Programming Assignment 5 - Part 1

Section 1

a. Include the Name of the Group Members.

Group Members: Oleg Tielushko (U48136789), Tyler Brown

b. Separate the data into the training set 1 (85%) and training set 2 (15%) and indicate the percentage of each letter in the set.

Letter	Percentage from Total Data in the Set (100%)	Percentage from Set 1 Data (85%)	Percentage from Set 2 Data (15%)
A	3.95%	3.98%	3.77%
B	3.83%	3.78%	4.13%
C	3.68%	3.68%	3.70%
D	4.03%	4.03%	4.00%
E	3.84%	3.81%	4.00%
F	3.88%	3.82%	4.17%
G	3.87%	3.81%	4.17%
H	3.67%	3.66%	3.70%
I	3.78%	3.75%	3.93%
J	3.74%	3.76%	3.57%
K	3.70%	3.66%	3.87%
L	3.81%	3.82%	3.73%
M	3.96%	4.01%	3.70%
N	3.92%	3.85%	4.27%
O	3.77%	3.74%	3.93%

P	4.02%	4.07%	3.70%
Q	3.92%	4.10%	2.87%
R	3.79%	3.74%	4.07%
S	3.74%	3.78%	3.53%
T	3.98%	3.98%	4.00%
U	4.07%	4.09%	3.90%
V	3.82%	3.87%	3.53%
W	3.76%	3.76%	3.77%
X	3.94%	3.90%	4.13%
Y	3.93%	3.90%	4.10%
Z	3.67%	3.65%	3.77%
TOTAL:	100.00%	100.00%	100.00%

Section 2

Should include the results that WEKA provides when you use data set 1 to train the network selecting “use training set” option in the “test options” window.
The following items are to be included and clearly described in this section 2:

a.The overall accuracy achieved by the trained network.

84.0471%

```
=== Summary ===

Correctly Classified Instances       14288               84.0471 %
Incorrectly Classified Instances      2712               15.9529 %
Kappa statistic                          0.8341
Mean absolute error                      0.0144
Root mean squared error                  0.1035
Relative absolute error                 19.518  %
Root relative squared error             53.8307 %
Total Number of Instances            17000
```

b.The accuracy by which each letter was classified. Use a format similar to what you used in section 1.

Letter	Total in the Data Set (85%)	Number Identified Correctly by NN	Percentage Correct
A	676	617	91.27%
B	642	546	85.05%
C	625	537	85.92%
D	685	590	86.13%
E	648	574	88.58%
F	650	538	82.77%
G	648	447	68.98%
H	623	455	73.03%
I	637	506	79.43%
J	640	515	80.47%
K	623	495	79.45%
L	649	571	87.98%
M	681	631	92.66%
N	655	565	86.26%
O	635	540	85.04%
P	692	594	85.84%
Q	697	539	77.33%
R	636	509	80.03%
S	642	472	73.52%
T	676	564	83.43%
U	696	624	89.66%
V	658	584	88.75%
W	639	578	90.45%
X	663	609	91.86%
Y	663	572	86.27%
Z	621	516	83.09%

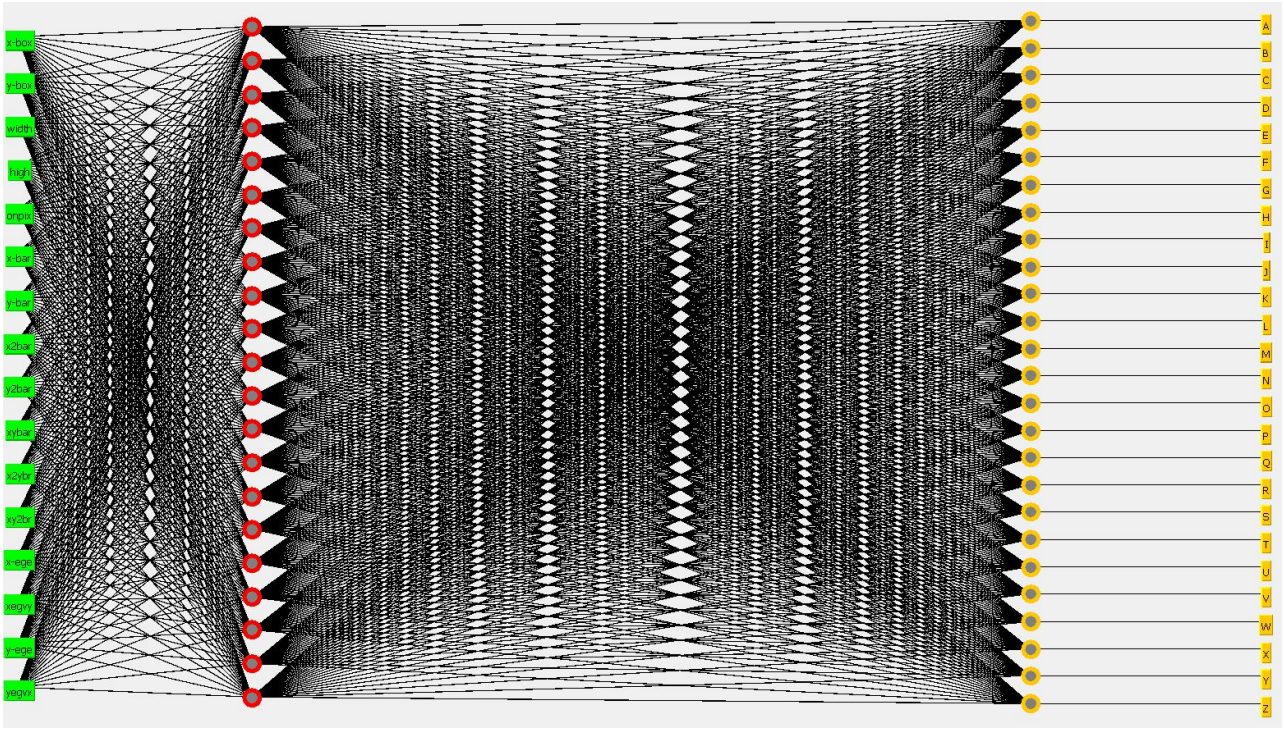
TOTAL:	17000	13671	83.97%

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   s   t   u   v   w   x   y   z   <-- classified as
 617   7   2   7   0   1   0   0   0   8   3   0   4   2   2   2   0   0   2   0  10   0   2   0   7   0 |   a = A
   2 546   1  12   8   6   1   1   1   0   3   0   0   0   8   1   0  27  21   0   0   0   0   3   1   0 |   b = B
   0   0 537   0  34   0   8   3   0   0  15   7   0   0   9   0   3   0   0   2   1   0   0   2   0   4 |   c = C
   5  15   0 590   0   3   0  18   0   2   1   0   3   4  14   4   0  19   2   0   2   0   0   3   0   0 |   d = D
   5  17   8   1 574  11   9   2   0   0   7   3   0   0   2   0   5   1   1   0   0   1   0   0   0   1 |   e = E
   1  40   2   3   7 538   0   0   1   1   1   0   0   0   0  13   0   2   4  21   0   0   2   2   7   5 |   f = F
  29   9  38   5   8   1 447   4   1   3  18   4   1   0   9  11  15  15   2   0   1   1   5   6   4  11 |   g = G
   4  30  12  22   1  10   3 455   0   1   9   0   2   4  11   4   2  25   0   0   7  10   1   4   6   0 |   h = H
  10   9   1  11   5  16   0   0 506   6   0   4   0   0   1  25   1   8  15   2   0   0   0  14   1   2 |   i = I
  16   8   0   9   3  13   0   2   6 515   0   0   0   0   2  40   0   1   5   6   0   0   0   7   4   3 |   j = J
   6  21  11   5   2   0   0  11   0   0 495   3   2   8   1   0   2  22   1   1   6   2   0  17   5   2 |   k = K
   0   9   0   3  14   0   9   4   0   0   7 571   0   0   3   0   2   5   4   0   0   0   0  16   1   1 |   l = L
   1   8  10   3   0   2   0   5   0   2   4   0 631   2   2   2   0   0   0   0   3   0   6   0   0   0 |   m = M
  11   4   4  26   0   4   0   3   0   3   3   2   4 565  13   3   0   2   0   0   2   0   4   0   2   0 |   n = N
   2   2  24   8   0   0   2   7   0   9   5   1   1   0 540   4   6   6   0   0   3   0  12   2   1   0 |   o = O
   1  17   5   5   1  43   3   0   1   0   4   0   0   1   2 594   2   1   0   1   0   0   2   0   9   0 |   p = P
   6  27   3   8  23   0   5   2   0   0   1   6   0   0  37   1 539   3   4   0   2   0   0   4   6  20 |   q = Q
  22  25   4  25   2   4   2   4   0   1  19   5   6   0   3   2   1 509   0   0   0   0   0   2   0   0 |   r = R
   5  44   0   6  12  20   0   0   2  11   0   5   0   0   6   9   1  11 472   2   0   6   0   5  10  15 |   s = S
   2   2   6  15  16   2   4   4   1   1   4  13   0   0   9   0   2   0   1 564   6   1   0  10   3  10 |   t = T
   5   2  13  11   0   0   0   6   0   0   3   3  13  11   0   0   1   0   0   0 624   0   4   0   0   0 |   u = U
   1  13  10  14   0   1   0   2   0   0   1   0   1   0   0   9   3   2   0   0   0 584  11   0   6   0 |   v = V
   1   4  11  21   0   0   0   6   0   2   5   0   4   1   0   0   1   0   0   0   3   0 578   0   2   0 |   w = W
   0   4   1   3   6   6   0   4   0   0   5   1   0   0   3   0   0   5   1   2   2   3   0 609   6   2 |   x = X
   0   3   0   7   5  12   0   0   0   0   0   1   1   0   0   1   7   0   1  21   3  22   0   6 572   1 |   y = Y
  16   7   0   0  28   2   0   0   0   0   0   4   0   0   0   2   0   5  24   0   0   0   0   6  11 516 |   z = Z
```

c.The number of layers in the network and the number of neurons in each layer
The Neural Network consists of the:

- Input Layer (Attributes)
- Hidden Layer: Node 26 through Node 46 (21 nodes total)
- Output Layer: Node 0 through Node 25 (26 nodes total)



d. Save the network in case this is the best network you will find.
DONE

e. A paragraph discussing the results obtained including -- best accuracy and worst accuracy on a letter, and any other points that you want to include to show that you understand the process of building the network.

The results that we have received have proven itself to be poorly accurate on determining the letter to be found. The average of almost 84% means that with the quite rich supply of records (17000), our neural network (NN) still fails to identify a significant amount of data and needs to be trained more accurately. The letter that the NN has most success in identifying is letter M, and in our opinion that could be because of its distinct shape and positioning, as such the NN has no problems identifying it correctly. Now, if we look at the cluster of letters with low recognition rate: G, H, S, Q we can clearly see how the neural network often confuses these letters with similar looking ones. G with most A and C; H with B, D, and R, and so on. This tells us we should be adding more intermediate layers and/or perform more training epochs for our neural network such that it can create better “intermediate” conclusions and differentiate the letters more efficiently.

Section 3

Should include the results that WEKA provides when data set 2 records are tested on the network described on section 2. The following items are to be included and clearly described in this section 3:

f.The overall accuracy achieved by the network on this test file
84.2667%

```
=== Summary ===

Correctly Classified Instances        2528               84.2667 %
Incorrectly Classified Instances       472               15.7333 %
Kappa statistic                          0.8363
Mean absolute error                      0.0143
Root mean squared error                  0.1016
Total Number of Instances             3000
```

g.The accuracy by which each letter was classified. Use a format similar to that used in sections1 and 2.

Letter	Percentage Accuracy on Test Set 2 (15%)	Percentage Correct on Training Set 1 (85%) for comparison
A	92.00%	91.27%
B	79.80%	85.05%
C	86.50%	85.92%
D	83.30%	86.13%
E	87.50%	88.58%
F	80.80%	82.77%
G	73.60%	68.98%
H	67.60%	73.03%
I	89.00%	79.43%
J	72.90%	80.47%
K	74.10%	79.45%

L	93.80%	87.98%
M	91.00%	92.66%
N	81.30%	86.26%
O	83.90%	85.04%
P	90.10%	85.84%
Q	75.60%	77.33%
R	86.10%	80.03%
S	67.00%	73.52%
T	87.50%	83.43%
U	94.90%	89.66%
V	89.60%	88.75%
W	92.90%	90.45%
X	92.70%	91.86%
Y	87.80%	86.27%
Z	86.70%	83.09%
MEAN:	84.3%	83.97%

h.A paragraph discussing the results obtained including – do the best and worst accuracies correspond to the same letters from section 2 and any other points that you want to include to show that you understand the process of building the network.

In general, the difference between the sets that we used for the training and for testing the network had a similar range of correct guessing. The general percent difference in training vs. testing set did not go above 4-5%. If we are to take the worst accuracy for the Training set, that would be G with accuracy of 68.98%, and for the Testing set that letter was H with a similar accuracy number of 67.60%. As far as the best accuracy letter, for the Training set, such letter was M with accuracy rate of 92.66% and for Testing it was U with accuracy of 94.9%. We can clearly see the difference in the letters represented by the best and worst accuracies, however it is important to point out that there is a very small percentage difference between the best and worst accuracies, which can indicate the level of the training of the network. In this case, it is clear that we need to modify the construction of our Neural Network and add additional epoch to the training time as well as increase the number of hidden layers within the network.