Digit Classifier

Classify images of hand-drawn digits

Learning Objectives

- 2D Classification
- Model confusion
- Dataset bias
- Data augmentation

Visualising the Data

The frequency of each digit in the dataset was determined. All digits were represented roughly equally with ~175 occurrences each.

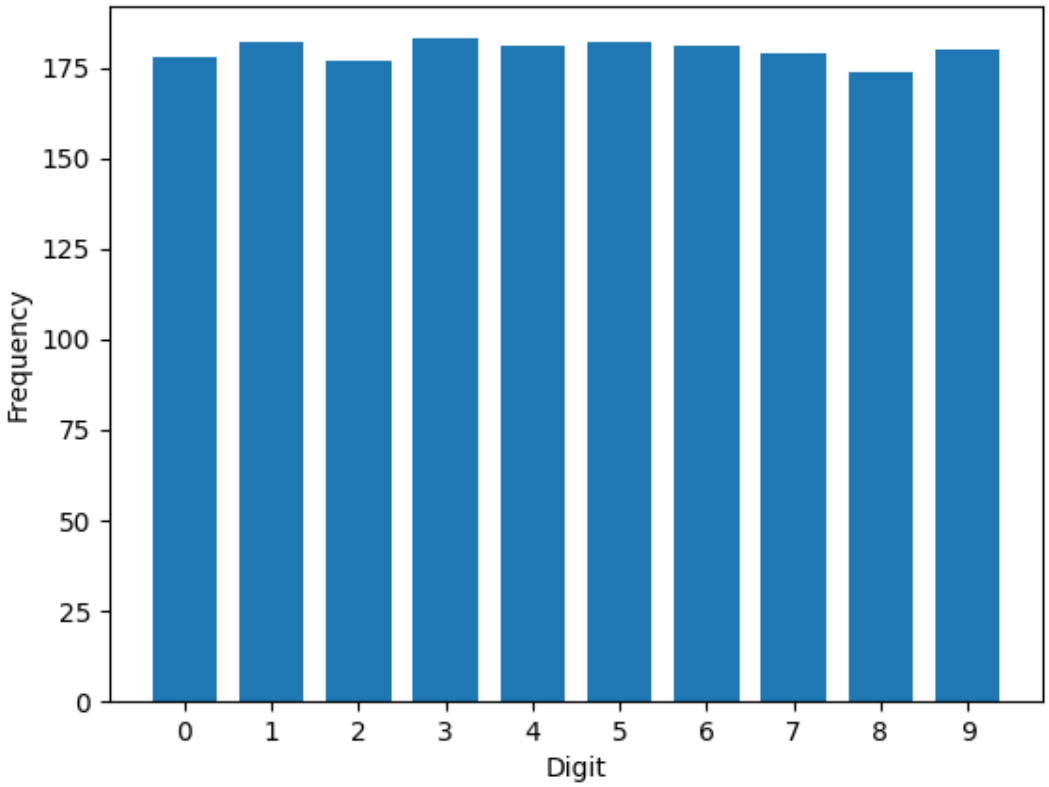
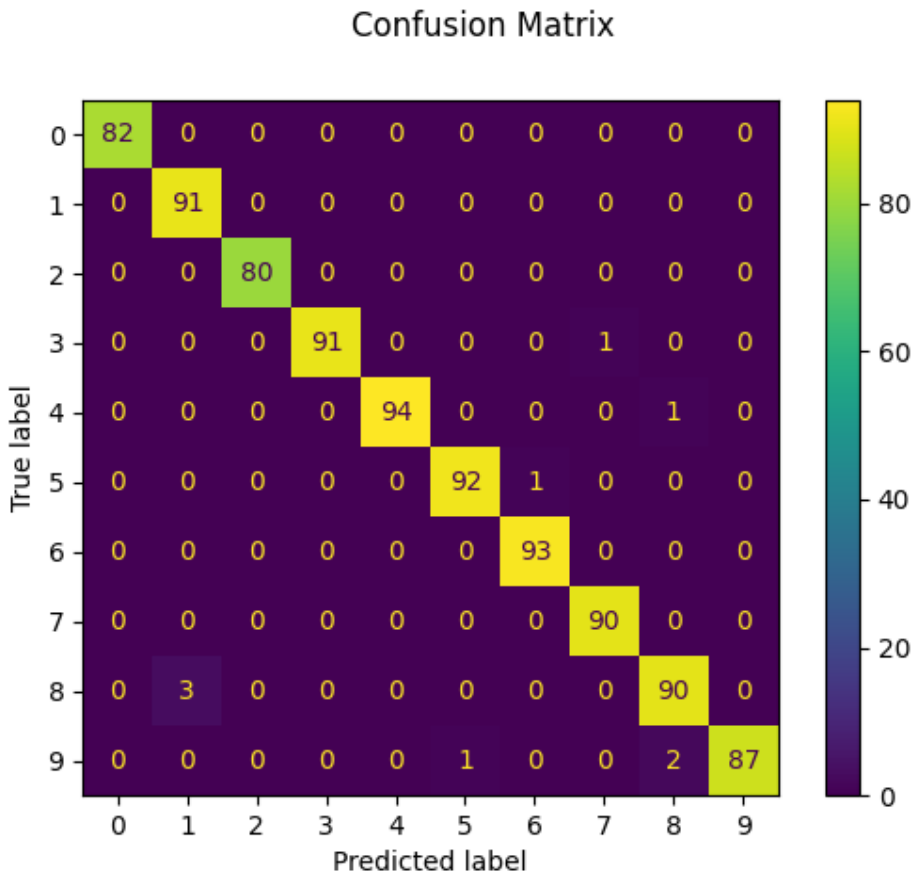
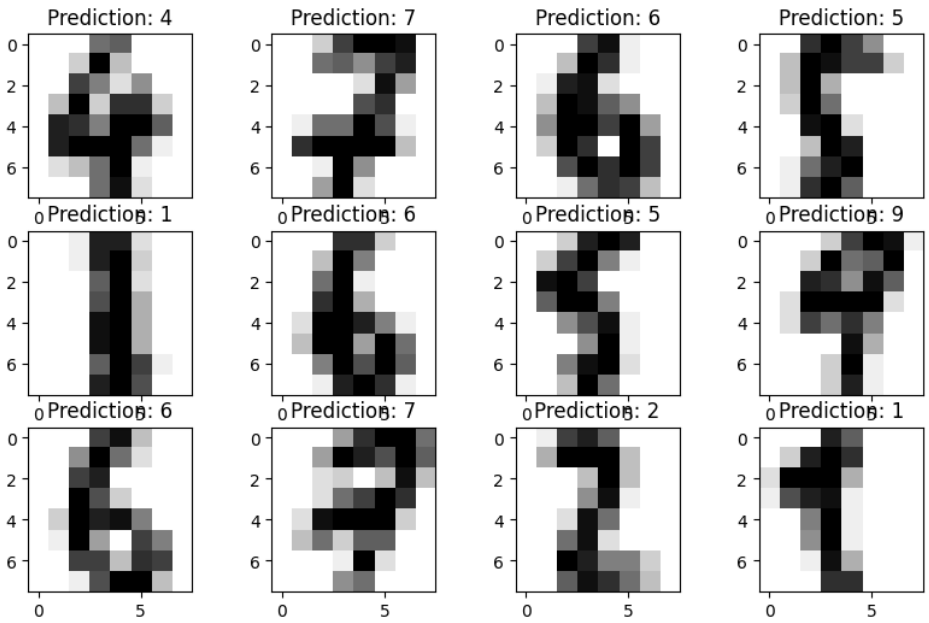


Image Classification

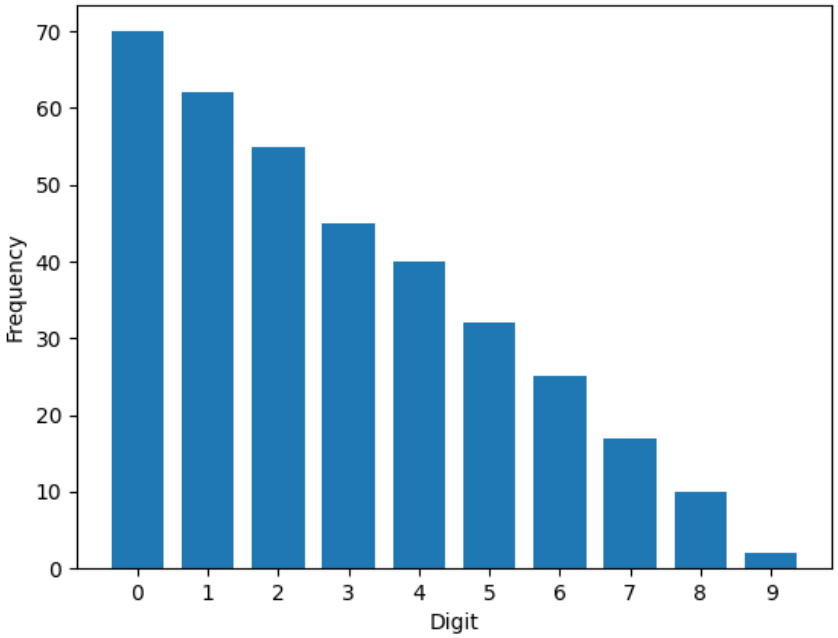
A SVM with an RBF kernel was trained on a dataset of 8x8 pixel grayscale images of hand-drawn digits (from 0-9). The dataset was pre-partitioned at random into training and testing sets. A confusion matrix from the testing phase is shown below, along with some of the images and the model's predictions. This model achieved a 99.0% accuracy.



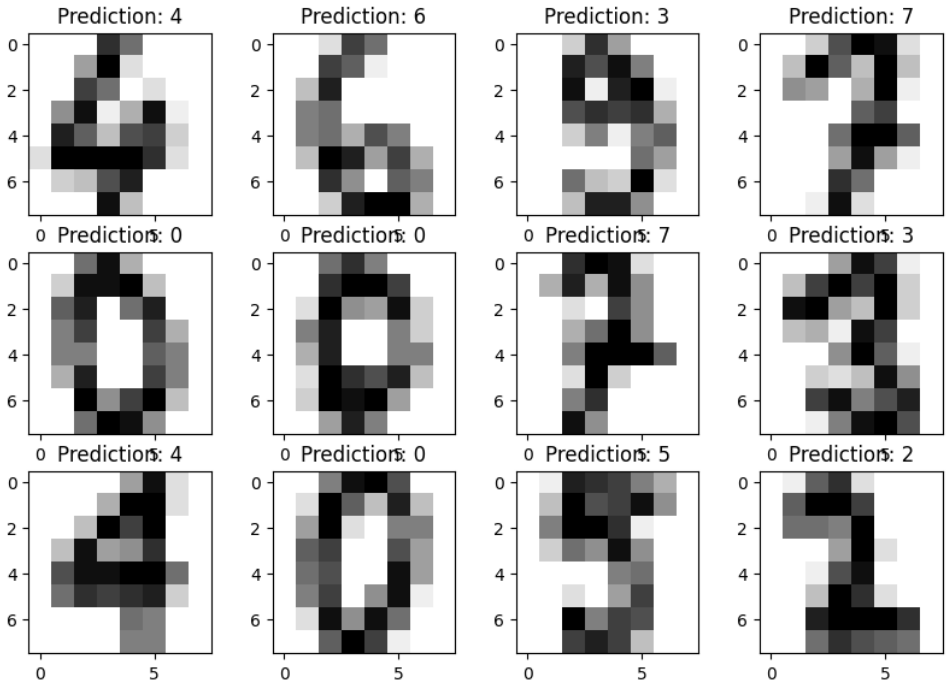


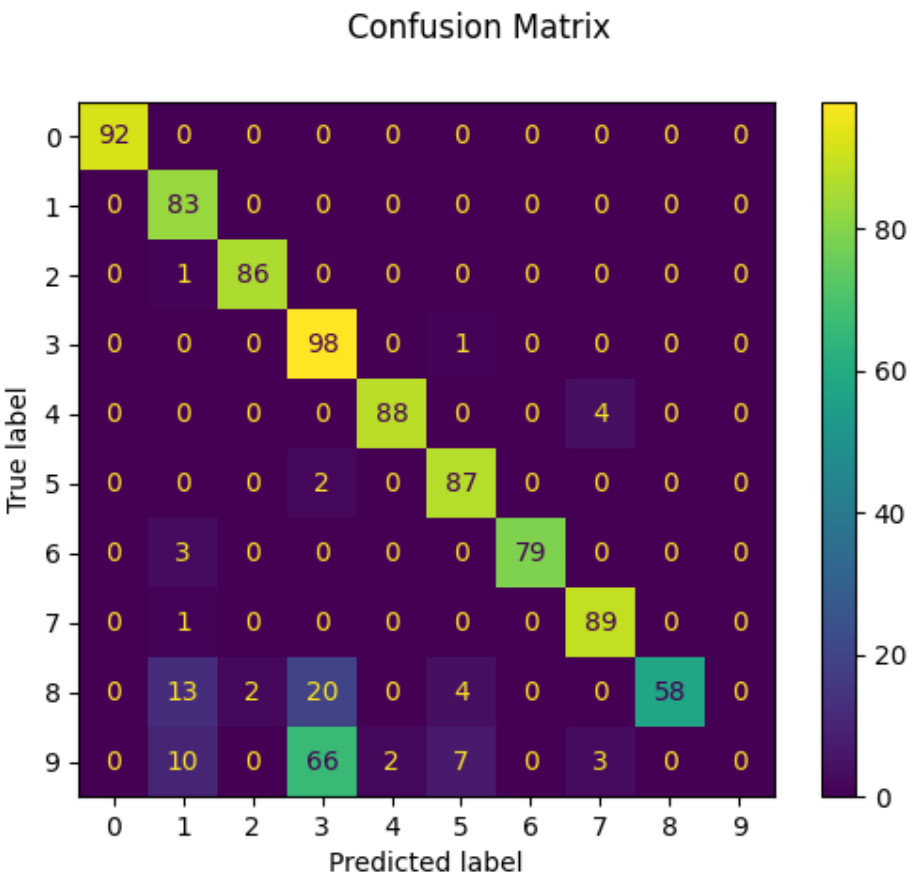
Dataset Bias

The training dataset was systematically biased by removing occurrences of specific digits. This resulted in some digits being underrepresented. Each increasing digit had fewer examples in the training dataset, as shown below.



As expected, this model misclassified the least represented digits as one of the more represented digits, achieving an overall accuracy of 84.5%. It particularly struggled with the digit 9 (which had only 3 occurrences in the training dataset), never correctly identifying the digit and misclassifying it as the digit 3 (which had 45 occurrences in the training dataset) 75% of the time.

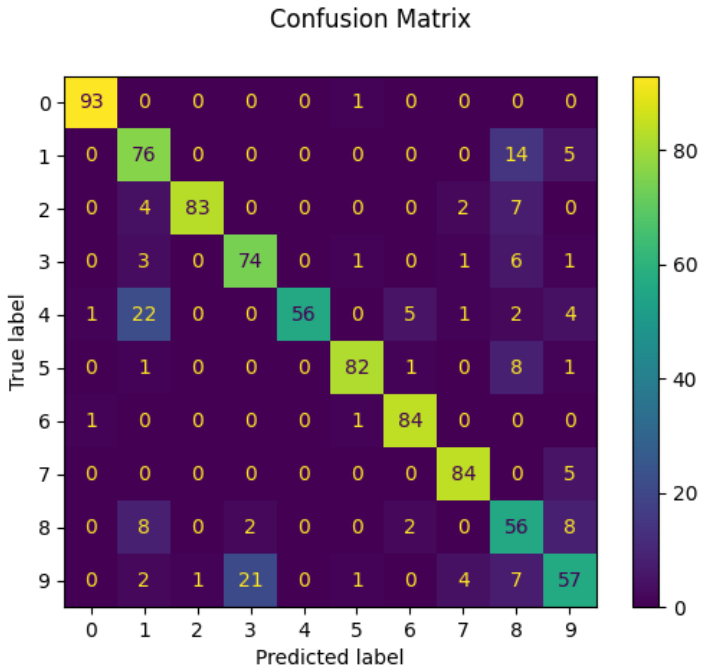



Confusion Matrix
True label
Predicted label
0 92 0 0 0 0 0 0 0 0 0
1 0 83 0 0 0 0 0 0 0 0
2 0 1 86 0 0 0 0 0 0 0
3 0 0 0 98 0 1 0 0 0 0
4 0 0 0 0 88 0 0 4 0 0
5 0 0 0 2 0 87 0 0 0 0
6 0 3 0 0 0 0 79 0 0 0
7 0 1 0 0 0 0 0 89 0 0
8 0 13 2 20 0 4 0 0 58 0
9 0 10 0 66 2 7 0 3 0 0
0 1 2 3 4 5 6 7 8 9
80
60
40
20
0

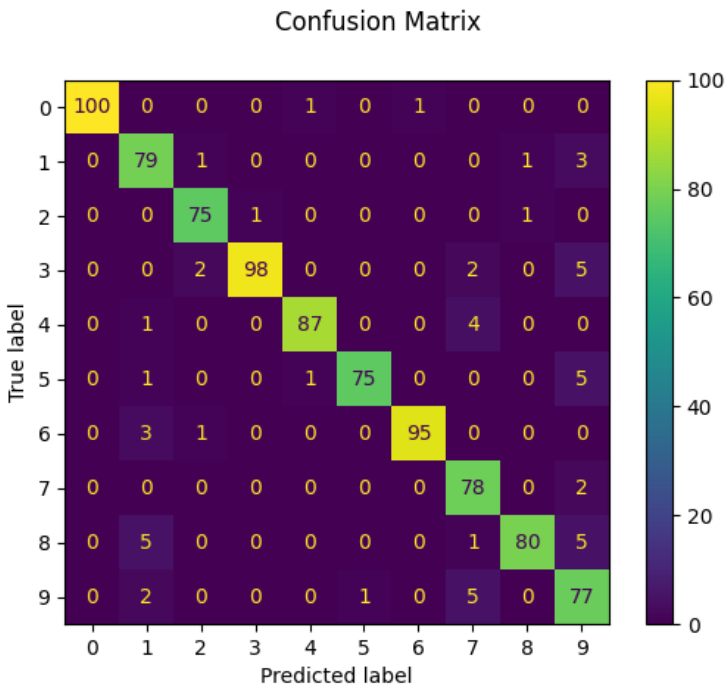
Data Augmentation

In data-poor regimes, creating synthetic data can improve the accuracy of predictive models.

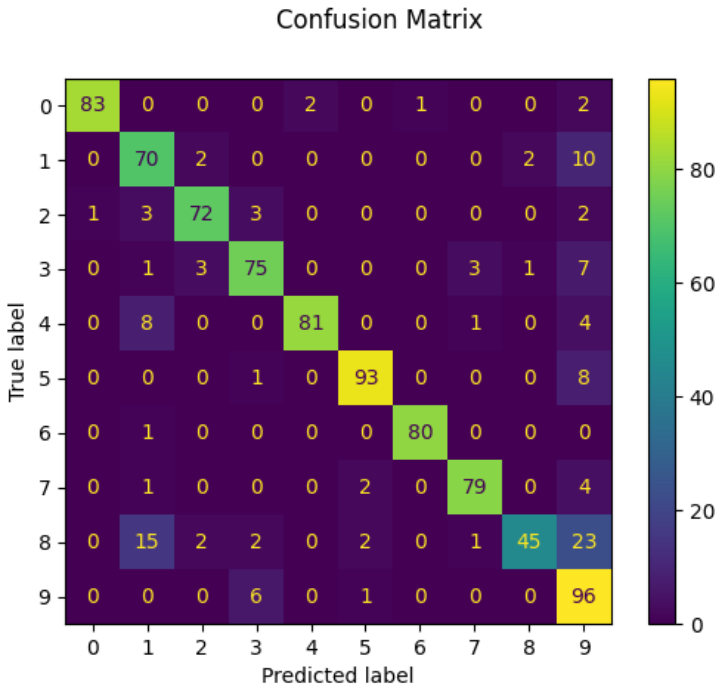
The training dataset was first cut down to only 10 examples of each digit, producing a model with 82.8% accuracy.

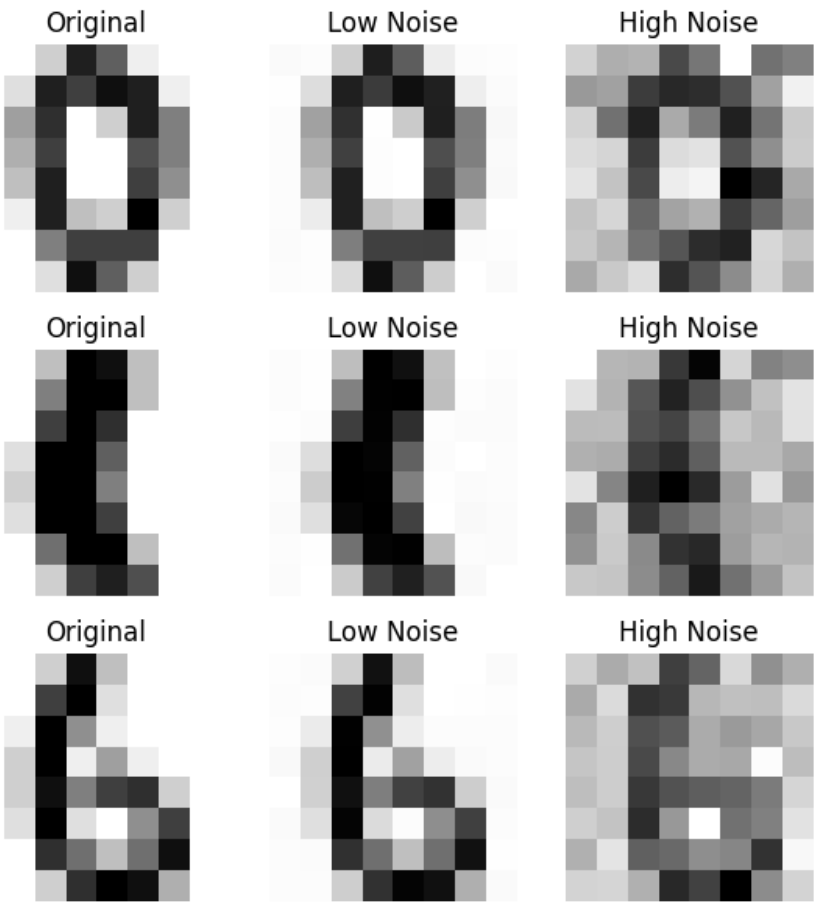


The training dataset was then expanded by generating 3 synthetic images from each original image, (giving a total of 40 examples of each digit). These synthetic images were obtained by adding Gaussian noise to the original 8x8 pixel image and normalising the result. When using a low level of noise (mean = 0.2, standard deviation = 0.1), this synthetic-data-boosted model achieved an accuracy of 93.7%.



However, when using a high level of noise (mean = 0.8, standard deviation = 0.2), the model only achieved an accuracy of 86.1%.





Conclusion

The quality and size of a dataset have a large impact in the performance of the resulting model. Biased datasets can produced skewed predictions, whilst models based on data-poor regimes often give an overall poor performance. Synthetic data can be used to increase the size of a dataset by slightly augmenting existing data, potentially improving the model's predictions. However, having a too high synthetic to original data ratio, or augmenting the underlying data too much could actually decrease the accuracy of the model.

Dataset	Testing Score
Full Dataset	0.9989
Biased Dataset	0.8445
Data-Poor Dataset	0.8277
3:1 Synthetic Dataset (Low Noise)	0.9367
3:1 Synthetic Dataset (High Noise)	0.8613