle with highly correlated features because they may redundantly split on similar information, leading to overfitting.

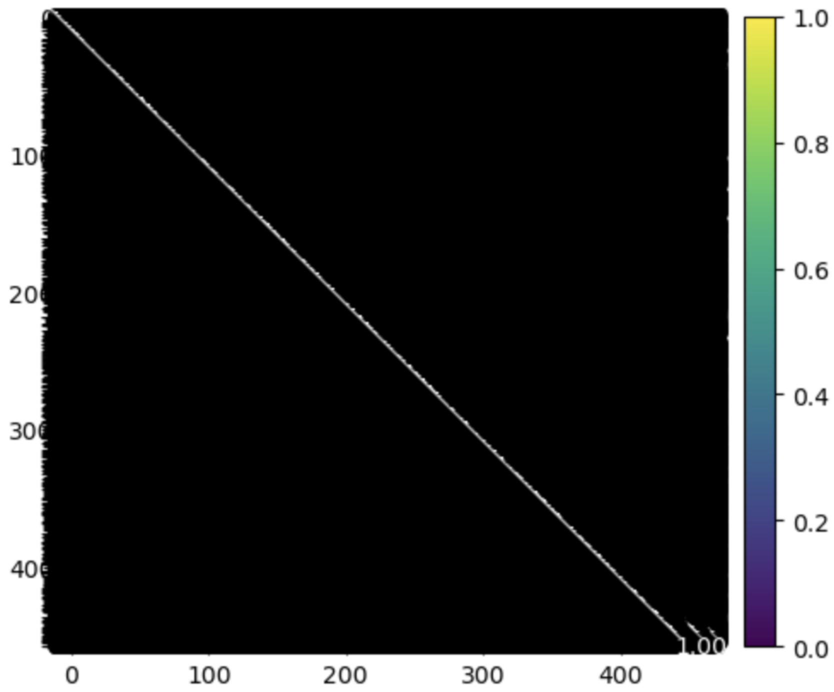


Figure 4: Correlation Matrix

However, this did not produce desirable results as we could not prevent overfitting (training accuracy=1.0). The accuracy on the test set was 0.8215.

- In order to improve on this number, we used grid search with a reduced depth of the DT (to prevent overfitting) to obtain best hyperparameters (ideal depth, which turned out to be 14). The test accuracy obtained in this case was 0.8333. This is only a slight improvement.

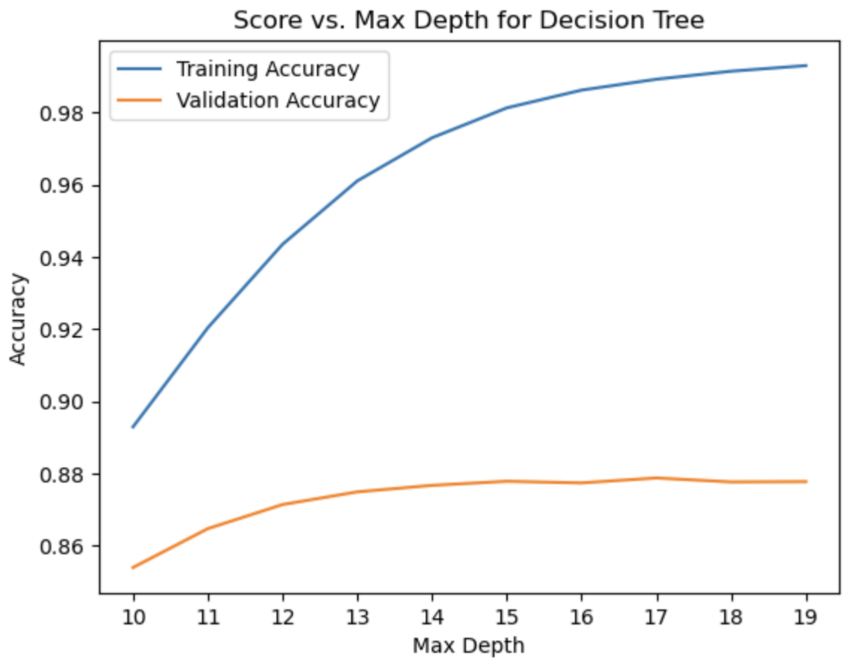


Figure 5

- As we can see in this graph, as the depth increases, there is clear overfitting while the validation accuracy stagnates. As a result, we rejected the combination of PCA+DT.

- To compare the performance of DT with and without PCA, we applied the classifier to our normalized dataset. The obtained accuracy was 0.882. This is not ideal either.
- Lastly, we used grid search on our custom dataset to obtain the ideal depth of the DT, which turned out to be 17, as can be seen in this figure.

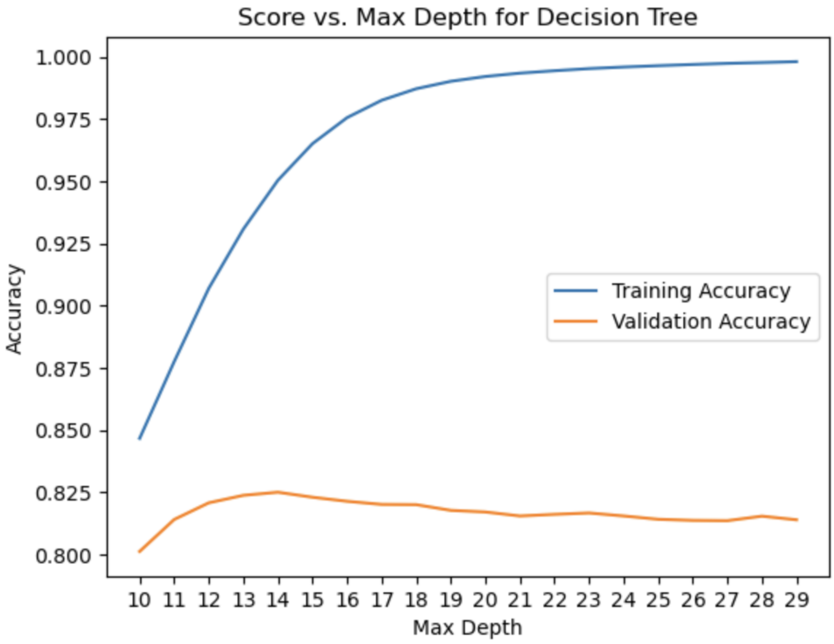


Figure 6

The accuracy of this combination was 0.8917, which, while better than other results, was still not nearly good enough.

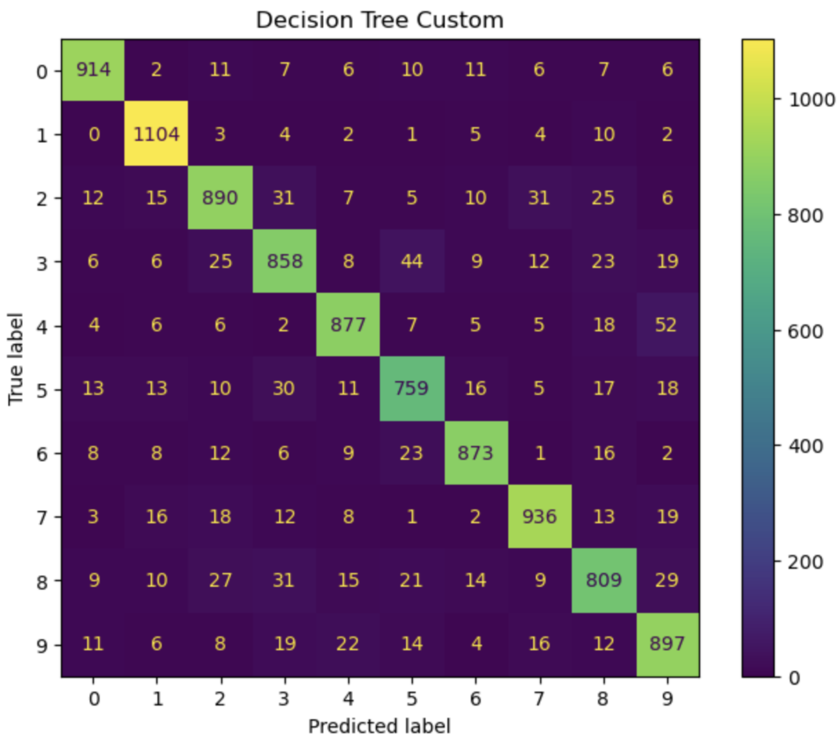


Figure 7: Confusion Matrix

Note: We have made similar confusion matrices for all classifiers, but only a few have been shown in this report to keep it concise.

2.2.3 Logistic Regression Classifier

- On the original normalized data, we got a testing accuracy of 0.9258.
- Edge detector feature extractor gave an abnormally low accuracy (0.1135) with logistic regression classifier. As a result, we rejected this feature extractor while moving forward.
- LR classifier gave a test accuracy of 0.925 on the PCA dataset. However, we rejected PCA feature extractor because:
  - The use of PCA on the DT classifier resulted in overfitting, thereby leading to undesirable accuracies. This showed that PCA is not consistent for all classifiers, and cannot be universally applied.
  - The time taken to train models on original dataset is not much longer as compared to that of PCA dataset.
- The LDA dataset was also rejected here as the dataset does not conform to the assumptions made by LDA (the features within classes are not normally distributed, and the classes do not have identical covariance matrices).
- With our custom feature extractor, LR produced a testing accuracy of 0.9242, which is higher than that by other feature extraction methods but the training time is also slightly higher.

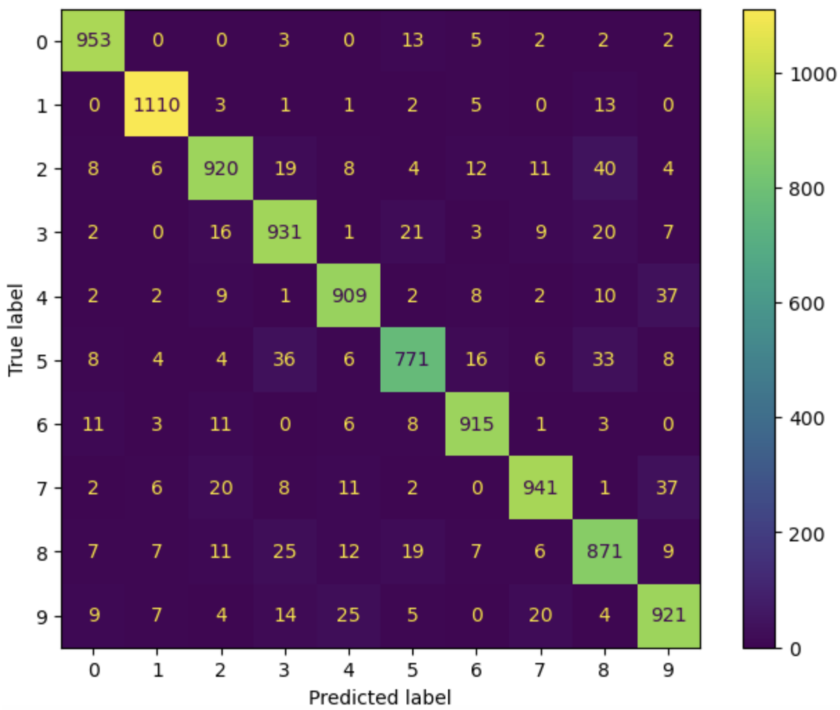


Figure 8: Confusion Matrix

After using logistic regression and rejecting feature extractors based on the aforementioned reasons, we are left with these 2 datasets:

- Original dataset (just normalization)
- Dataset which uses our custom feature extractor

  These 2 datasets consistently provide great accuracy with all models. Hence, only these will be used with models which take longer time to train.

2.2.4 Naive Bayes Classifiers

- Both Gaussian Naive Bayes and Multinomial Naive Bayes Classifiers produce remarkably low accuracies. This is because the fundamental assumptions followed by these classifiers are violated in the dataset. These assumptions include:
  - Features within each class follow normal / multinomial distribution.
  - Features are conditionally independent.
- Accuracies of Gaussian Naive Bayes classifier on:
  - Original dataset: 0.5558
  - Custom dataset: 0.5484
- Accuracies of Multinomial Naive Bayes classifier on:
  - Original dataset: 0.8236
  - Custom dataset: 0.8365

2.2.5 Random Forest Classifier

- The Random Forest Classifier was instantiated with the following parameters:
  - nestimators=200: This parameter sets the number of decision trees in the random forest ensemble to 200. Each tree in the forest is trained on a random subset of the training data and contributes to the final prediction through a voting or averaging mechanism.
  - bootstrap=True: When bootstrap is set to True, each decision tree in the random forest is trained on a bootstrap sample of the training data. This means that the algorithm randomly samples data points with replacement, creating multiple subsets for training each tree.
  - maxsamples=0.9: The maxsamples parameter specifies the maximum proportion of samples (data points) to be used for training each decision tree. Here, maxsamples=0.9 means that each tree is trained on a random subset of 90% of the training data, selected with replacement.
  - njobs=12: The njobs parameter controls the number of CPU cores to use during training. Setting njobs=12 means that the random forest classifier will utilize 12 parallel processes (if available) for training, potentially speeding up the training process significantly, especially for large datasets.
- Testing accuracy of this classifier on:
  - Original dataset: 0.9613
  - Custom dataset: 0.9718
- **Random Forest classifier has produced the best accuracy so far, with a peak of 97.18% when coupled with our custom feature extractor.**

2.2.6 AdaBoost Classifier

- AdaBoost (Adaptive Boosting) is an ensemble learning algorithm designed for classification tasks. It operates by sequentially training a series of weak learners, such as decision trees, on the dataset. Each weak learner focuses on instances that previous learners classified incorrectly, adjusting their weights to prioritize these instances. The final prediction is made by combining the predictions of all weak learners, weighted by their individual accuracy. By iteratively emphasizing difficult-to-classify instances, AdaBoost aims to create a strong learner from a collection of weak learners, thereby improving classification accuracy.
- Testing accuracy of AdaBoost classifier on:
  - Original dataset: 0.6201
  - Custom dataset: 0.8542

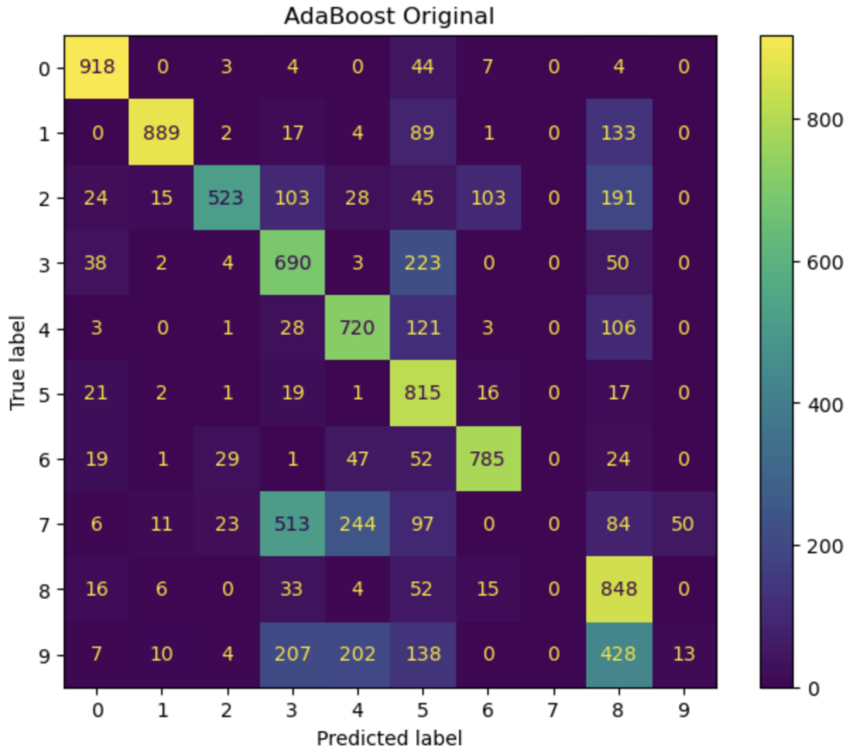


Figure 9: Confusion Matrix

- The accuracies obtained by this classifier are significantly lower than what we have already achieved.

2.2.7 Histogram Gradient Boosting Classifier

- The Histogram Gradient Boosting Classifier is a variant of the Gradient Boosting algorithm, optimized for large datasets. It operates by iteratively fitting decision trees to the residuals of the previous trees, gradually reducing prediction errors. Unlike traditional gradient boosting, which splits nodes based on the exact feature values, the histogram approach bins features into histograms, reducing computational complexity. The final prediction is a weighted sum of the predictions from all trees, with each tree correcting errors from previous trees. This algorithm is efficient for big data scenarios and can achieve high predictive accuracy.
- Testing accuracy of this classifier on:
  - Original dataset: 0.9452
  - Custom dataset: 0.9808
- **The accuracy of this classifier on the dataset produced by using our custom feature extractor is the highest we have achieved.**

2.2.8 Support Vector Machine (SVM) Classifier

- The RBF kernel was used.
- Testing accuracy of RBF kernel on:
  - Original dataset: 0.9612
  - Custom dataset: 0.9791

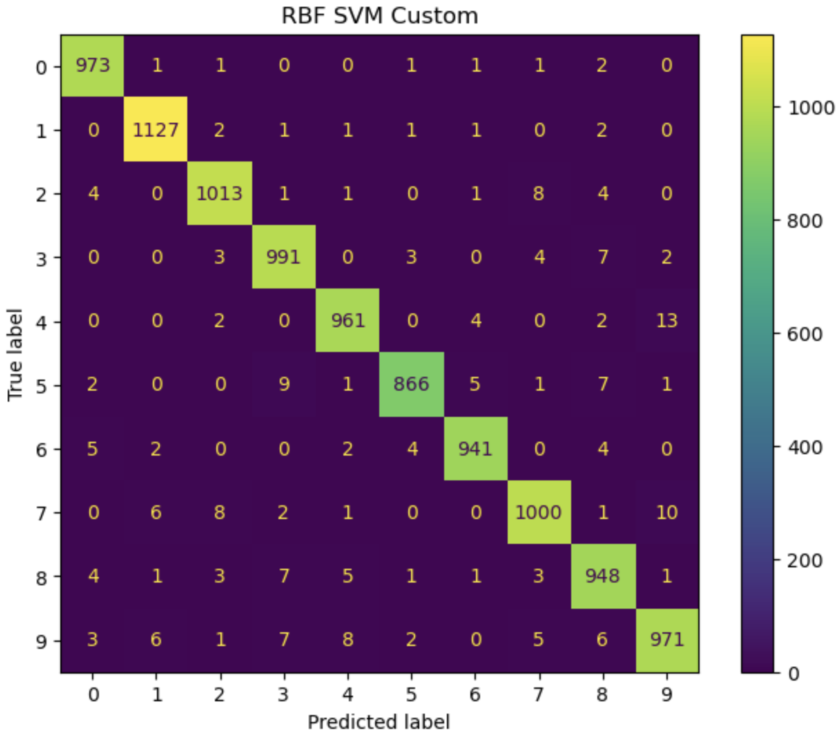


Figure 10: Confusion Matrix

- While these accuracies are appreciable, we have even better results at our disposal.

2.3 Optimal Models

- The following models produced acceptable testing accuracies:
  - Custom feature extractor + KNN: 0.9672
  - Custom feature extractor + Random Forest: 0.9718
  - Custom feature extractor + SVM Classifier (RBF Kernel): 0.9791
  - Custom feature extractor + Histogram Gradient Boost Classifier: 0.9808
- Among these, the model using SVM was rejected because the training time was too high (868.2s). Furthermore, if we are to use augmented data to increase the number of training samples, the time taken to train would significantly rise. This is not feasible for practical purposes.
- To finalize a model, we tried out ensemble voting and even augmenting the training set. This has been discussed in section 3 2 of this report.

3 Experiments and Results

3.1 Ensemble Voting

- Ensemble voting is a technique used in ensemble learning, where multiple individual models (e.g., classifiers or regressors) make predictions on the same dataset, and the final prediction is determined by combining or voting on the predictions of these individual models.
  The types of ensemble voting algorithms we used include:
  - Hard voting: Each individual model in the ensemble makes a prediction, and the final prediction is the class that receives the most votes (i.e., the majority of models predict that class).
  - Soft voting: Instead of making a binary decision based on a majority vote, soft voting takes into account the probability scores predicted by each model. The final prediction is often the class with the highest average probability across all models.
- We performed both of these on the final 3 models which we had obtained. The results were as follows:
  - Hard voting: 0.977
  - Soft voting: 0.9806
- As can be seen, while the final testing accuracies obtained by both types of ensemble voting were high enough to be accepted, they were not the highest values we had obtained. Hence, we discarded this concept.

3.2 Image Augmentation

- Image augmentation refers to the process of artificially expanding a dataset for training machine learning models by applying various transformations to existing images.
  These are the primary objectives of image augmentation:
  - Increasing the size of the dataset: Image augmentation effectively increases the size of the training dataset by generating new images featuring variations. This often results in enhanced model performance and generalization.
  - Enhancing robustness: The inclusion of augmented images in the training data introduces diversity and variability, thereby fortifying the model against the fluctuations and noise that are inherent in real-world images.
  - Preventing overfitting: Augmentation can function as a regularization technique to mitigate overfitting by introducing variability and noise into the training data. This introduces complexity and discourages the model from retaining minute details of the training images, thereby preventing overfitting.
  - Improving generalization: Augmenting the diversity of the training dataset enables the model to acquire a greater number of generalized features and patterns, thereby enhancing its capability to process novel, unseen images during the inference process.
- Our approach to image augmentation:
  - Every image will be randomly rotated in either clockwise or counterclockwise direction within an angle of 15 degrees from the original axes.
  - This rotation will be coupled with a translation in one of 8 directions (up, down, left, right, left along main diagonal, right along main diagonal, left along counter diagonal, right along counter diagonal), or no translation at all. This leads to 9 possible combinations.

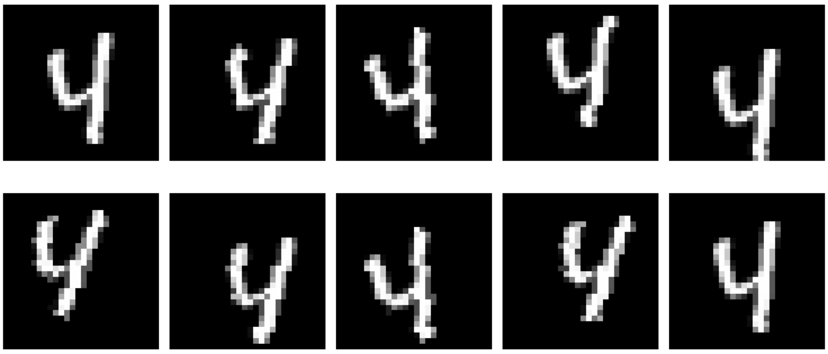


Figure 11: Variations observed in a sample after augmentation

  - Given that the size of the training dataset is 60,000 samples, this image augmentation procedure produced 5,40,000 additional samples, resulting in a total of 6,00,000 training samples.

- We had arrived at 3 optimal models at the end of section 2. Post image augmentation, we rejected the model which used KNN classifier as the memory requirement for it would be tremendously high in this new training dataset which had 10 times the number of samples as the original dataset.

- We loaded the remaining 2 models on this new training dataset and computed the testing accuracies, which turned out to be as follows:

  - Random forest model: 0.9778
  - Histogram gradient boosting model: 0.9671

Model	Image Augmentation	Testing Accuracy
Random Forest	Yes	0.9778
Random Forest	No	0.9718
Histogram Gradient Boost	Yes	0.9671
Histogram Gradient Boost	No	0.9808

Figure 12: Comparative Table

- Figure 12 summarizes the testing accuracies on the 4 models we have at hand currently.
  As we can see, the accuracy increases for the Random Forest model on performing image augmentation but it decreases for the Histogram Gradient Boost model.

- Between the 2 models producing the highest accuracies, we preferred the Random Forest model because its accuracy is post image augmentation, unlike in the case of the Histogram Gradient Boost model. This would mean that the former is more suitable for generalized inputs and is the more robust model.

Final Result: The best model we have obtained involves using our custom feature extractor along with the Random Forest classifier and performing image augmentation on the training dataset. This model produces a testing accuracy of 97.78%.

3.3 Predictions on Real Images

3.3.1 Preprocessing

- The RGB image was first converted to greyscale the luminance method, which takes into account the human eye's sensitivity to different color channels.
  The formula used was: Grayscale=$0.2989\times$Red$+0.5870\times$Green$+0.1140\times$Blue
- Next, the negative of the entire image was taken. This involved subtracting the value of each pixel from the maximum value across all pixels.
- Once this was done, we performed the min-max scaling in a manner such that the pixel with least intensity gets assigned the value "0" and the one with the maximum intensity gets assigned the value "255".

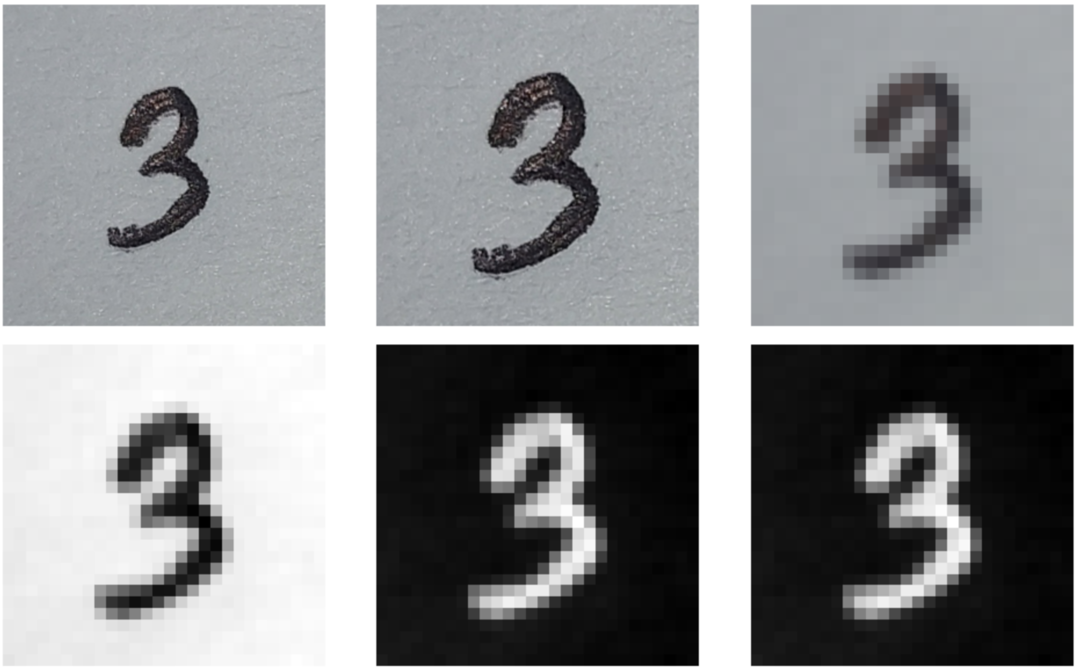


Figure 13

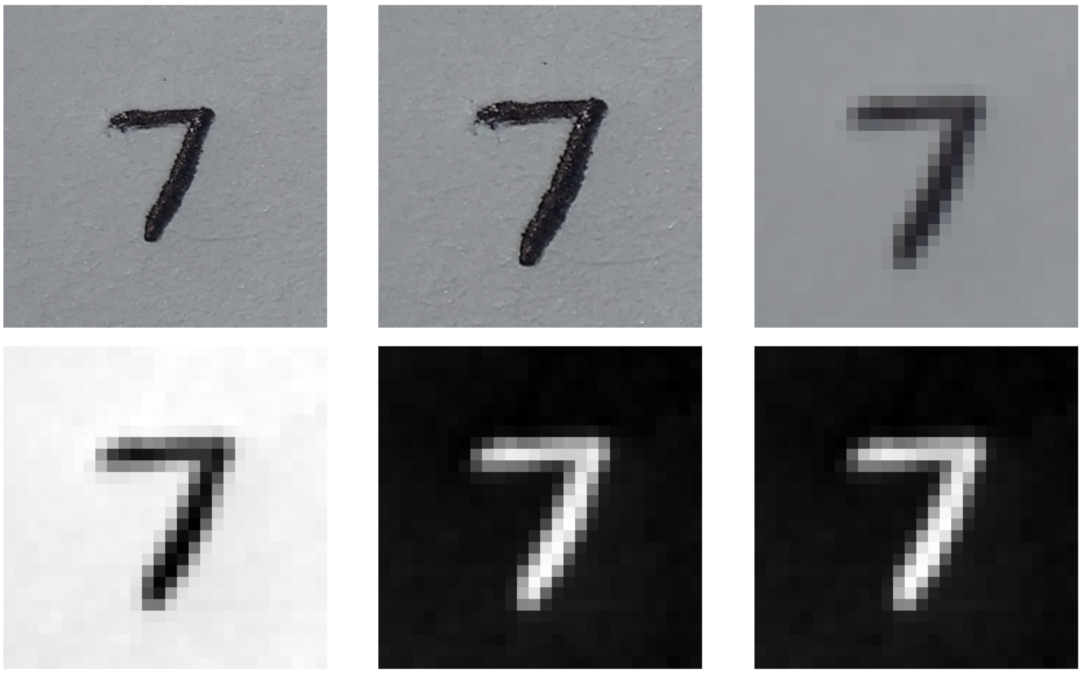


Figure 14

Each step of preprocessing is highlighted in image 13 and image 14

3.3.2 Feature Extraction

- Our custom feature extractor was deployed on the preprocessed images.

Figure 15

Figure 16

A visualization of the working of our feature extractor is shown in figures 15 and 16.

3.4 Predictions

- Lastly, our final model was used to make predictions on these images.
- The value with the highest probability is going to be predicted by the model.

	Figure 14	Figure 15
0	0.0166667	0.0
1	0.05666667	0.01333333
2	0.04	0.01
3	0.32	0.02666667
4	0.03666667	0.00666667
5	0.20666667	0.0
6	0.01333333	0.0
7	0.08	0.9333333
8	0.07666667	0.0
9	0.15333333	0.01

Figure 17: Probability of Each Digit

Figure 17 shows the probability of each digit when the model is deployed on figures 15 and 16.

- **The digit with the highest probability is predicted. Hence, the predictions are "3" and "7" respectively. Both predictions are correct.**

3.5 Failure Case Analysis

- We tried to analyze where our model is misclassifying images. To do so, we observed the model's confusion matrix, which is shown in figure 18

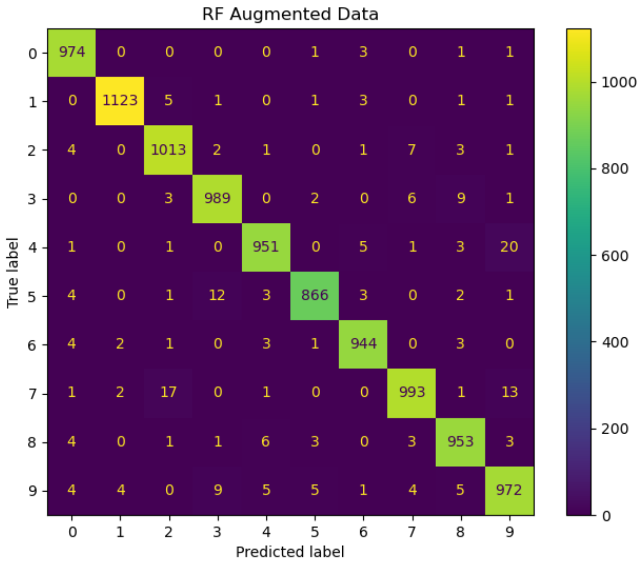


Figure 18: Confusion matrix of best model

- Based on the confusion matrix, we plotted the images of those categories where larger misclassifications were occurring.
- Our model could not distinguish between 5 and 3 in the images shown in figure 19, and ended up predicting them to be 3, which was incorrect. A possible explanation for this is that the images were very blurry and ambiguous, even to the naked eye.

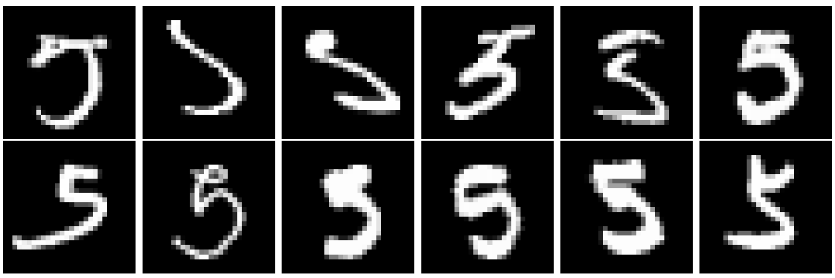


Figure 19: Samples in which 5 was misclassified as 3

- In the samples shown in figure 20, our model gave the output as 2 while it was supposed to be 7. This is perhaps due to the horizontal line in the samples, which might have been confusing to the classifier.

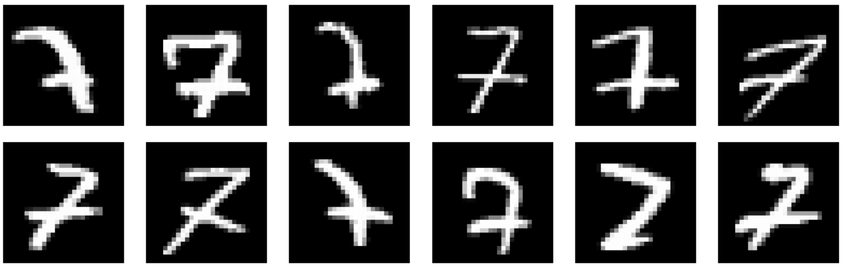
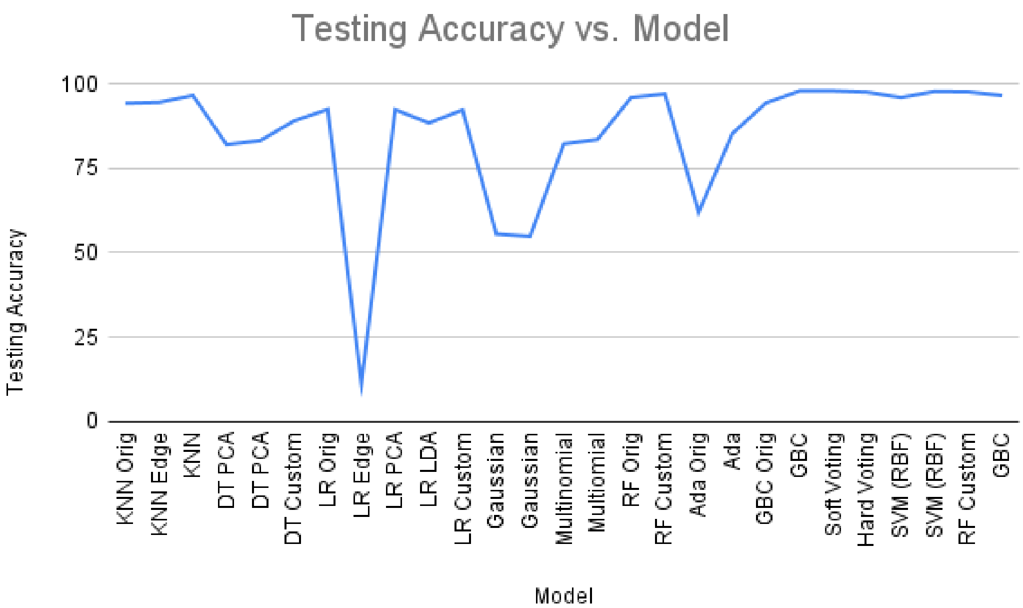


Figure 20: Samples in which 7 was misclassified as 2

- To conclude, our model primarily misclassified only those images which were unclear and ambiguous due to different handwriting styles.

3.6 Graphs and Tables



-

Figure 21: Testing Accuracy vs Model

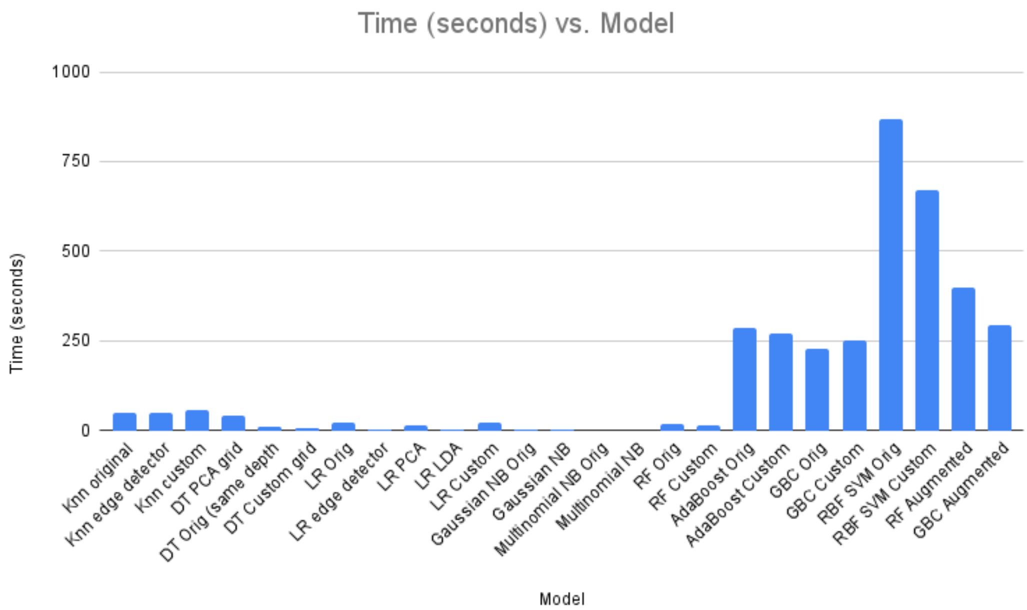


Figure 22: Training Time (s) vs Model

Label	Precision	Recall	F1 score	Support
0	0.98	0.99	0.99	980
1	0.99	0.99	0.99	1135
2	0.97	0.98	0.98	1032
3	0.98	0.98	0.98	1010
4	0.98	0.97	0.97	982
5	0.99	0.97	0.98	892
6	0.98	0.99	0.98	958
7	0.98	0.97	0.97	1028
8	0.97	0.98	0.97	974
9	0.96	0.96	0.96	1009
Macro average	**0.98**	**0.98**	**0.98**	**10000**
Weighted average	**0.98**	**0.98**	**0.98**	**10000**

Figure 23: Classification report of best model

4 Interface

5 Summary

Brief summary of the project:

- Samples from the MNIST dataset were pre-processed initially.
- Using 4 feature extractors (LDA, PCA, Edge Detection and Custom Feature Extractor), we created 5 datasets.
- We checked the performance of various classifiers on each dataset, and then rejected different feature extractors based on either the testing accuracy scores or the fact that the hypothesis for some of the classifiers was violated in the MNIST dataset.
- Finally, we selected 4 models (combination of a feature extractor and a classifier) that were giving the best accuracies. One of these was rejected due to very high time consumption.
- To determine one best model, we tried ensemble voting (hard voting and soft voting). However, we did not obtain an improvement in the accuracy.
- Next, we performed data augmentation, and produced 5,40,000 additional training samples. Due to a very large training sample size, we were able to reject another model for its extremely high memory consumption was undesirable.
- We tested the remaining 2 models on this augmented training dataset, and compared the accuracies produced of both models with and without data augmentation. Keeping in mind the fact that augmentation makes the model more robust and generalized, we reached a final model.
- **Best model obtained: Custom feature extractor+ Random Forest classifier+ Data Augmentation. This model produced an accuracy of 97.78%.**
- Lastly, we tried this model on real images and displayed the predictions. Failure case analysis was also performed.

References

A Contribution of each member

1. **Rishabh Acharya**: Ideation and implementation of custom feature extractor; Data Augmentation; Training best models, Designing project page
2. **Ayush Pekamwar**: Training with Decision Trees and Naive Bayes models; Designed the website which works as an interface for showing working
3. **Raj Nandan Singh**: Training with Histogram Gradient Boosting and RBF SVM; Report documentation
4. **Ankit Kumar**: Training with Random Forest and AdaBoost; Making the presentation
5. **Pujit Jha**: Making new datasets with different feature extractors; Training with KNN and Logistic Regression classifiers; Preparing the report