Experiment 1: Varying Hidden Units

In this experiment, we will be running our MLP on 3 different values of hidden units (n = 20, 50, 100) for 50 epochs. We will be fixing the momentum to 0.9 and the learning rate to 0.1. The results and analyses of the MLPs are below:

Final Test Accuracy:
For this table, I used the average of the final 3 epochs for each MLP to get an idea of the 'ending' test accuracy.

Hidden Units	Test Accuracy
20	94.91%
50	96.95%
100	97.36%

We can see that as the number of hidden units increases, the final accuracy inceases, though there are diminishing returns. I expect that an increase in hidden units beyond the 100 will stop improving the test accuracy by any significant margin.

Convergence:
In Figure 1, we can see that the increase in number of hidden units increases the number of epochs needed to reach a convergence or plateau in our accuracy. For the first MLP (n = 20), the training set accuracy converges at around 30 epochs while the test set converges very quickly at around 15-20 epochs. On the latter two MLPS, the training set continues to slowly increase in accuracy until around 40-45 epochs, while the test set plateaus at around 20-25 epochs. The tradeoff is that the latter two take about 50% more epochs to converge, however they have a higher test set final accuracy as seen above. In Figure 2, I have graphed the test set accuracy results on a single graph, in order to clearly see the effects of the increase in hidden units. We can clearly see the initial jump when we moved from 20 to 50 hidden units, as well as diminishing returns on the increase to 100 hidden units. We have slightly better results at the cost of more computations per epoch.

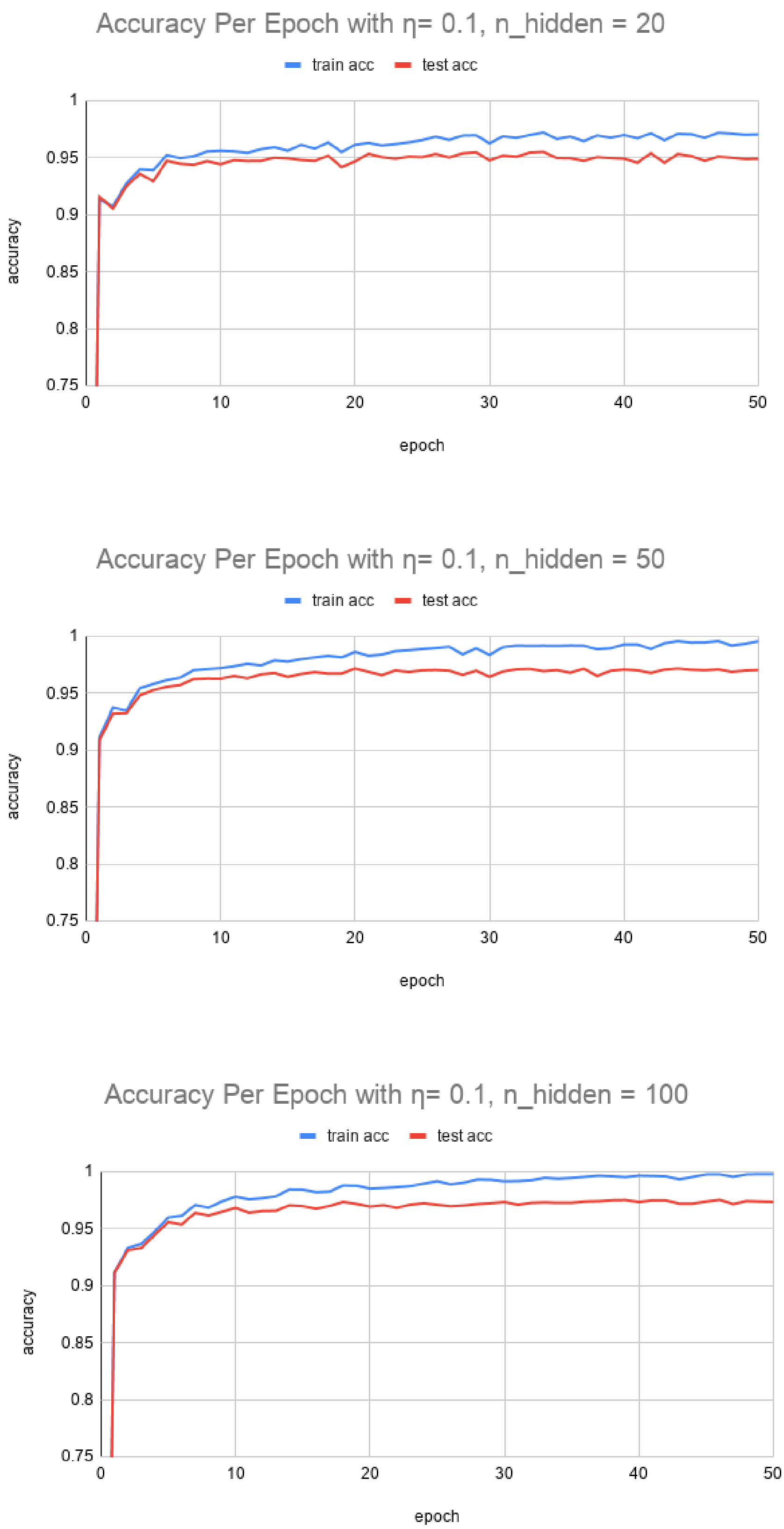

FIGURE 1. Varying Number of Hidden Units

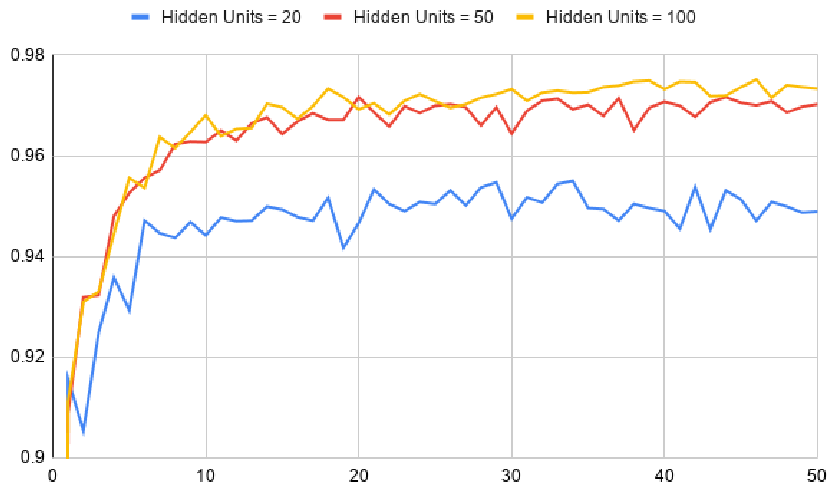

FIGURE 2. Comparing Test Set Accuracy

Overfitting:
In Figure 1, we can see that no overfitting occurred on any of the MLPs since we do not see the test set accuracy dropping off as the MLP 'memorizes' the training set data.

SLP Comparison:
My best test set accuracy using an SLP was only 83.10%, more than 14% worse than the MLP with 100 hidden units. We can see the significant difference in Figure 3, where the SLP plateaus quickly, but at the cost of test and training set accuracy.

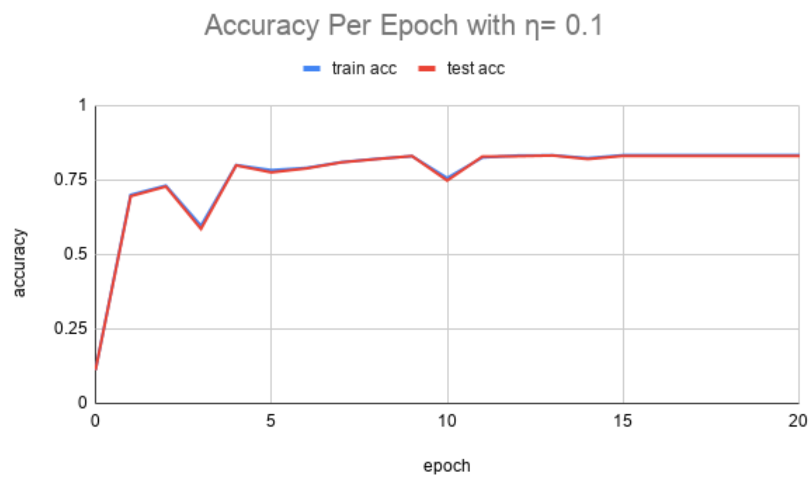

FIGURE 3. Single Layer Perceptron Results

	0	1	2	3	4	5	6	7	8	9
0	962	0	0	1	2	1	6	3	2	3
1	0	1118	3	1	0	1	3	4	5	0
2	9	3	972	16	0	0	0	8	21	3
3	4	3	8	947	0	13	2	6	24	3
4	3	1	5	0	886	0	4	4	6	73
5	8	1	0	21	2	824	7	2	21	6
6	11	3	2	1	9	11	910	0	11	0
7	1	12	12	7	0	3	0	974	2	17
8	7	1	1	10	3	4	6	5	936	1
9	3	3	0	6	12	6	0	8	11	960

Figure 4a. Confusion Matrix for 20 Hidden Units

	0	1	2	3	4	5	6	7	8	9
0	959	0	6	1	0	2	5	4	2	1
1	0	1119	3	1	0	1	2	2	6	1
2	4	3	994	6	3	0	3	11	6	2
3	1	0	2	976	0	15	0	6	7	3
4	2	0	2	1	947	1	6	0	2	21
5	3	2	0	8	3	863	6	1	4	2
6	6	1	1	0	5	9	935	0	1	0
7	1	4	9	5	1	2	0	990	3	13
8	5	2	3	7	2	12	2	3	935	3
9	3	1	1	5	7	3	0	4	1	984

Figure 4b. Confusion Matrix for 50 Hidden Units

	0	1	2	3	4	5	6	7	8	9
0	965	0	1	1	0	3	4	1	4	1
1	0	1124	2	1	0	1	2	1	4	0
2	4	2	999	7	1	0	5	5	8	1
3	0	0	3	988	0	8	0	1	4	6
4	2	0	6	1	947	1	5	3	0	17
5	4	1	0	9	2	860	5	1	8	2
6	7	2	1	1	3	4	936	1	3	0
7	0	6	11	3	1	0	0	992	3	12
8	3	0	1	10	1	5	2	2	947	3
9	2	3	1	5	8	5	1	7	2	975

Figure 4c. Confusion Matrix for 100 Hidden Units

In the confusion matricies in Figure 4, we see that the most common mistake is predicting a 9 when the target was a 4. I believe that is a very reasonable mistake, something even humans make from time to time. Seeing the high values on the diagonal is quite satisfying and the mistakes the MLP does make are reasonable, even for a machine.

Experiment 2: Varying Momentum

In this experiment, we will be running our MLP on 4 different momentum values (m = 0, 0.25, 0.5, 0.9) for 50 epochs. We will be fixing the number of hidden units to 100 and the learning rate to 0.1. The results and analyses of the MLPs are below:

Final Test Accuracy:
For this table, I used the average of the final 3 epochs for each MLP to get an idea of the 'ending' test accuracy.

Momentum	Test Accuracy
0	96.51%
0.25	96.73%
0.5	97.10%
0.9	97.36%

We can see that the momentum changes the final test set accuracy after 50 epochs only slightly. The main reason we see that difference is because the higher momentum values reach a plateau faster, which means it reaches a higher accuracy faster as we'll see below.

Convergence:
In Figure 5 below, we have the separate plotted results of varying the momentum term. However, it is easiest to see the power of the momentum term in Figure 6 below, where I've graphed the test set accuracy of 4 different MLPs with different momentum values. The highest momentum value, in green, reaches 97% after only 14 epochs, while the second highest momentum value, in yellow, takes 29 epochs. The two lower values don't even reach 97% after the full 50 epochs. Including a meaningful momentum term allows us to reach convergence more quickly, as well as giving us a higher final accuracy rate.

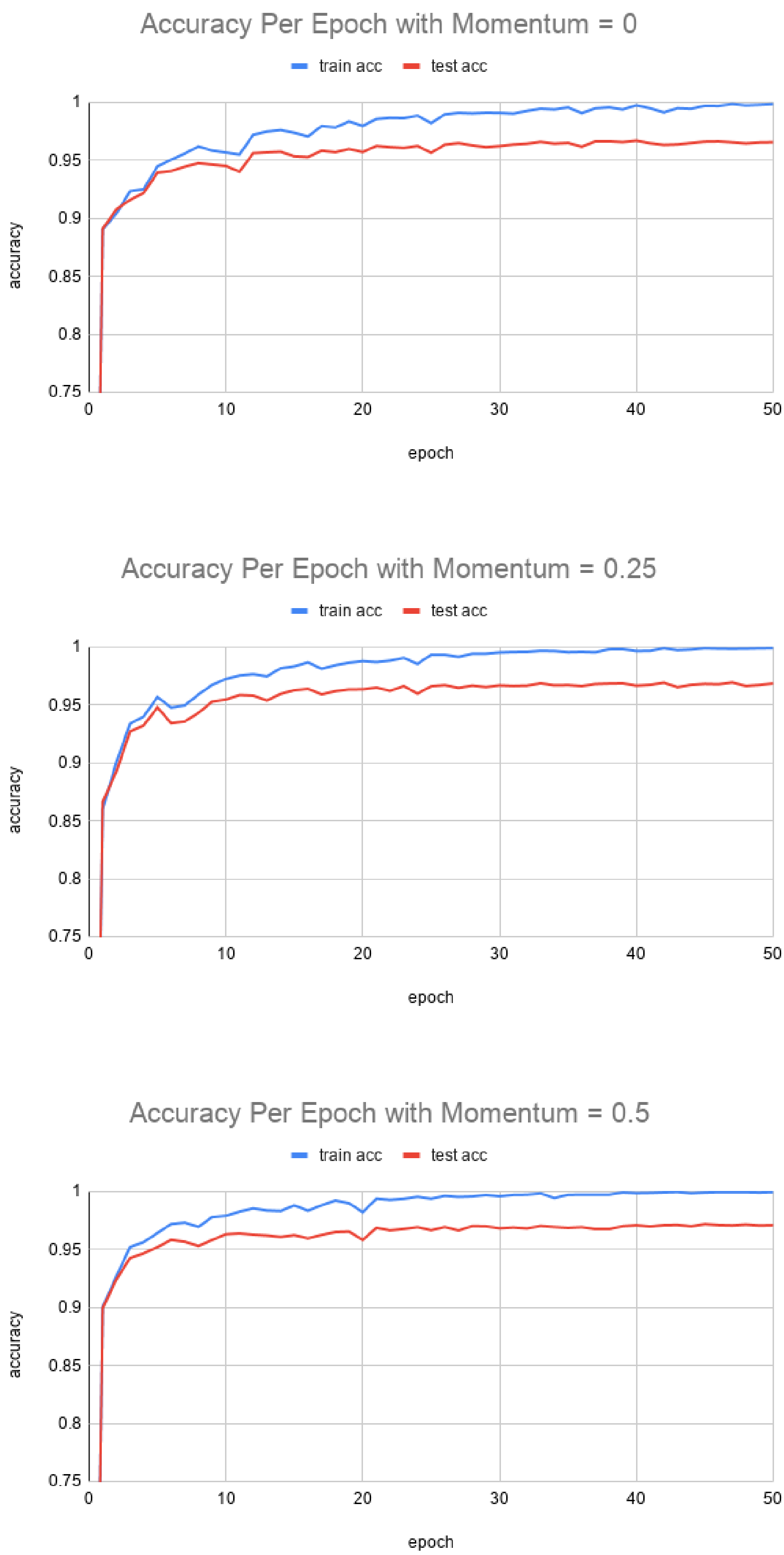

FIGURE 5. Varying Momentum Value

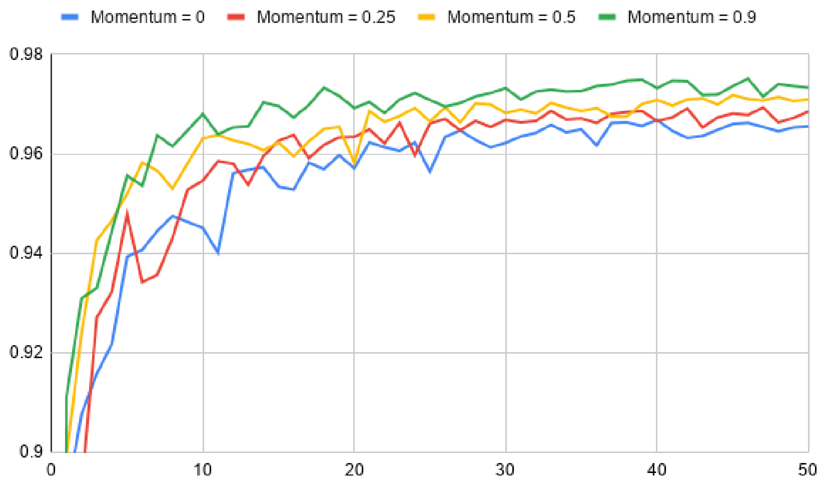

FIGURE 6. Comparing Test Set Accuracy

Overfitting:
Again, no evidence of overfitting was found in any iteration of these MLPs. No MLP had a significant decrease in test set accuracy as the number of epochs increased.

	0	1	2	3	4	5	6	7	8	9
0	962	0	1	1	1	5	6	1	3	0
1	0	1119	4	1	0	0	2	3	6	0
2	4	0	992	6	4	1	4	8	13	0
3	1	0	5	975	0	5	0	9	13	2
4	3	0	5	1	953	0	2	1	2	15
5	4	1	0	16	2	849	9	2	7	2
6	12	3	3	1	4	6	923	1	5	0
7	0	4	11	6	1	2	0	994	2	8
8	5	0	2	5	1	5	2	7	945	2
9	5	5	0	8	25	8	0	8	7	943

FIGURE 7A. Confusion Matrix for Momentum = 0

	0	1	2	3	4	5	6	7	8	9
0	965	0	1	0	0	4	4	3	3	0
1	0	1121	5	0	0	0	2	2	5	0
2	6	1	1001	2	2	0	1	9	10	0
3	0	0	7	981	0	4	0	9	7	2
4	1	0	5	1	947	2	3	3	2	18
5	4	1	1	7	4	856	8	1	7	3
6	8	3	4	0	8	9	922	0	4	0
7	0	5	6	5	1	0	0	1000	3	8
8	5	0	6	4	2	8	3	6	938	2
9	2	5	0	7	21	6	0	11	3	954

FIGURE 7B. Confusion Matrix for for Momentum = 0.25

	0	1	2	3	4	5	6	7	8	9
0	963	0	0	0	0	2	5	1	4	5
1	0	1119	4	1	0	0	2	3	6	0
2	7	2	1000	5	2	0	1	5	9	1
3	0	1	4	990	0	5	0	5	4	1
4	2	0	3	1	945	0	7	6	3	15
5	4	1	0	18	1	848	6	2	9	3
6	6	2	3	1	2	5	932	1	6	0
7	1	5	8	8	0	1	0	996	1	8
8	4	0	1	14	2	6	1	5	940	1
9	2	3	1	8	9	4	0	4	2	976

FIGURE 7C. Confusion Matrix for for Momentum = 0.5

As usual, 4s being mistaken as 9s (and vice-versa) is the biggest contributor to error in our MLPs.

Experiment 3: Varying Number of Training Examples

In this experiment, we will be running our MLP on 3 different amounts of training examples. (d = 15000, 30000, 60000) for 50 epochs. We will be fixing the number of hidden units to 100, momentum to 0.9, and the learning rate to 0.1. The results and analyses of the MLPs are below:

Final Test Accuracy:
For this table, I used the average of the final 3 epochs for each MLP to get an idea of the 'ending' test accuracy.

Number of Data Points	Test Accuracy
15k	95.25%
30k	96.46%
60k	97.36%

We can see that the number of data points significantly increases the final test set accuracy. Of the three variables that we tweaked in these experiments, the number of data points as well as number of hidden units were the most significant in terms of final test set accuracy. Even considering that the 60k test sees 4x as much data as the 15k test, we see that the 60k test at 10 epochs is still significantly better than the 15k test at 40 epochs.

Convergence:
In Figure 9 below, we can see that the number of data points does not significantly impact the number of epochs needed to converge. All three tests take approximately 20-30 epochs before they start to plateau, with the exception of the 15k example, which continues changing until around 40 epochs. It appears like more data points does contribute slightly to how fast the MLP converges, but it is not as significant of a factor as others. In Figure 8, we can see the comparison of the test set accuracies, which vary the most in accuracy compared to the first two experiments. This comparison shows that more data leads to a more accurate MLP in the most significant way.

Overfitting:
Again, no evidence of overfitting was found in any iteration of these MLPs. No MLP had a significant decrease in test set accuracy as the number of epochs increased.

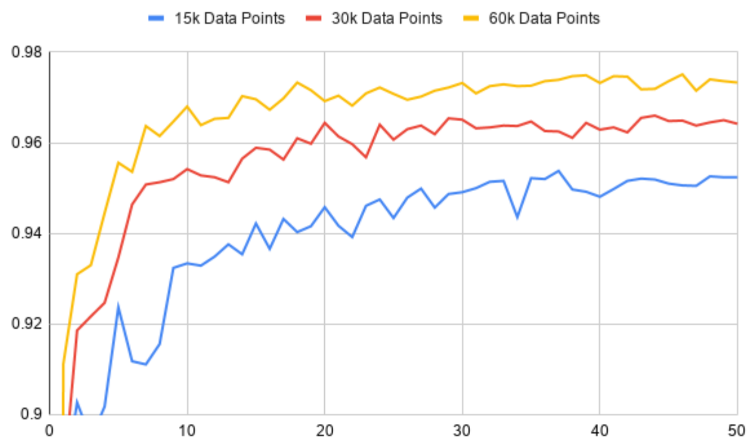

FIGURE 8. Comparing Varying Number of Data Points

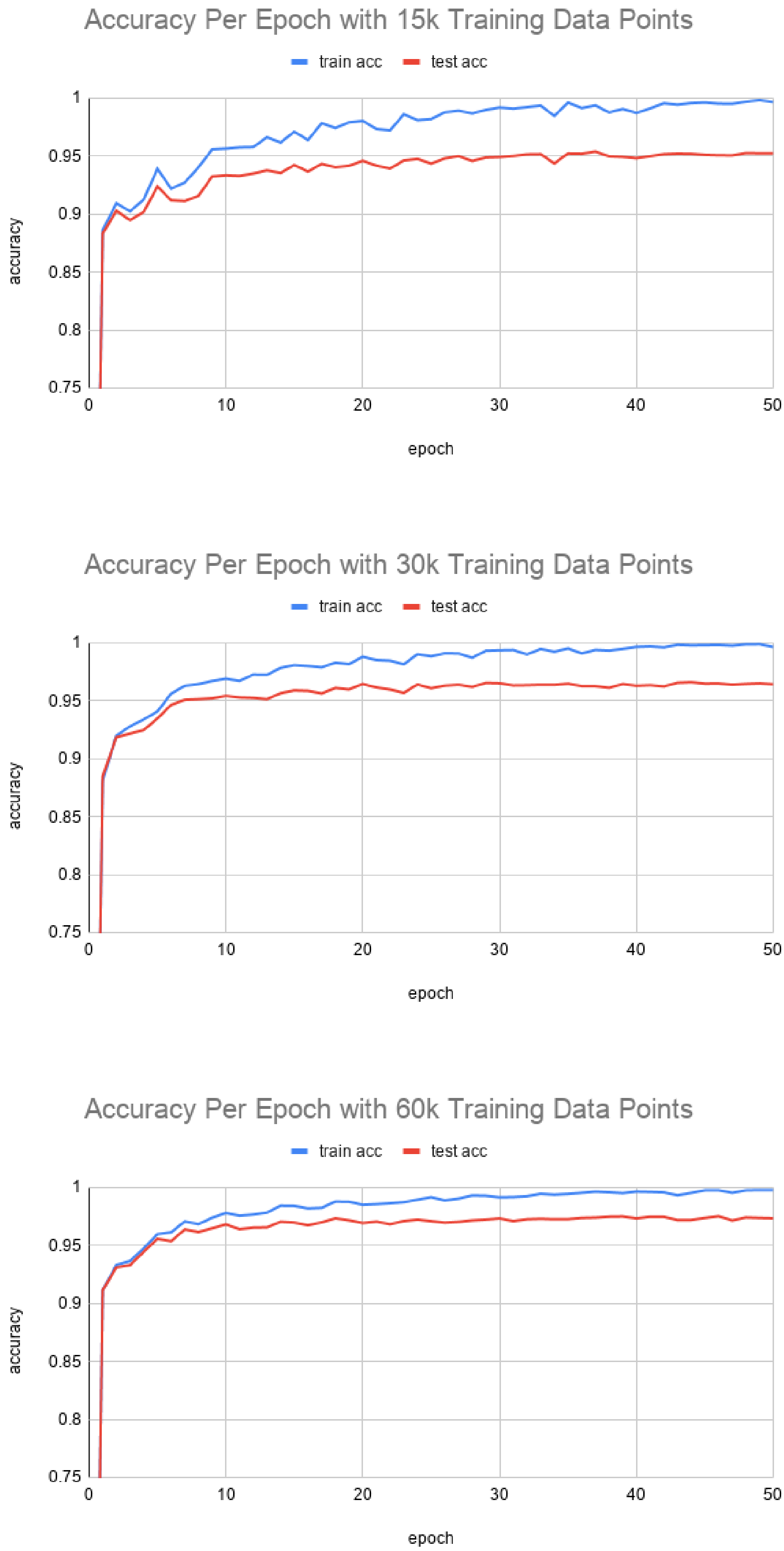

Figure 9. Varying Number of Data Points

	0	1	2	3	4	5	6	7	8	9
0	949	0	2	1	0	9	14	1	2	2
1	0	1112	8	4	0	0	2	5	4	0
2	7	1	994	7	6	0	5	8	3	1
3	1	0	9	976	0	9	0	5	8	2
4	1	0	7	0	957	0	4	1	1	11
5	8	1	1	33	4	813	12	2	9	9
6	6	3	11	2	7	6	917	3	3	0
7	0	8	17	9	3	0	0	970	2	19
8	5	0	9	22	5	9	11	5	894	14
9	4	1	1	10	36	3	0	7	5	942

FIGURE 10A. Confusion Matrix for 15k Data Points

	0	1	2	3	4	5	6	7	8	9
0	965	0	1	3	1	4	1	1	3	1
1	0	1123	3	2	0	1	2	2	2	0
2	5	2	986	14	2	2	1	10	10	0
3	0	0	1	996	1	5	0	3	2	2
4	1	0	5	1	961	0	5	3	1	5
5	6	3	0	21	2	834	13	1	9	3
6	7	2	5	2	4	8	927	1	2	0
7	1	6	8	9	4	0	0	983	1	16
8	4	1	9	23	2	10	2	11	910	2
9	3	5	1	13	20	5	0	5	0	957

FIGURE 10B. Confusion Matrix for for 30k Data Points

Surprisingly, these MLPs predicted 3s instead of 5s, 8s, and 9s as its most common mistake. This could be due to the limited number of data points, which means the MLP isn't trained against sloppily written numbers which leads to a lower success rate on the test set.