CSC730: Report for Assignment 1

South Dakota School of Mines and Technology

Jacob James, David Mathews, Kris Jensen

January 15, 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.

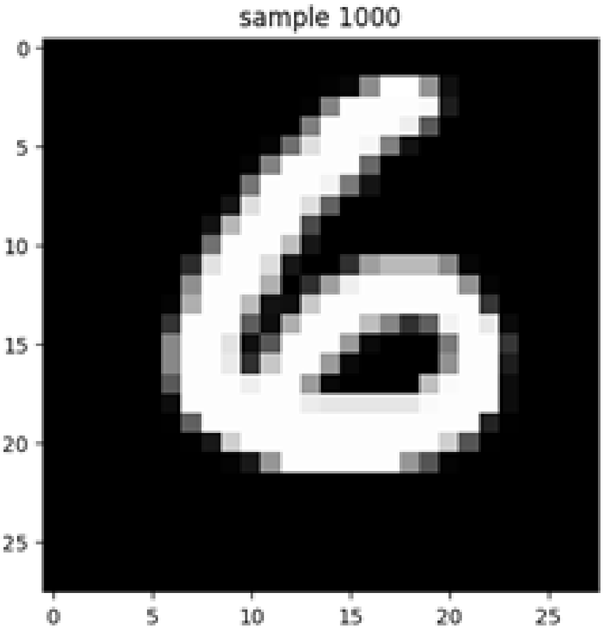


Figure 1: Handwriting Sample 1000

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. 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.

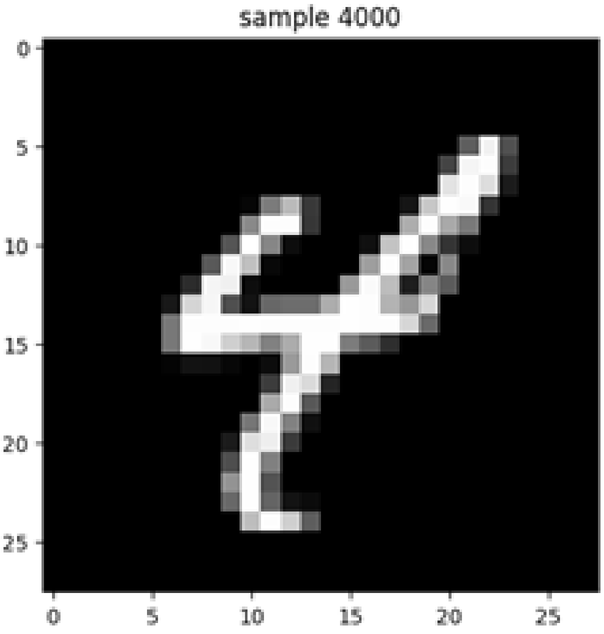


Figure 2: Handwriting Sample 4000

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.

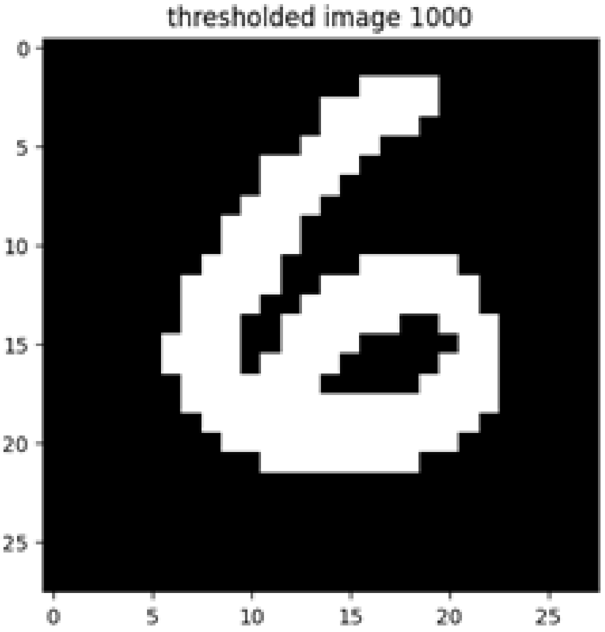


Figure 3: Threshold set to 128 of handwriting sample 1000

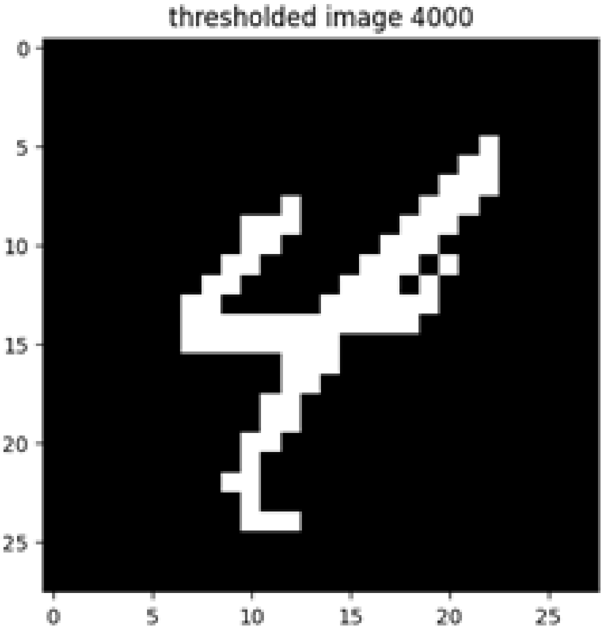


Figure 4: Threshold set to 128 of handwriting sample 4000

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.

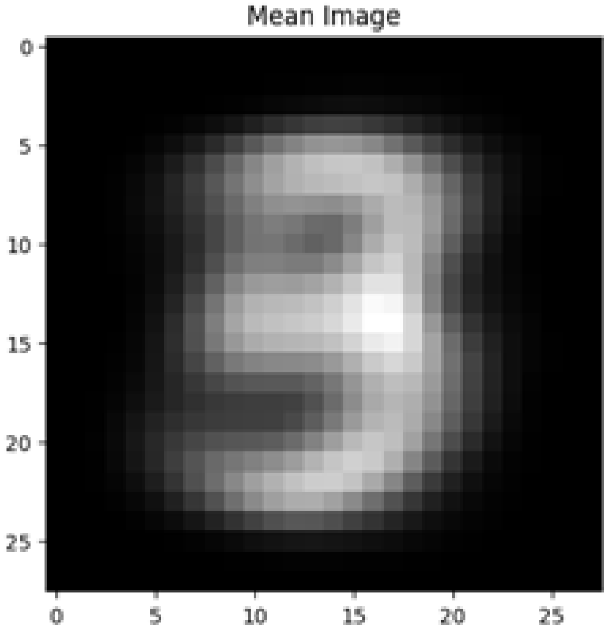


Figure 5: Mean Image

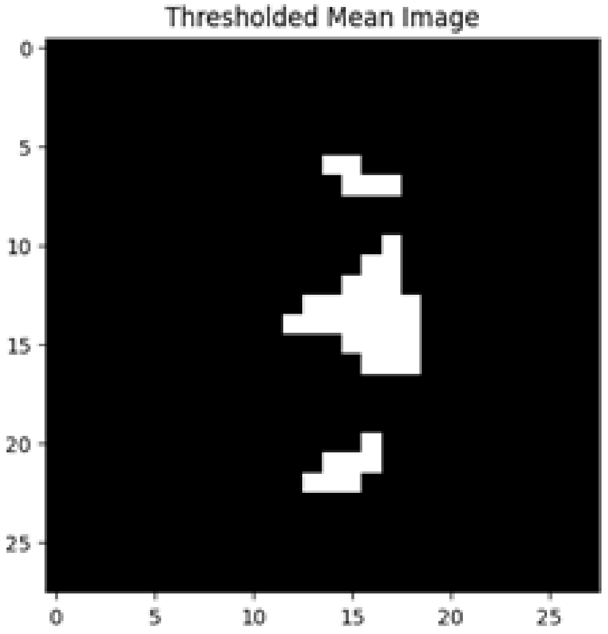


Figure 6: Threshold set to 128 of mean image

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

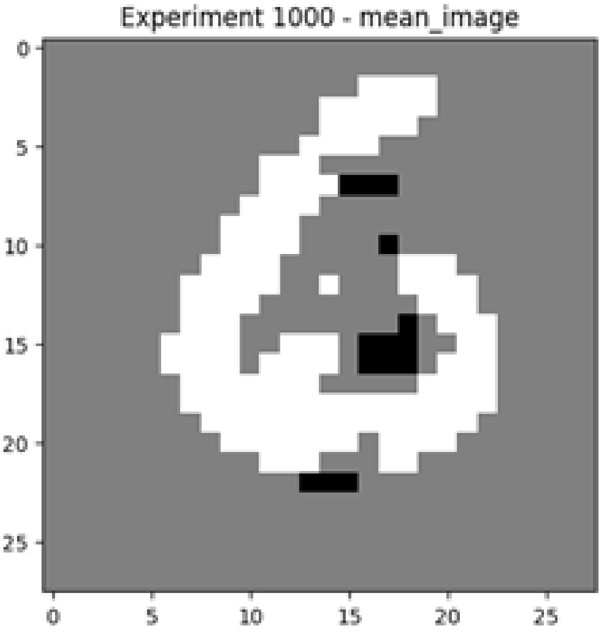


Figure 7: Subtraction of experiment image and mean image for experiment 1000

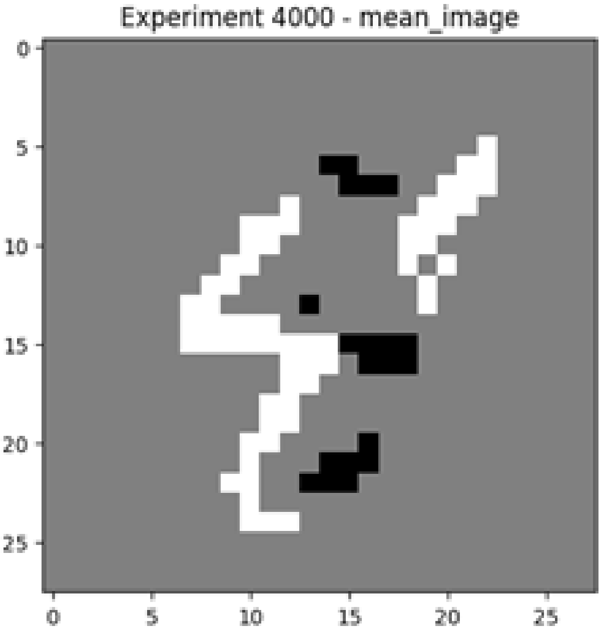


Figure 8: Subtraction of experiment image and mean image for experiment 4000

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.

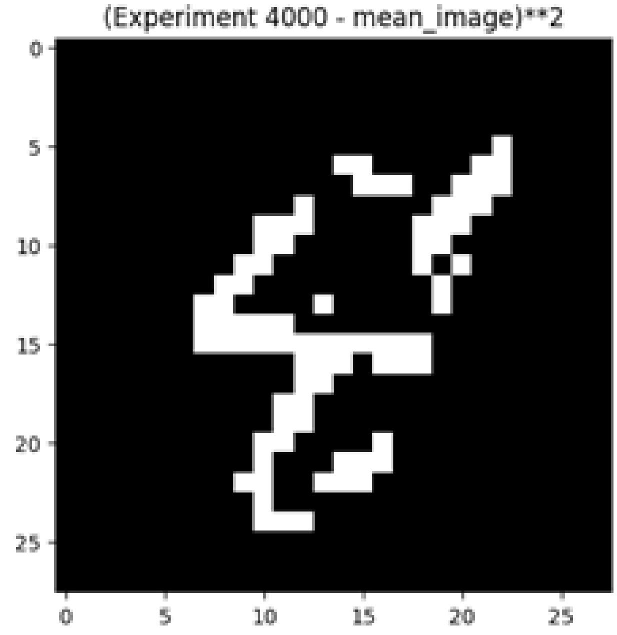


Figure 9: Square of the subtraction of experiment image and mean image for experiment 1000

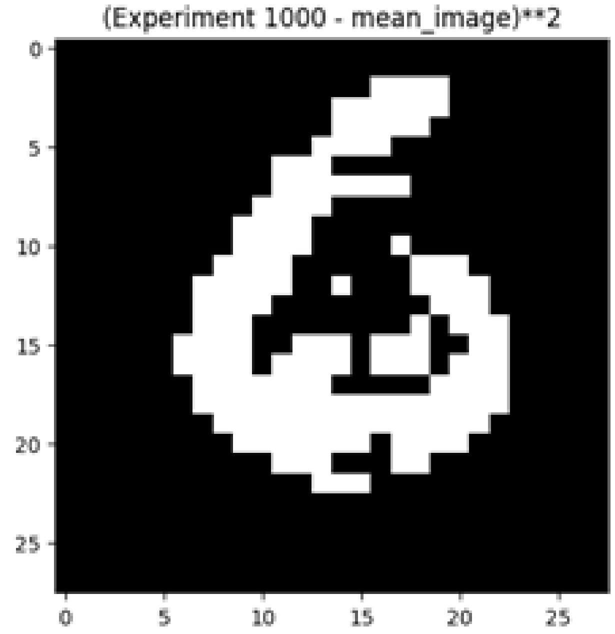
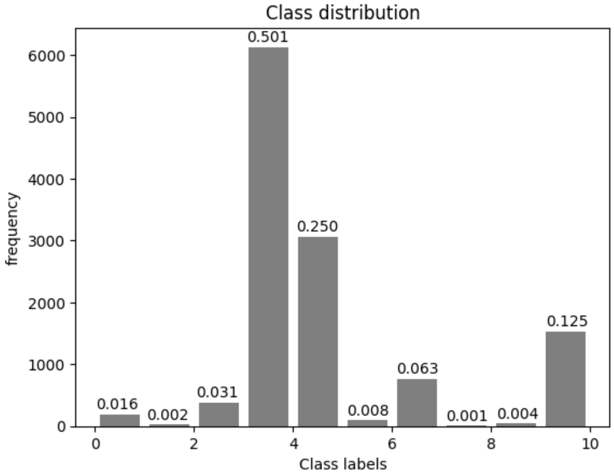


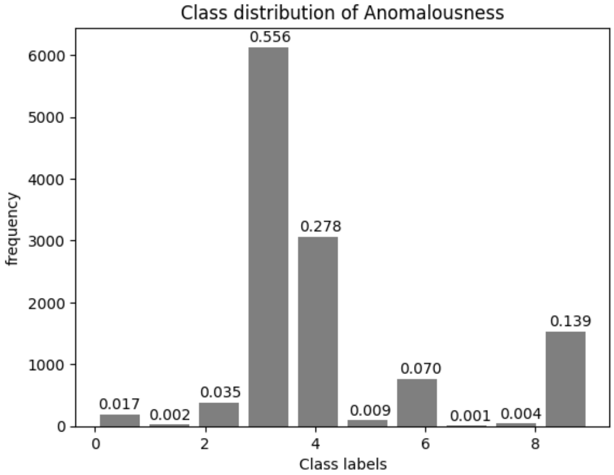
Figure 10: Square of the subtraction of experiment image and mean image for experiment 4000

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 **??** and figure **??**.

3 Results

The probabilities from the data set are in the following output excerpt: Probabilities for Each Class:





The accuracy of the method was 0.65485, or 65.5 %.

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 65.5 %. 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 method achieved nearly double the accuracy of a baseline random guess.

5 Conclusion