STA5076Z Supervised Learning Assignment 3

Adam Mosam (MSMADA002)

11 November 2023

Contents

1 Introduction

The burgeoning field of data science continuously seeks innovative methodologies to extract meaningful insights from various forms of data. This assignment delves into the comparative analysis of two distinct datasets utilizing advanced machine learning techniques to demonstrate the robustness and adaptability of these methods across different data types and classification challenges.

1.1 Aim

The primary objective of this assignment is to apply different dimensional reduction methods to two types of data to evaluate the effectiveness of these methods in enhancing classification accuracy. The datasets in question are diverse, consisting of the notMNIST database of letters and the Wisconsin Diagnostic Breast Cancer dataset (WDBC). The notMNIST dataset, comprising PNG images of handwritten letters "C" and "G", poses the challenge of image recognition and classification. In contrast, the WDBC dataset, stored as a CSV file, encapsulates clinical measurements reflective of cancer diagnosis. The exploration of these datasets involves the k-NN algorithm for the initial classification, followed by the PCA,Autoencoder and SOM techniques to reduce dimensionality and to potentially unveil latent patterns beneficial for enhancing the predictive models.

1.2 Dataset

The notMNIST dataset is a collection of grayscale PNG images of letters, specifically "C" and "G," each encapsulated in a 28x28 pixel grid. The essence of each image is captured in the pixel values, which serve as features for classification. The distinctiveness of each letter emerges from the pixel's intensity, where each image is a flat array of 784 pixel values ranging from 0 to 255 (standardidized to 0-1). There includes 1872 images for each letter.

Figure 1: Image of notMNIST C and G letters represented on a 28x28 pixel grid.

The WDBC dataset is a structured collection of biophysical markers in CSV format designed to assist in the diagnosis of breast cancer. There are 569 records, with each record in the dataset representing a case with a unique ID and diagnosis result, accompanied by a suite of features including radius, texture,

perimeter, area, and various attributes characterizing the cellular nuclei present in the images. The dataset's comprehensive nature, with features ranging from mean values to worst-case (maximum) values of the cell nuclei characteristics, provides a rich foundation for sophisticated analyses.

The analysis of these datasets not only underscores the utility of machine learning in practical applications but also challenges the adaptability of algorithms across varying data types, from the pixel values of images to the clinical attributes in cancer diagnosis. The judicious application of dimensional reduction offers a pathway to distill these complex datasets into a more manageable form, potentially enhancing the predictive capabilities of the k-NN classification algorithm.

2 Methodology

2.1 Initial k-NN Model Development

The methodology begins with the preprocessing of two distinct datasets: the noMnist dataset containing images of handwritten letters classified as 'C' or 'G', and the Wisconsin Breast Cancer Diagnosis (WBDC) dataset, labeled as 'malignant' or 'benign'. Preprocessing will involve normalizing the pixel intensity values for the noMnist dataset and standardizing the biophysical marker features in the WBDC dataset to ensure that the scale of the measurements does not bias the k-NN algorithm.

After preprocessing, both datasets will be divided into training and test sets. The k-NN classifier will then be applied to the training set, and its performance will be validated on the test set. To optimize the k-NN model, a range of values for 'K' (the number of nearest neighbors) will be explored to determine the value that yields the highest accuracy in classifying the test data. This process will establish baseline accuracies for each dataset without dimensionality reduction.

2.2 Dimensionality Reduction and k-NN Classification

Following the establishment of a baseline model, we will apply three dimensionality reduction techniques: Principal Component Analysis (PCA), Autoencoders, and Self-Organizing Maps (SOM). These techniques will be used to transform the original datasets into new feature spaces with reduced dimensions, specifically considering 2, 3, 4, 5, and 6 dimensions for each method.

For PCA, we will compute the principal components and retain the top components as per the specified dimensionalities. For Autoencoders, a neural network-based approach will be utilized to learn a compressed representation of the input data. SOM's will be used to map the high-dimensional data into a two-dimensional grid that captures the topological characteristics of the input space.

Once the datasets are transformed by each dimensionality reduction technique, the k-NN classifier will be retrained on these reduced features. The training process will be similar to the initial model development, optimizing for the number of neighbors 'K' that achieves the best accuracy on the validation set. The test accuracy of these dimensionally-reduced k-NN models will be compared against the initial models' accuracies.

The comparison of accuracies will allow us to evaluate the impact of dimensionality reduction on the k-NN classifier's performance. We expect to observe changes in accuracy, which will provide insights into the trade-off between dimensionality reduction and classification performance.

3 k-NN Model Construction on the Full Dataset

The k-NN model development proceeds as follows:

Data Splitting: The datasets are firstly divided into a training set and a test set, using an 80-20 split. The split is carried out to ensure a representative distribution of both classes in each subset.

Hyperparameter Tuning: The k-NN algorithm is sensitive to the choice of the hyperparameter 'K', which dictates the number of nearest neighbors to consider for classification. To fine-tune this parameter, a range of 'K' values (1 to 10) is tested through cross-validation on the training data. The 'K' value that results in the highest cross-validation accuracy is selected for the final model.

Model Training: Employing the chosen 'K', the k-NN model is trained over the entire feature space of the training dataset.

Performance Evaluation: The efficacy of the k-NN model is quantified by its accuracy on the test set. This evaluation metric establishes a baseline performance for the original, high-dimensional dataset, which will be crucial for comparative analysis against models trained on dimensionality-reduced data.

3.1 noMnist

Prior to model development, the PNG images within the noMnist dataset, consisting of 'C' and 'G' letters need to be converted into a structured pixel grid format. The 28x28 pixel grid is unwrapped into a 784-dimensional vector, where each element represents the grayscale intensity of a single pixel.

For both 'C' and 'G' images, a categorical response variable is created and assigned respectively. The datasets for 'C' and 'G' are then amalgamated into a single dataset with labels, ready for the k-NN classifier to process.

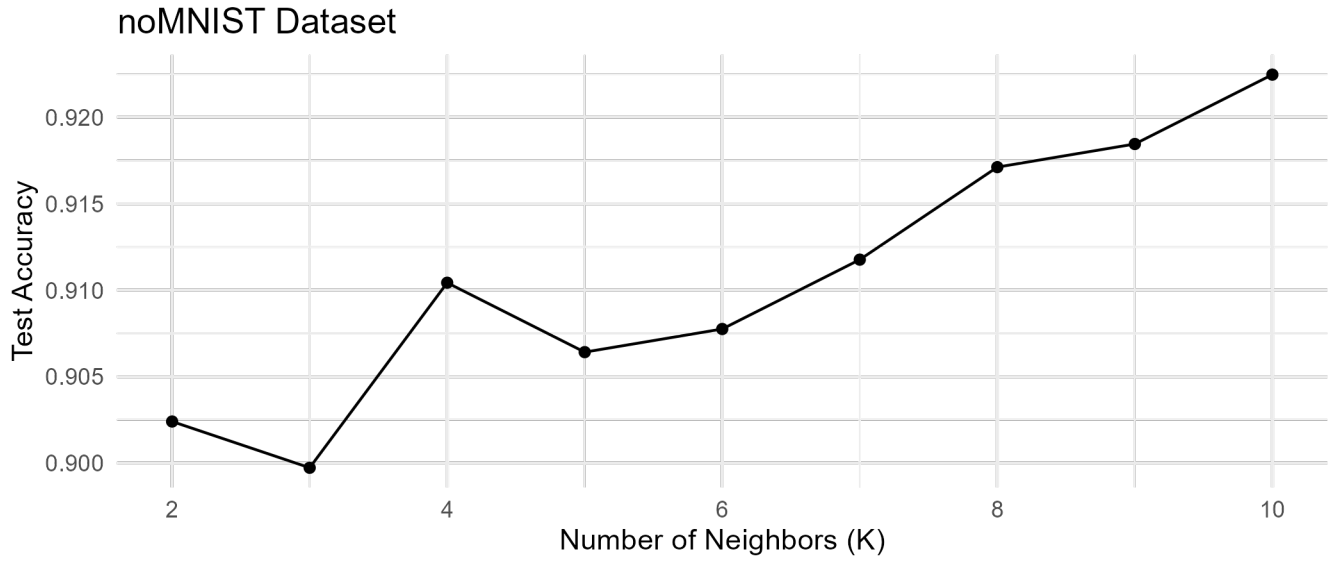

Figure 2: Test accuracy of full noMNIST dataset, using various K values.

As shown the maximum accuracy achieved, 92%, is found with a K value of 10

3.2 WBDC

For the WBDC dataset, the features required scaling first, with a standard deviation of 1 and mean of 0.

From the 3 it may be shown that the best accuracy comes with a K value of 3, resulting in an accuracy score of 99%.

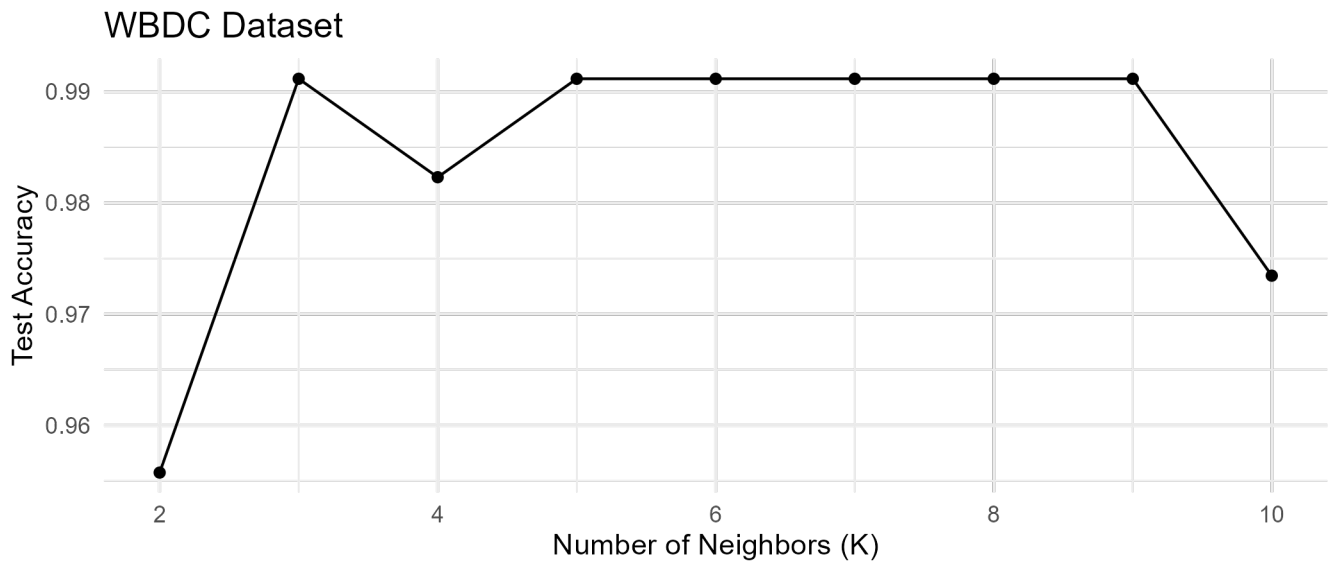

Figure 3: Test accuracy of full WBDC dataset, using various K values.

4 PCA Model

In this section, we aim to ascertain the most effective Principal Component Analysis (PCA) model variant for our dataset. We will explore three distinct kernels: linear, polynomial, and radial basis function (RBF), to perform dimensionality reduction. The dataset will be reduced to six dimensions using each kernelized PCA model. Subsequently, we will employ the k-Nearest Neighbors (k-NN) classifier to evaluate the test accuracy for each reduced dataset, following the methodology outlined previously. This comparative analysis will enable us to identify the kernel that maximizes classification performance within the context of our study.

4.1 Radial PCA

For the radial or Gaussian kernal PCA, a grid search of the σ parameter over the range 0.1, 1 and 10 will be performed. Both datasets are standardized, so this range of values is appropriate. The k-NN test accuracies for both datasets are shown in Figure 4.

4.2 Polynomial PCA

For the polynomial kernal PCA, a grid search of the polynomial degree parameter over the range 2-6 will be performed using the KPCA function. A scale value of 1 and offset value of 0 will be assumed within the KPCA function.

The polynomial kernel applied to both the noMNIST and WBDS datasets are shown in Figure 5:

4.3 Linear PCA

For the linear PCA analysis, PRCOMP will be used. The k-NN test accuracies for both datasets are shown in Table 1.

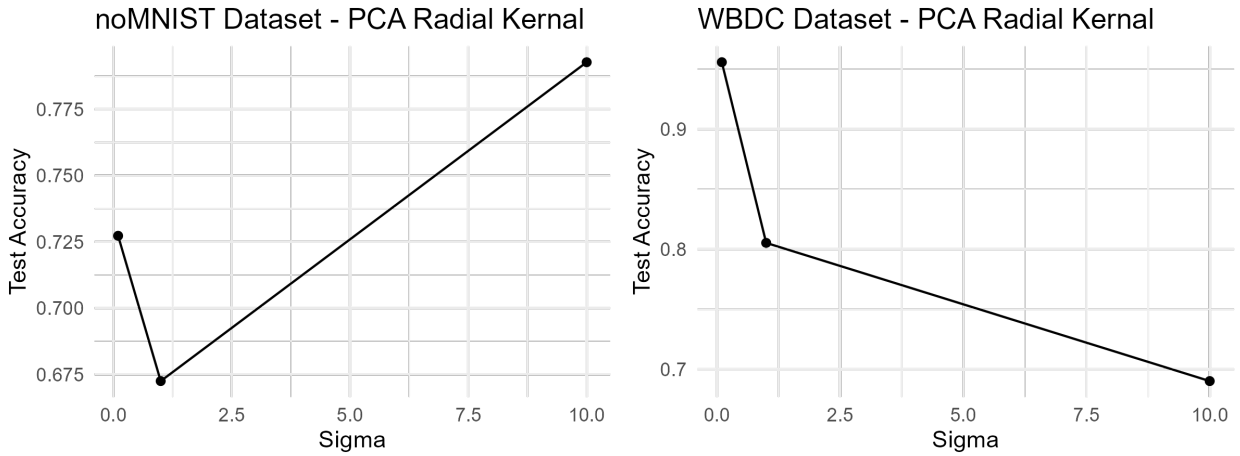

Figure 4: Test accuracy of reduced MNIST (left) and WBDC (right) datasets, using radial kernel with various sigma values.

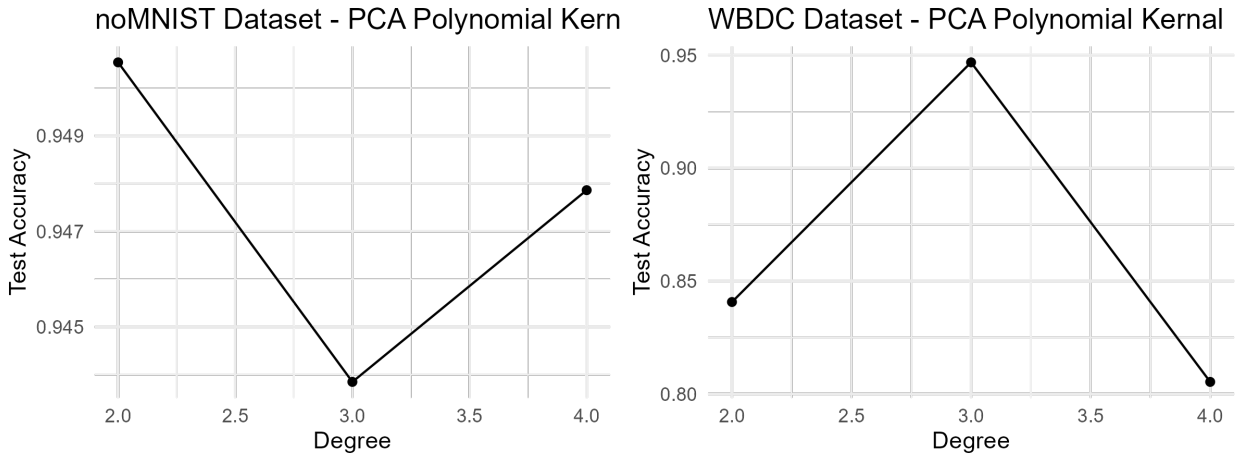

Figure 5: Test accuracy of reduced MNIST (left) and WBDC (right) datasets, using polynomial kernel with various degrees.

Table 1: Test accuracy of reduced MNIST and WBDC datasets, using lienar PCA.

noMNIST	WBDC
96.3	95.6

4.4 Model Selection

From the sections above, it may be observed that

- WBDC - Linear PCA and Polynomial (3rd degree) perform the best
- noMNIST - Linear and Polynomial (2nd degree) perform the best

To select the optimal model we will additionally look at the two most significant principle components in order to undertsand how the data is partitioned. For the WBDC data set get the following

From Figure 6 it may be viewed that the linear shows a more distinct seperation between the first two principle components. In contrast, from Figure 7, the polynomial PCA appears to have a better fit with respect to the distinction between the principle components.

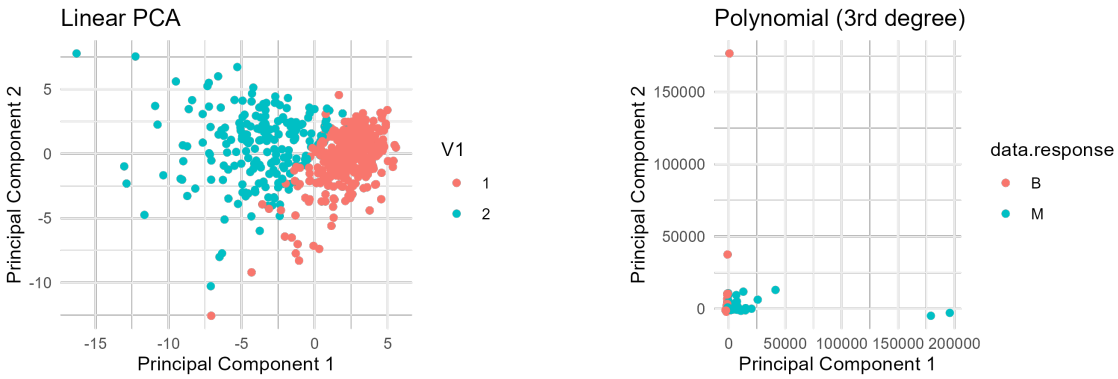

Figure 6: First two principle components for WBDC using linear and polynomial (3rd degree) PCA.

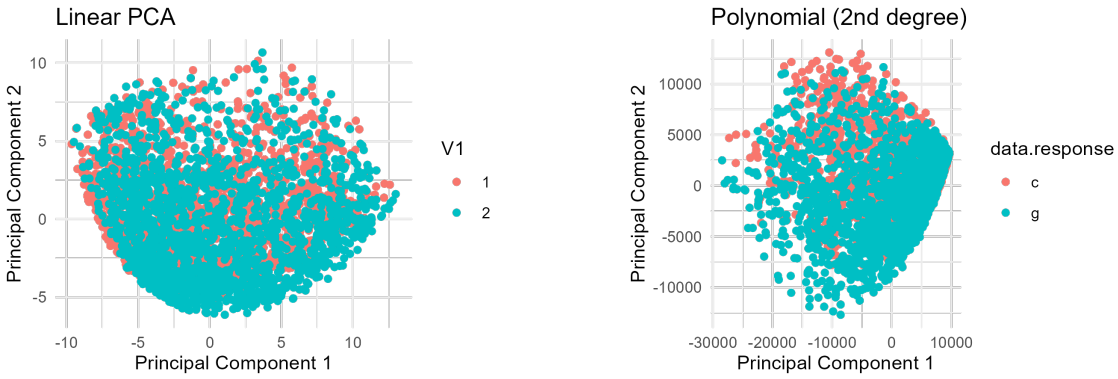

Figure 7: First two principle components for noMNIST using linear and polynomial (2nd degree) PCA.

5 Autoencoders Model

This section introduces the application of autoencoders to the noMNIST and WBDC datasets for dimensionality reduction. The architecture features two hidden layers: a latent layer and an additional layer to enhance data representation. We employ autoencoders with 100 epochs and a batch size of 256, using ReLU activation in the encoder for efficient gradient propagation and Sigmoid in the decoder to match the normalized data range. The subsequent analysis will focus on hyperparameter tuning, specifically adjusting the hidden layer size and L2 regularization to optimize the latent space representation for k-NN classification performance.

5.1 Hidden Neuron Layer Size

The impact of the hidden layer's neuron count will be assessed with configurations of (32, 64, 128), and (128, 256, 384) neurons for the WBDC and noMNIST datasets, respectively. Accuracy metrics for each configuration are detailed below for both the noMNIST and WBDC datasets.

From Figure 8, the following optimal hyperparameters are observed:

- WBDC - a hidden layer size of 64 neurons.
- noMNIST - a hidden layer size of 256 neurons.

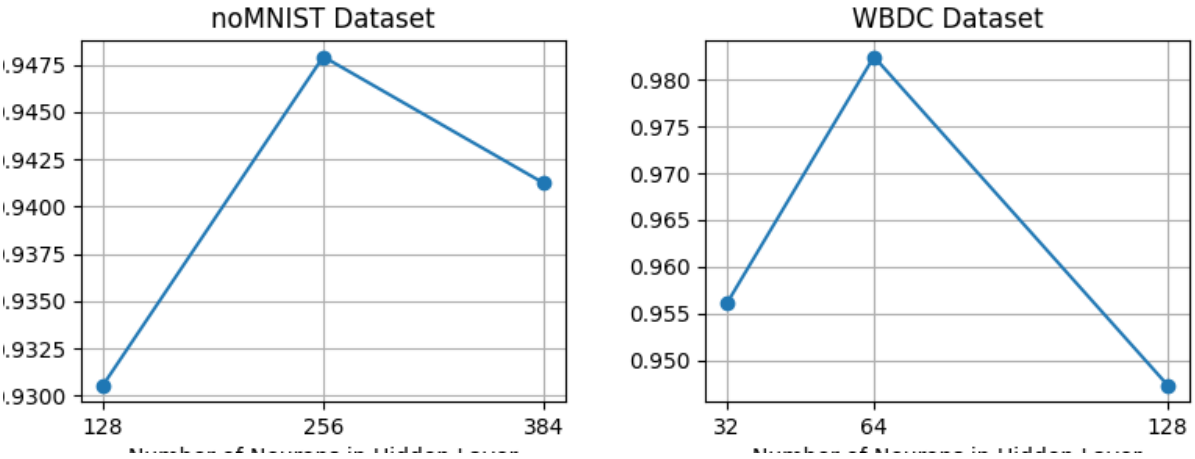

Figure 8: Autoencoder hidden layer neuron size optimization

5.2 L2 Regularization

We will evaluate the effect of L2 regularization on model performance by testing lambda values of (0.001, 0.01, 0.1) and (0.00001, 0001, 0.001), for the WBDC and noMNIST datasets, repsectively. The accuracy results for both the noMNIST and WBDC datasets, across these regularization parameters, are presented in Figure 9.

From Figure 9, the following optimal hyperparameters are observed:

- WBDC - λ value of 0.01.
- noMNIST - λ value of 0.00001.

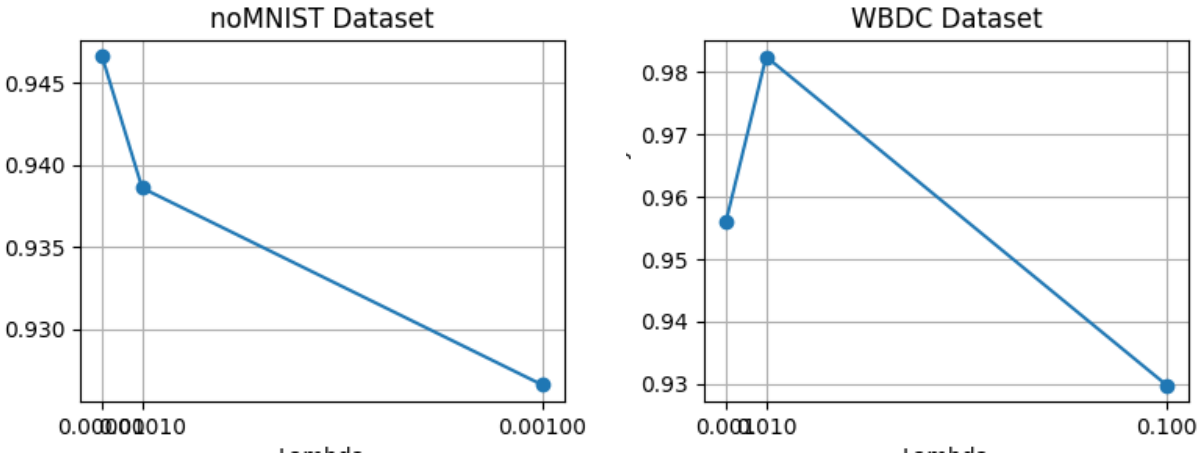

Figure 9: Autoencoder L2 regularization paramater, lambda, optimization.

6 SOM Model

In this analysis, we integrate Self-Organizing Maps (SOMs) to preprocess the data prior to classification with the k-NN algorithm. SOMs offer a distinctive approach to high-dimensional data comprehension, arranging it within a two-dimensional grid framework. This technique contrasts with PCA and autoencoders, which reduce dimensionality quantitatively, as SOMs prioritize the relational layout of data points.

For consistency with previous methodologies, we maintain a 6×6 grid size throughout the SOM analysis. Our primary objective is to optimize the SOM's learning rate to enhance kNN classification accuracy. In the next section, we will investigate the effects of varying the grid dimensions.

The effect of the learning rate on the k-NN accuracy may be viewed in Figure 10. From the figure, the following optimal learning rates are noted

- WBDC - learning rate of 0.0001.
- noMNIST - learning rate of 0.0001.

7 Final Model Performance

In the final results section, we'll present the outcomes of using different dimensionality reduction techniques, specifically PCA, autoencoders, and SOMs, and their effects on k-NN classification accuracy. With the best hyperparameters in hand for each method, we'll explore how different sizes of the reduced feature space, ranging from 2 to 6 dimensions, impact the performance. For SOMs, we'll test grid sizes from 2x2 to 6×6 to see which size offers the best results with kNN.

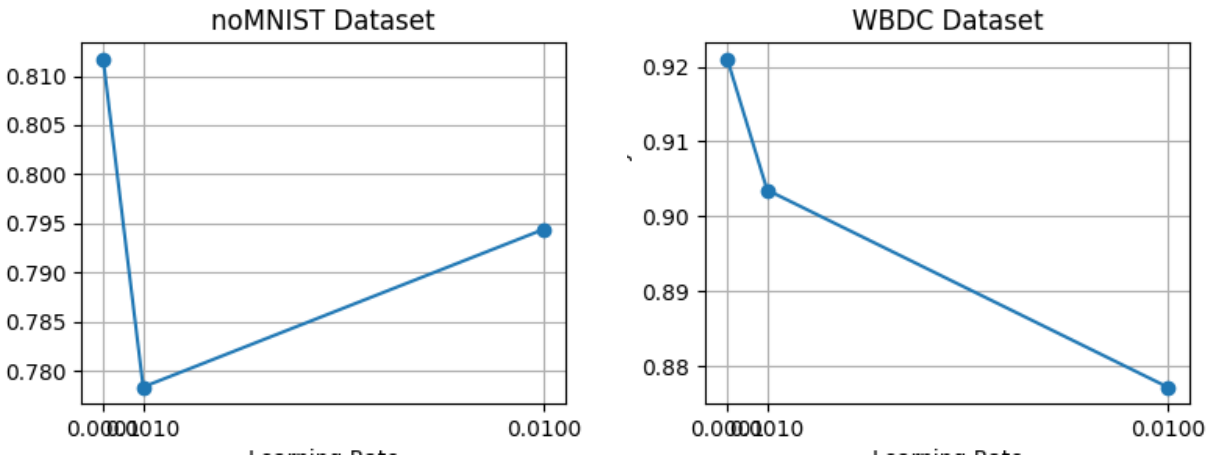

Figure 10: Slef organizing maps, learning rate optimization.

Through this analysis, we'll determine which dimensionality reduction approach and settings give us the highest classification accuracy.

The size of the reduced feature on the test accuracy will be investigated in this section for the PCA, Autoencoder and SOM models. The results are shown in Figure 11.

The best performing models from each dimension reduction technique is displayed in Table 2 and 3. In addition, the accuracy, determined from the full dataset is also provided.

Table 2: PCA, AE and SOM, optimal model accuracies for nonMNIST dataset

Result	Full	PCA	AE	SOM
Accuracy	92.246	95.053	94.126	81.175
Feature Size		6.000	5.000	5.000

Table 3: PCA, AE and SOM, optimal model accuracies for WBDC dataset

Result	Full	PCA	AE	SOM
Accuracy	99.1	95.575	98.246	95.614
Feature Size		6.000	6.000	5.000

The results from the dimensionality reduction techniques applied to the noMNIST and WBDC datasets, when used in conjunction with the k-NN classifier, show interesting trends. For the noMNIST dataset, using PCA and autoencoders (AE) sometimes led to better accuracy compared to the full dataset. This could mean

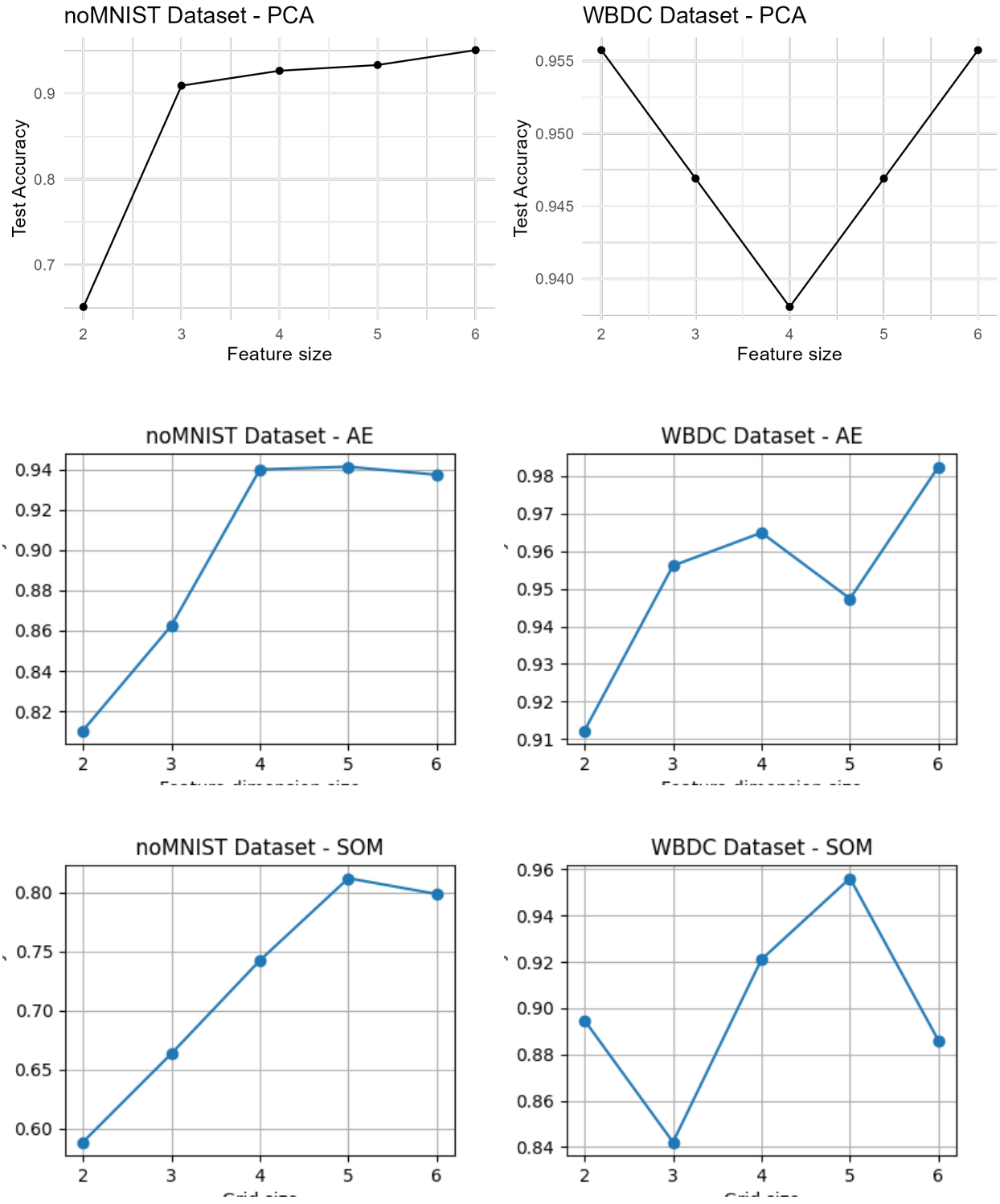

Figure 11: PCA, AE and SOM Models, simulated across a range of feature space sizes.

that these methods effectively highlighted the most important features for distinguishing between different characters, possibly by removing noise from the data.

In contrast, the Self-Organizing Maps (SOM) method and some instances of PCA and AE resulted in a slight drop in accuracy for the WBDC dataset. This decrease in accuracy is likely due to the loss of information.

Despite the variations, it's noteworthy that all accuracies remained relatively high, with most results exceeding 90%. This high level of accuracy across all models indicates that the key patterns within the data are robust enough to be captured even with a reduced set of features.

8 Conclusions

The application of dimensionality reduction techniques to the noMNIST and WBDC datasets reveals that while PCA and autoencoders can enhance kNN classification accuracy by isolating the most impactful features, especially in image-based data, caution is necessary to avoid losing vital information in more complex, clinical datasets. Despite the occasional reduction in accuracy, performance remains commendably high, suggesting that these techniques are capable of preserving the essential characteristics needed for effective classification.