An Empirical Comparison of Supervised Learning in Handwritten Digits Classification

Yage G. Xin
A17304338
COGS 118A Final Report
`yaxin@ucsd.edu`

Abstract

We[1] conduct an empirical comparison of three supervised learning algorithms (logistic regression, decision trees, and linear support vector machines) reported in Caruana et al. (2008) on four handwritten digit datasets (Semeion, Optical Recognition, Pen-Based, and MNIST). Using multiple training/test splits and cross-validation for hyperparameter tuning, we evaluate model performance in terms of accuracy, generalization, and overfitting. Our results show that SVM generally achieves the highest accuracy, while decision trees exhibit greater overfitting, particularly on smaller or more variable datasets. Overall, model performance is robust across datasets, highlighting the influence of feature representation and dataset characteristics on classifier effectiveness.

1 Introduction

Handwritten digit recognition is a canonical problem in machine learning, providing a controlled setting for evaluating supervised classification algorithms. In this study, we systematically compare the performance of logistic regression, decision trees, and linear SVMs across four publicly available datasets: Semeion, Optical Recognition, Pen-Based, and MNIST. These datasets vary in size, input representation, and preprocessing, providing a range of challenges in terms of resolution, writer variability, and feature structure.

We focus on a binary classification task, distinguishing the digit "7" from all other digits, in order to provide a consistent evaluation framework and simplify comparison across datasets. By using multiple training/test splits and cross-validation for hyperparameter tuning, we investigate not only model accuracy but also the effects of dataset characteristics, hyperparameters, and overfitting on classifier performance. This study aims to identify patterns of generalization across diverse handwritten digit datasets and highlight practical considerations in model selection.

2 Methodology

2.1 Learning algorithms and training procedure

We compare logistic regression, decision trees, and linear SVMs across four datasets and three train/test splits: 20/80, 50/50, and 80/20. For each dataset and split, classifiers are trained and evaluated using 3-fold cross-validation to select optimal hyperparameters. After tuning, each model is retrained on the full training set and evaluated on the corresponding test set. Each experiment is repeated three times to account for variability in random splits, and mean performance metrics are reported. This procedure allows us to compare model performance, hyperparameter effects, and generalization patterns systematically across datasets.

2.1.1 Logistic regression

We search different regularization term C from the range {0.001,0.01,0.1,1,10}.

---

[1]First person plural is used in this report considering conventions of academic writing. I would like to clarify here that this is an individual project.

2.1.2 DECISION TREE

For comparison sake, we only fine tune the maximum depth parameter in the decision tree once. The parameter is set to 5 by cross validation on the first dataset, using a 50/50 training/testing split from the range of (1,11). By fixing the maximum depth parameter, we ensure that these decision trees used in different experiments are the same algorithm.

2.1.3 LINEAR SVM

We search different regularization term C from the range {0.001,0.01,0.1,1,10}.

2.2 DATASET

To evaluate how the classifiers respond to changes in task difficulty, we chose four handwritten-digit datasets that differ systematically in structure and complexity. These datasets are chosen in order to provide a consistent task across different sources.

Using multiple datasets allows us to test how well models generalize across variations in handwriting styles, digit resolution, and dataset size. It also provides a more robust assessment than relying on a single dataset, as models might overfit to specific dataset characteristics. By including all four datasets, we can compare model performance across datasets of varying difficulty, evaluate hyperparameter effects consistently, and draw conclusions that are not dataset-specific but generalizable to the handwritten digit recognition problem.

We converted the multi-class digit labels into a binary classification problem. Specifically, the labels are converted from 0-9 to "7" and "not 7". This allows us to examine model performance on a simplified task. This binary conversion also facilitates comparison of model behavior across datasets, highlighting differences in learning patterns and sensitivity to hyperparameters.

2.2.1 SEMEION HANDWRITTEN DIGIT

The Semeion dataset contains 1,593 handwritten digit images collected from approximately 80 writers (sem, 1998). Each digit was scanned and normalized to a 16×16 grayscale image, then binarized using a fixed threshold to convert each pixel into a 0/1 value. For the rest of the report, this dataset is referred to as "Semeion".

2.2.2 OPTICAL RECOGNITION OF HANDWRITTEN DIGITS

This is a dataset with size 1797 (Alpaydin & Kaynak, 1998), samples being handwritten digits. The raw 32×32 bitmaps were divided into 4×4 non-overlapping blocks, and the number of active ("on") pixels in each block was computed. This produces an 8×8 representation where each pixel takes an integer value from 0 to 16, providing a compact and distortion-tolerant encoding. For the rest of the report, this dataset is referred to as "Optical".

2.2.3 PEN-BASED RECOGNITION OF HANDWRITTEN DIGITS

This is a dataset of size 10992(Alpaydin & Alimoglu, 1996). The data were collected using a WACOM tablet, capturing (x,y) coordinates of the pen. Each digit trajectory is normalized for translation and scale, and resampled to a fixed number of points to create feature vectors of equal length. For the rest of the report, this dataset is referred to as "Pen".

2.2.4 MNIST DATABASE OF HANDWRITTEN DIGITS

This is one of the most commonly used database for assessing supervised learning models (Deng, 2012). The dataset contains centered 28×28 grayscale images of handwritten digits. There are 70000 samples in this dataset. For the rest of the report, this dataset is referred to as "MNIST".

These four datasets vary in size, input type, and preprocessing, providing a diverse set of challenges for handwritten digit recognition.
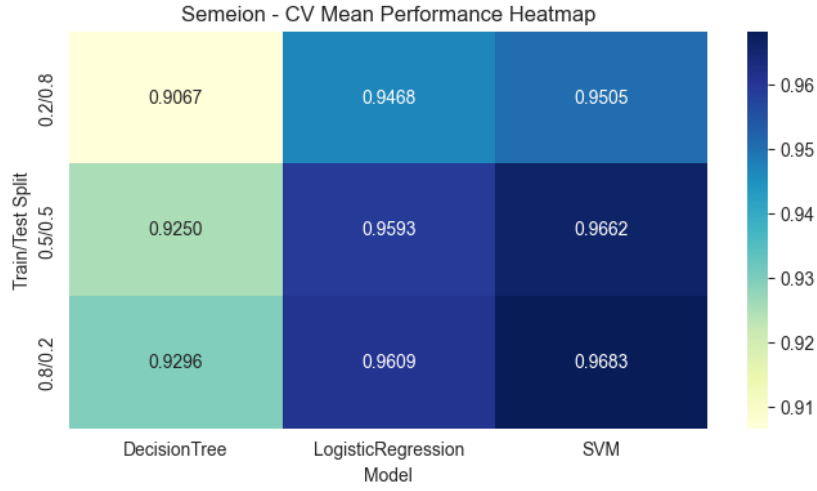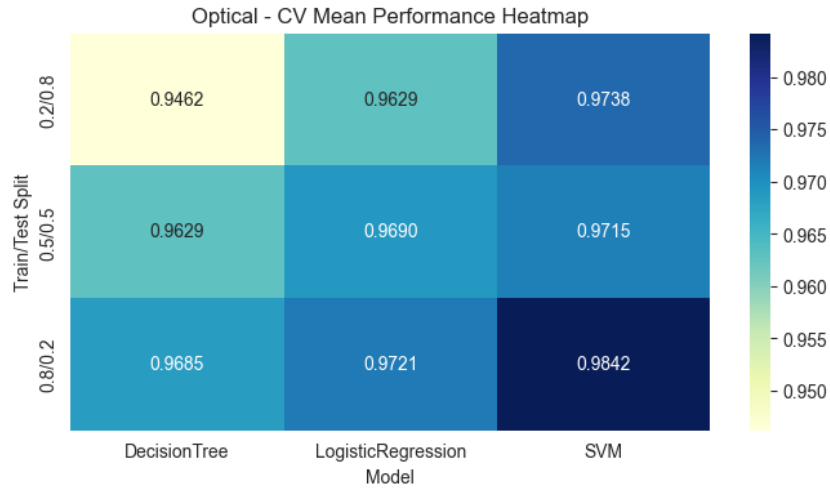
Figure 1: CV mean performance heatmap for Semeion

Figure 2: CV mean performance heatmap for Optical

3 EXPERIMENTAL RESULTS

3.1 HYPERPARAMETERS

Tables 1, 2, 3 show the cross-validation mean accuracy per model / dataset / split. We also included the heatmaps (Figures 1, 2, 3, 4) for the CV mean performances for each dataset.

3.2 TEST ACCURACY

Tables 4, 5, 6 show the test accuracy and f1-score of each classifier on three different datasets. Bold font means that it is the best in the same column.
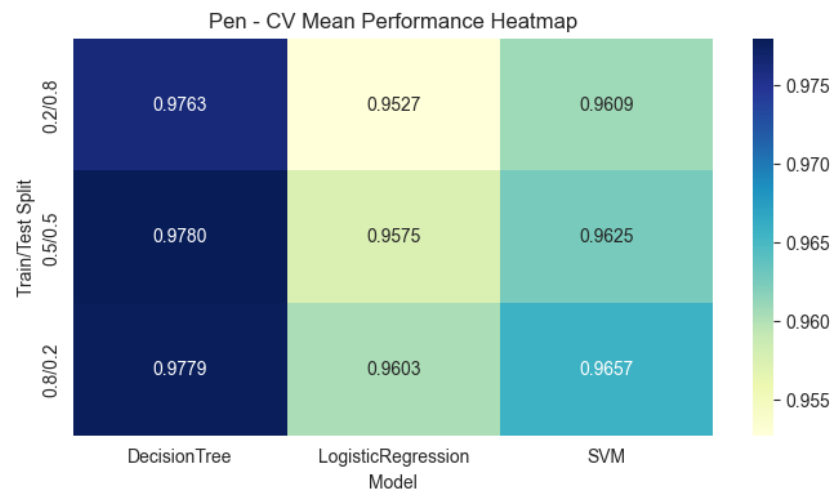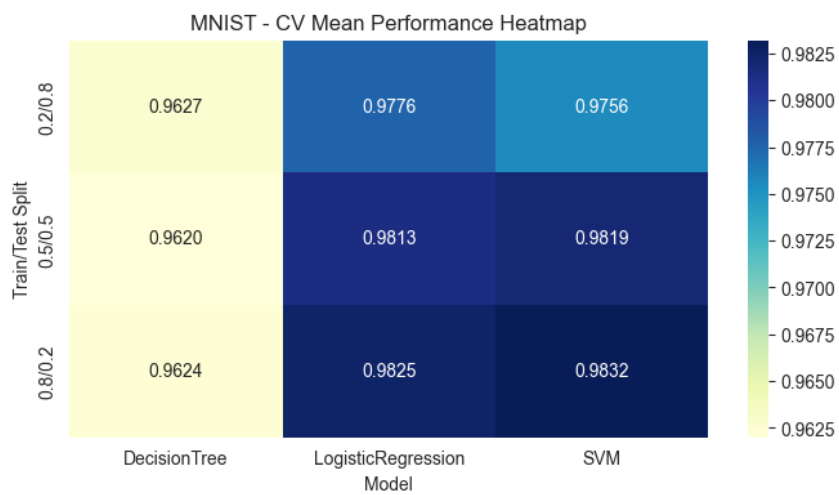
Figure 3: CV mean performance heatmap for Pen

Figure 4: CV mean performance heatmap for MNIST

Table 1: CV mean performance for decision tree

Dataset	Split	C	CV
MNIST	0.2/0.8	-	0.9627 ± 0.0038
MNIST	0.5/0.5	-	0.9620 ± 0.0022
MNIST	0.8/0.2	-	0.9624 ± 0.0011
Pen	0.2/0.8	-	0.9763 ± 0.0043
Pen	0.5/0.5	-	0.9780 ± 0.0022
Pen	0.8/0.2	-	0.9779 ± 0.0012
Semeion	0.2/0.8	-	0.9067 ± 0.0239
Semeion	0.5/0.5	-	0.9250 ± 0.0192
Semeion	0.8/0.2	-	0.9296 ± 0.0099
Optical	0.2/0.8	-	0.9462 ± 0.0133
Optical	0.5/0.5	-	0.9629 ± 0.0111
Optical	0.8/0.2	-	0.9685 ± 0.0082

Table 2: CV mean performance for logistic regression

Dataset	Split	C	CV
MNIST	0.2/0.8	0.01	0.9776 ± 0.0015
MNIST	0.5/0.5	0.01	0.9813 ± 0.0006
MNIST	0.8/0.2	0.01	0.9825 ± 0.0006
Pen	0.2/0.8	1.00	0.9527 ± 0.0027
Pen	0.5/0.5	10.00	0.9575 ± 0.0012
Pen	0.8/0.2	10.00	0.9603 ± 0.0013
Semeion	0.2/0.8	10.00	0.9468 ± 0.0095
Semeion	0.5/0.5	0.10	0.9593 ± 0.0037
Semeion	0.8/0.2	0.10	0.9609 ± 0.0046
Optical	0.2/0.8	10.00	0.9629 ± 0.0056
Optical	0.5/0.5	10.00	0.9690 ± 0.0035
Optical	0.8/0.2	0.10	0.9721 ± 0.0026

3.3 OVERFITTING ANALYSIS

We calculated the overfitting by dataset and model. Specifically, the training and validation accuracy across all splits and trials for each dataset and model is calculated, and the extent of overfitting is calculated by Train - Validation Accuracy.

3.4 DATASET DIFFICULTY

The average test accuracy is also calculated in order to compare model performances across datasets (See Figures 6 and 7).

4 DISCUSSION

4.1 HYPERPARAMETER EFFECTS

Based on the CV mean performance heatmaps, SVM generally achieves the highest CV mean across datasets, except for the Pen dataset, where the decision tree performs best. This may reflect the nature of the feature representation in Pen, which benefits from the tree's partitioning approach. Additionally, we observe slightly higher CV mean scores for the 0.8/0.2 split compared to other splits, likely because a larger training set allows the models to learn more effectively, resulting in improved cross-validation performance.
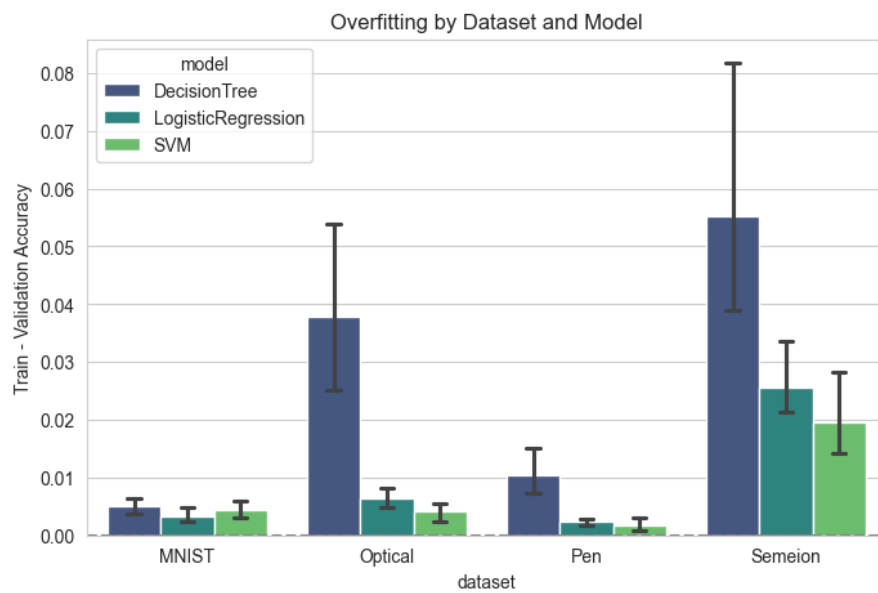
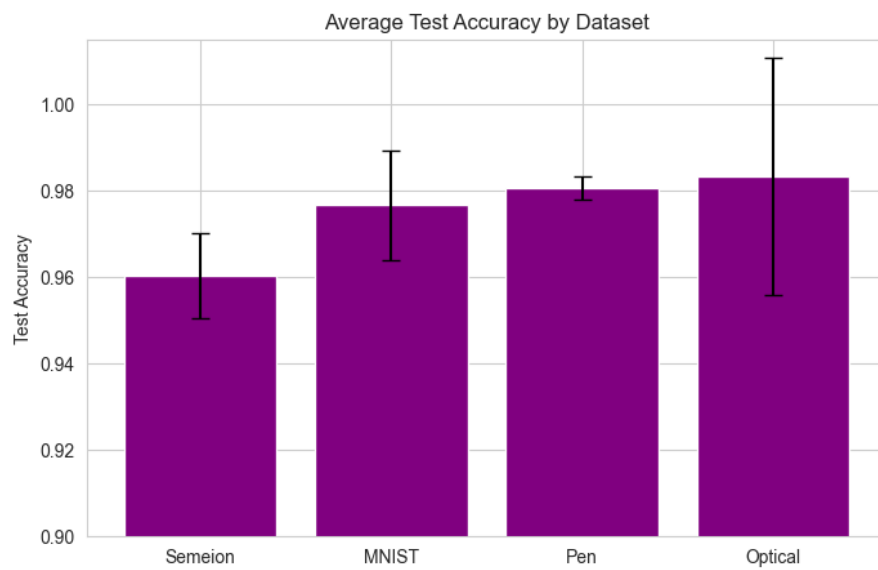Figure 5: Overfitting analysis across dataset and model

Figure 6: Average test accuracy by dataset

Table 3: CV mean performance for SVM

Dataset	Split	C	CV
MNIST	0.2/0.8	0.01	0.9756 ± 0.0018
MNIST	0.5/0.5	0.01	0.9819 ± 0.0006
MNIST	0.8/0.2	0.01	0.9832 ± 0.0006
Pen	0.2/0.8	1.00	0.9609 ± 0.0026
Pen	0.5/0.5	1.00	0.9625 ± 0.0013
Pen	0.8/0.2	1.00	0.9657 ± 0.0017
Semeion	0.2/0.8	0.10	0.9505 ± 0.0100
Semeion	0.5/0.5	0.01	0.9662 ± 0.0040
Semeion	0.8/0.2	0.01	0.9683 ± 0.0037
Optical	0.2/0.8	0.10	0.9738 ± 0.0066
Optical	0.5/0.5	1.00	0.9715 ± 0.0052
Optical	0.8/0.2	0.01	0.9842 ± 0.0036

Table 4: Results for Split 0.2/0.8

Model	MNIST F1	MNIST Test	Pen F1	Pen Test	Semeion F1	Semeion Test	Optical F1	Optical Test
DecisionTree	0.813	0.964	0.880	0.976	0.543	0.911	0.800	0.958
LogisticRegression	0.912	0.982	0.898	0.979	0.827	0.968	**0.958**	**0.991**
SVM	**0.916**	**0.983**	**0.905**	**0.981**	**0.836**	**0.970**	0.941	0.988

Table 5: Results for Split 0.5/0.5

Model	MNIST F1	MNIST Test	Pen F1	Pen Test	Semeion F1	Semeion Test	Optical F1	Optical Test
DecisionTree	0.814	0.963	0.902	0.981	0.582	0.921	0.846	0.969
LogisticRegression	0.919	**0.984**	0.909	0.982	**0.892**	**0.980**	0.961	0.992
SVM	**0.921**	**0.984**	**0.914**	**0.983**	0.888	0.979	**0.962**	**0.993**

Table 6: Results for Split 0.8/0.2

Model	MNIST F1	MNIST Test	Pen F1	Pen Test	Semeion F1	Semeion Test	Optical F1	Optical Test
DecisionTree	0.810	0.963	0.891	0.979	0.714	0.947	0.877	0.977
LogisticRegression	0.922	**0.984**	0.909	0.982	0.912	0.983	0.951	0.991
SVM	**0.924**	**0.984**	**0.914**	**0.983**	**0.921**	**0.984**	**0.961**	**0.993**

4.2 TEST ACCURACY ACROSS MODELS

We do not provide a ranking here since it is not statistically meaningful, and all test accuracy exceeded 0.9. Overall, these three classifiers perform similarly on accuracy metrics. Linear SVM seemed to achieve the highest accuracy, particularly in MNIST and Pen, whereas decision tree performed relatively worse on all datasets, especially when comparing the F1 scores.

4.3 OVERFITTING

Overall, decision tree shows a pronounced gap between training and validation accuracy, especially in the Optical, Pen, and Semeion datasets. This indicates its strong overfitting, especially on small datasets. Logistic regression and SVM demonstrate smaller gaps, suggesting better generalization.

4.4 DATASET DIFFICULTY

While the average test accuracies were similar across datasets, subtle differences may reflect variations in dataset characteristics. For instance, Semeion exhibits higher writer variability and less standardized preprocessing, whereas MNIST and Optical contain more uniform, centered images. These factors could make some datasets slightly more challenging, even if overall accuracy appears comparable.
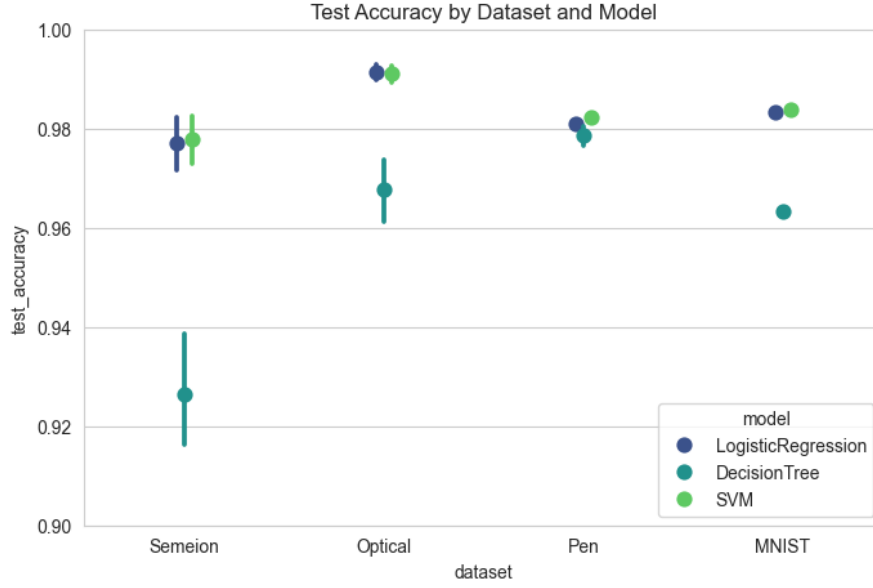
Figure 7: Average test accuracy by dataset and model

Table 7: Training and validation accuracy for split 0.2/0.8

Model	MNIST_train	MNIST_val	Optical_train	Optical_val	Pen_train	Pen_val	Semeion_train	Semeion_val
DecisionTree	0.969	0.963	1.000	0.946	0.991	0.976	0.988	0.907
LogisticRegression	0.987	0.982	1.000	0.993	0.983	0.980	1.000	0.966
SVM	0.988	0.982	0.996	0.992	0.984	0.981	0.996	0.968

5 CONCLUSION

Taken together, our experiments demonstrate that SVM generally achieves the highest accuracy across datasets, with logistic regression performing similarly, while decision trees show greater overfitting, particularly on smaller or more variable datasets such as Semeion and Pen. Differences in dataset characteristics (such as input representation, preprocessing, and writer variability) modulate model performance, though overall accuracy remains high for all classifiers.

Cross-validation heatmaps reveal that hyperparameter selection can influence performance, with larger training sets (e.g., 0.8/0.2 split) producing slightly higher CV mean accuracy due to improved learning. While average test accuracy was similar across datasets, subtle distinctions in overfitting and variance highlight the importance of dataset structure in practical model evaluation. Overall, this study provides a comprehensive comparison of supervised classifiers for handwritten digit recognition, illustrating how algorithm choice and dataset properties jointly determine performance.

A APPENDIX

A.1 TRAINING AND VALIDATION ACCURACY

Tables 7, 8, 9 show the training and validation accuracy for each experiment.

A.2 SOURCE CODE

The code for our experiments is available at `https://github.com/yagregx/COGS118A_final`.

Table 8: Training and validation accuracy for split 0.5/0.5

Model	MNIST_train	MNIST_val	Optical_train	Optical_val	Pen_train	Pen_val	Semeion_train	Semeion_val
DecisionTree	0.967	0.962	0.997	0.963	0.987	0.978	0.970	0.925
LogisticRegression	0.986	0.984	0.999	0.991	0.983	0.980	1.000	0.979
SVM	0.988	0.984	0.997	0.991	0.983	0.982	0.992	0.978

Table 9: Training and validation accuracy for split 0.8/0.2

Model	MNIST_train	MNIST_val	Optical_train	Optical_val	Pen_train	Pen_val	Semeion_train	Semeion_val
DecisionTree	0.966	0.962	0.994	0.968	0.985	0.978	0.969	0.930
LogisticRegression	0.986	0.984	0.997	0.992	0.984	0.982	0.999	0.977
SVM	0.987	0.984	0.996	0.994	0.984	0.983	0.991	0.975

A.3 BONUS POINTS

In addition to the minimum three datasets required, I included the MNIST dataset, providing a significantly larger and more complex benchmark for handwritten digit classification. This allows for a more comprehensive comparison across datasets of varying size and difficulty.

REFERENCES

Semeion Handwritten Digit. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5SC8V.

E. Alpaydin and Fevzi. Alimoglu. Pen-Based Recognition of Handwritten Digits. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5MG6K.

Ethem Alpaydin and Cune Kaynak. Optical recognition of handwritten digits [dataset]. UCI Machine Learning Repository, 1998.

Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012. 2211477.