CSC730: Report for Assignment 1

South Dakota School of Mines and Technology

Jacob James, David Mathews, Kris Jensen

January 17, 2024

1 Introduction

The first assignment of this course tasked us with analyzing a set of skewed data from the MNIST dataset. This skewed dataset contained eight classes from a total set of ten classes. Also, each class was not represented with equivalent frequency. The dataset contains 12244 records of data. Each record contains a 784-element list that represents a 28x28 image of a handwriting sample.

This report will detail the steps taken by our team to generate an anomaly score for each image and compare our results for accuracy.

We used two primary toolsets to perform the analysis. One toolset was python executed on VS code and the other was python executed on Google Colab.

2 Methodology

The methodology that provided the best results for this assignment counted the the pixels that exceeded an arbitrarily chosen threshold value of 128. The maximum gray scale intensity being 255, this value is the midpoint of the intensity range. This count was subtracted from the mean value of counts from all images in the dataset and finally squared. This choice of anomaly score generation proved to be reasonable. The results are presented later in the paper.

Another method that was attempted was to uses an FFT to generate an anomaly score. This score was generated by taking the FFT of each image and then computing the power spectral density (PSD). After the PSD is computed, the values are scaled by the mean and standard deviation. After scaling, peaks are identified and the top four selected. The four peaks are used to generate the anomaly score, by using their locations to compute an average location then finally the compute the distance from the origin. This distance from the origin is the anomaly score.

The remainder of this section will provide graphical details of the algorithm in action.

An example of two handwriting samples is shown as images in figure 1 and figure 2. These images were produced using the imshow function from the matplotlib python library.

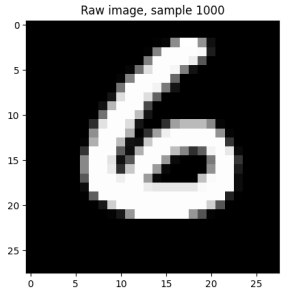

Figure 1: Handwriting Sample 1000

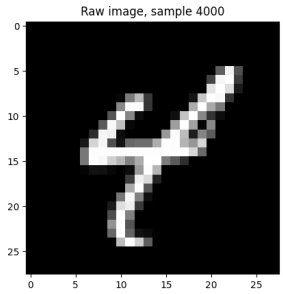

Figure 2: Handwriting Sample 4000

The first step is producing a thresholded image of each sample as shown in figure 3 and figure 4 . Experiment 1000 resulted in 615 black pixels and 169 white pixels. Experiment 4000 resulted in 705 black pixels and 79 white pixels.

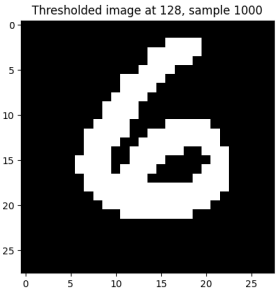

Figure 3: Threshold set to 128 of handwriting sample 1000

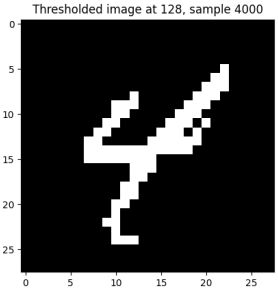

Figure 4: Threshold set to 128 of handwriting sample 4000

Then a mean image is produced, figure 5, to generate the mean count. The count of white pixels for the thresholded mean image is 38, figure 6.

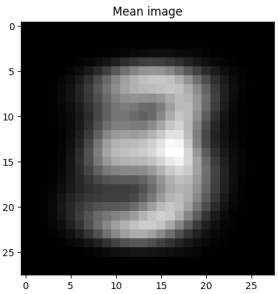

Figure 5: Mean Image

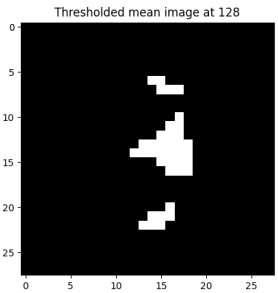

Figure 6: Threshold set to 128 of mean image

The next step requires subtracting the experiment data from the mean data. The graphical results are shown in figure 7 and figure 8.

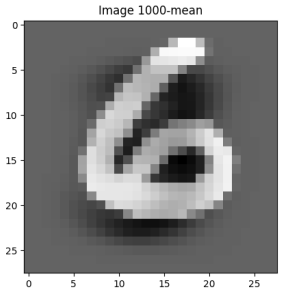

Figure 7: Subtraction of experiment image and mean image for experiment 1000

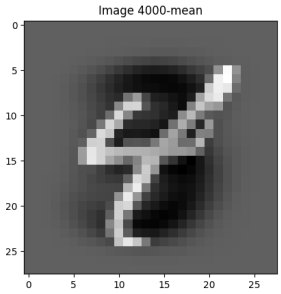

Figure 8: Subtraction of experiment image and mean image for experiment 4000

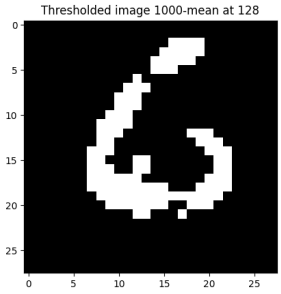

Figure 9: Square of the subtraction of experiment image and mean image for experiment 1000

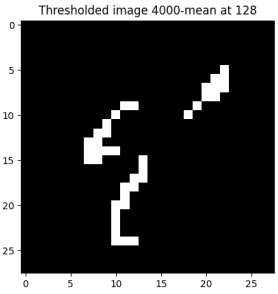

Figure 10: Square of the subtraction of experiment image and mean image for experiment 4000

Now that the subtraction of the experiment image and the mean image has occurred, the resultant value is squared. Refer to figure 9 and Figure 10 for the graphical results.

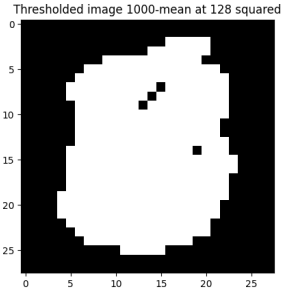

Figure 11: Square of the subtraction of experiment image and mean image for experiment 1000

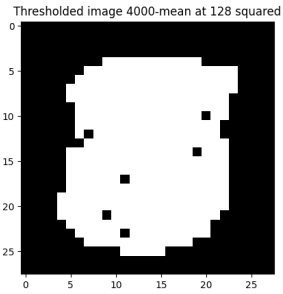

Figure 12: Square of the subtraction of experiment image and mean image for experiment 4000

The final step is to sum the values of the squared subtraction of the experiment image and the mean image. The results are shown in figure 11 and figure 12.

The images of the power spectral density plots by class will listed in the following ten figures.

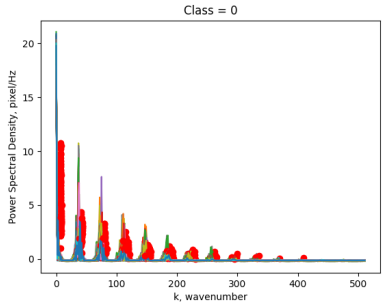

Figure 13: PSD of Class 0 handwriting data

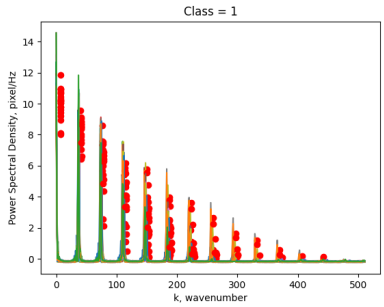

Figure 14: PSD of Class 1 handwriting data

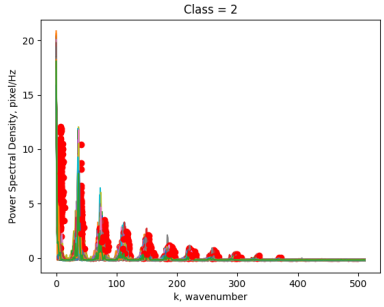

Figure 15: PSD of Class 2 handwriting data

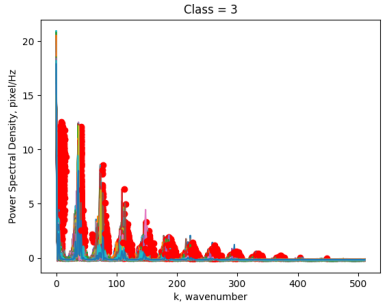

Figure 16: PSD of Class 3 handwriting data

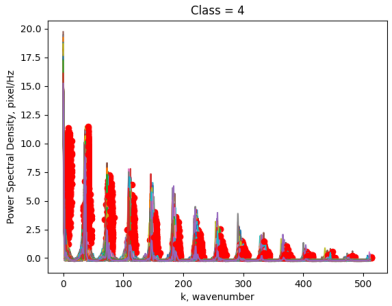

Figure 17: PSD of Class 4 handwriting data

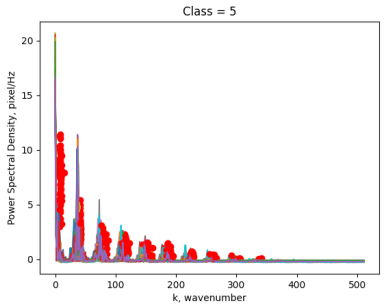

Figure 18: PSD of Class 5 handwriting data

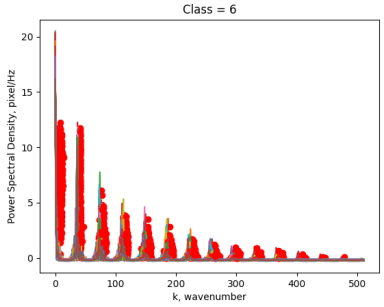

Figure 19: PSD of Class 6 handwriting data

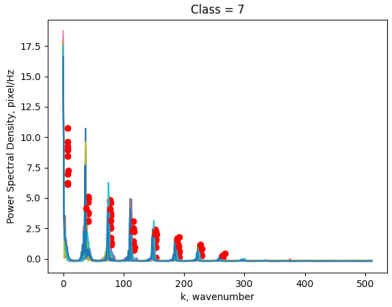

Figure 20: PSD of Class 7 handwriting data

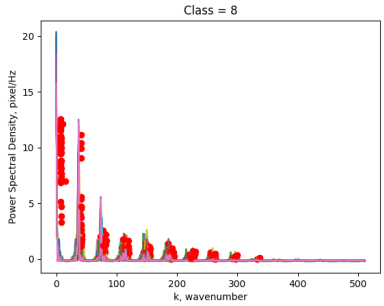

Figure 21: PSD of Class 8 handwriting data

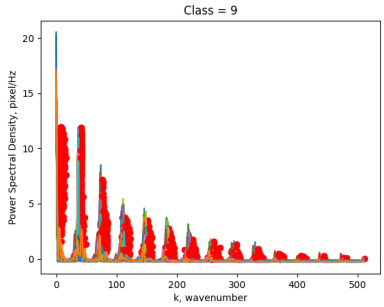

Figure 22: PSD of Class 9 handwriting data

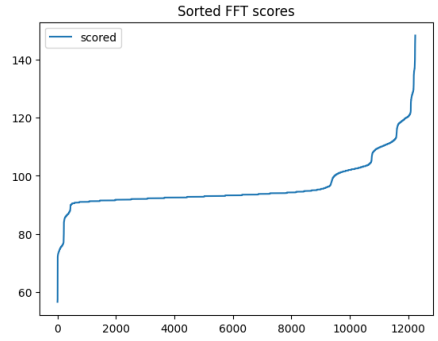

Figure 23: Sorted FFT scores

3 Results

The probabilities from the data set are in the following figure.

Table of results

Method	Accuracy
Random	33.3%
Threshold Sum	39.2%
Threshold Sum and Square	32.7%
Single-Axis Guassian	38.5%
FFT distance	47.5%

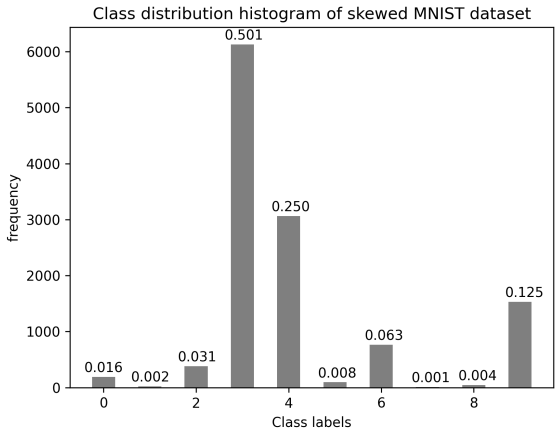

Figure 24: Histogram of True Classes

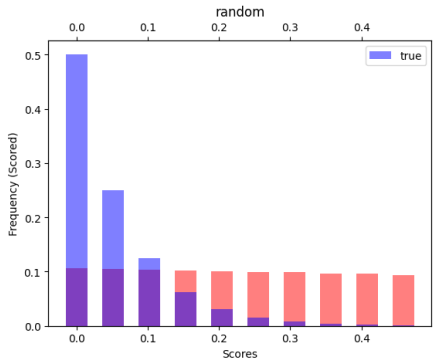

Figure 25: Sorted histogram of random method scores

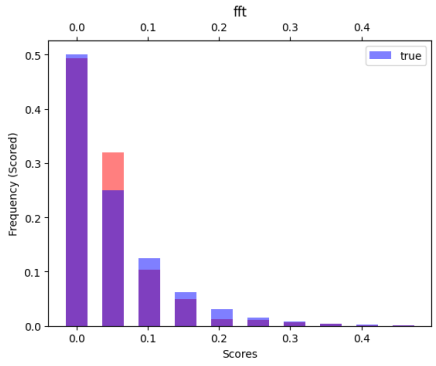

Figure 26: Sorted histogram of FFT method scores

4 Discussion

Rather than reducing the dimensions using previous statistical methods like PCA, our group's hypothesis was that most of the information contained in the image was found in the surface area of given pixel values. The skewed MNIST dataset contained images with only white and black pixel values, thus the surface area was captured using a specified median criterion of a typical [0,255] domain. Our second hypothesis was that each class followed a distribution, and thus a difference from the mean could be used as an anomaly score, thus the mean was subtracted from each count of white pixels contained in each image. Lastly, this value was squared to ensure values further away from the mean would give a larger score regardless of being greater or less than the mean. As discussed in the results section, the accuracy for this method was 32.7 %. To develop a baseline comparison, a simulation was created which assigned random anomalous scores for each sample achieved an anomalous accuracy of 33.3 %, which shows our hypothesis for this method was incorrect.

We then examined another method that utilized the FFT as the primary data analysis tool. This method utlimately calculatesd a score from the peaks that result in each power sprectral density. Our results for this method exceeded the random baseline as well as the other methods we examined.

5 Conclusion

Multiple methods were explored throughout this assignment including: count difference from the mean squared and unsquared, univariate Gaussian distribution likelihood, fft. All three methods performed considerably better than a random guess approach, as discussed in the results section. Additionally, when reconstructing class counts for each method, the fft approach resulted in a 93% simularity with the original data, while resulting in a 48% identification accuracy, which shows significant improvements to the fft anomolous accuracy can be obtained with future research. This is significantly better than the random guess, difference from the mean, and univariate Gaussian which recieved simularity scores of 43%, 78%, and 80% respectively.