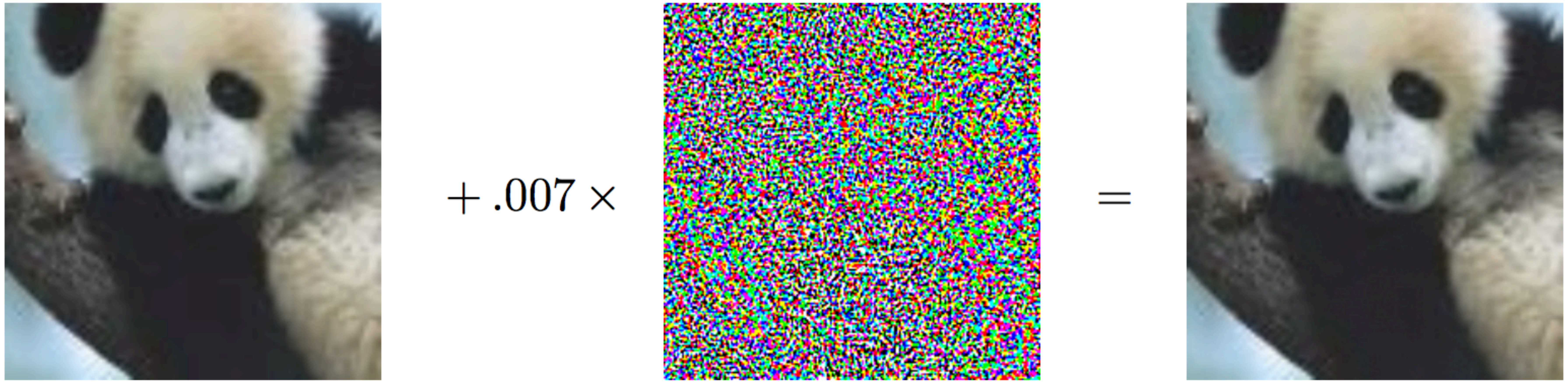
Adversarial Attack With Overfitting

Adversarial Attack



$+ .007 \times$ $=$

$\boldsymbol{x}$

"panda"
57.7% confidence

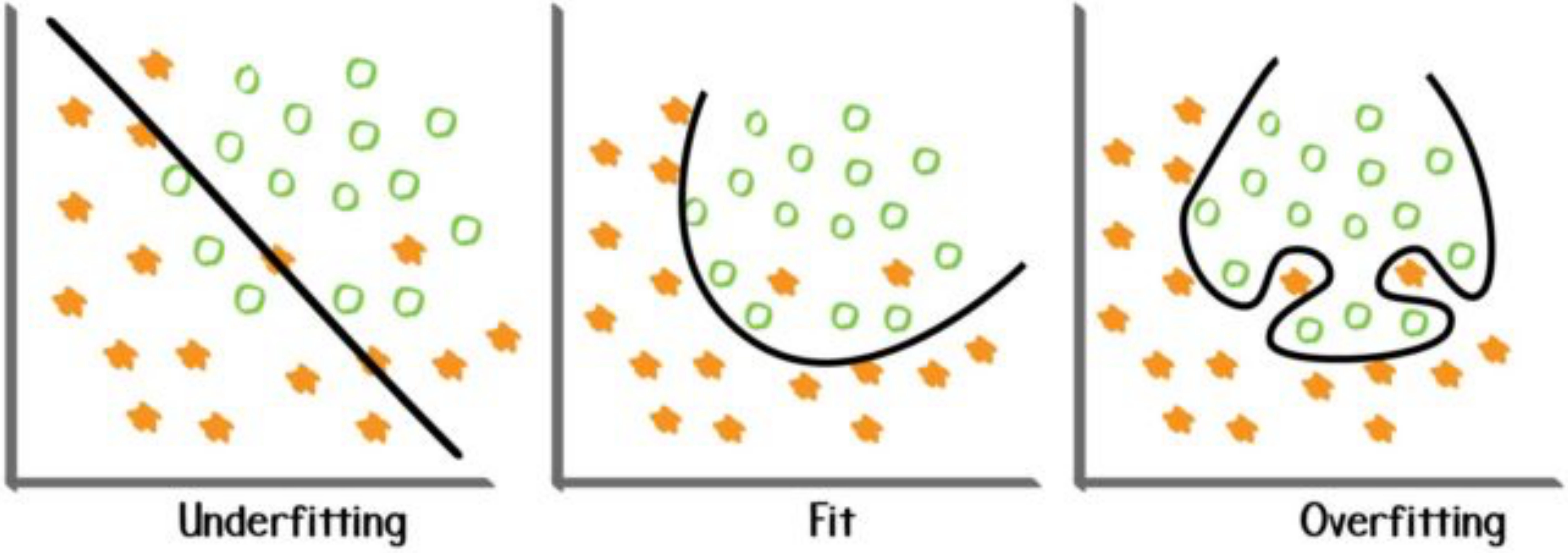
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

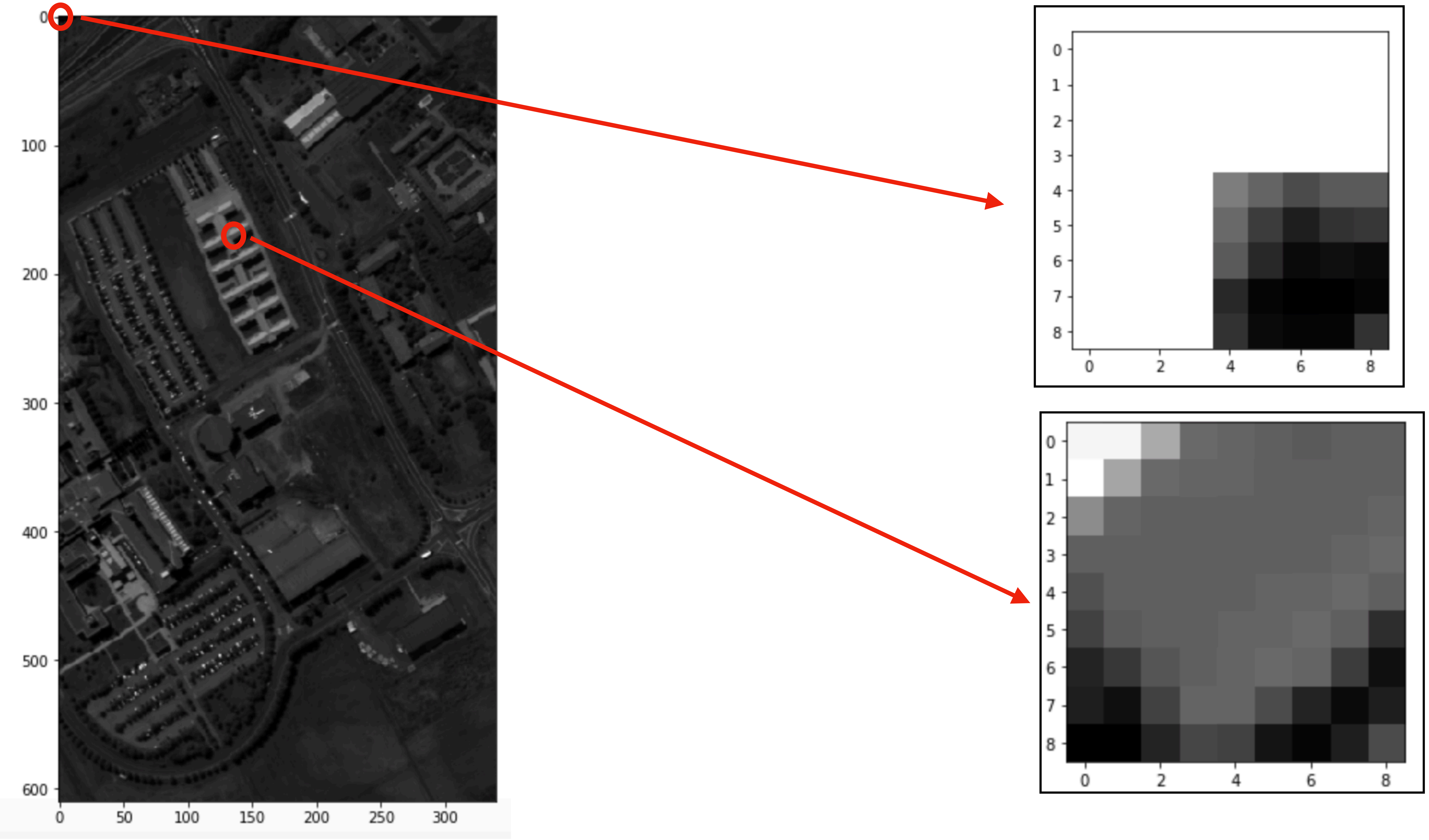
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Overfitting

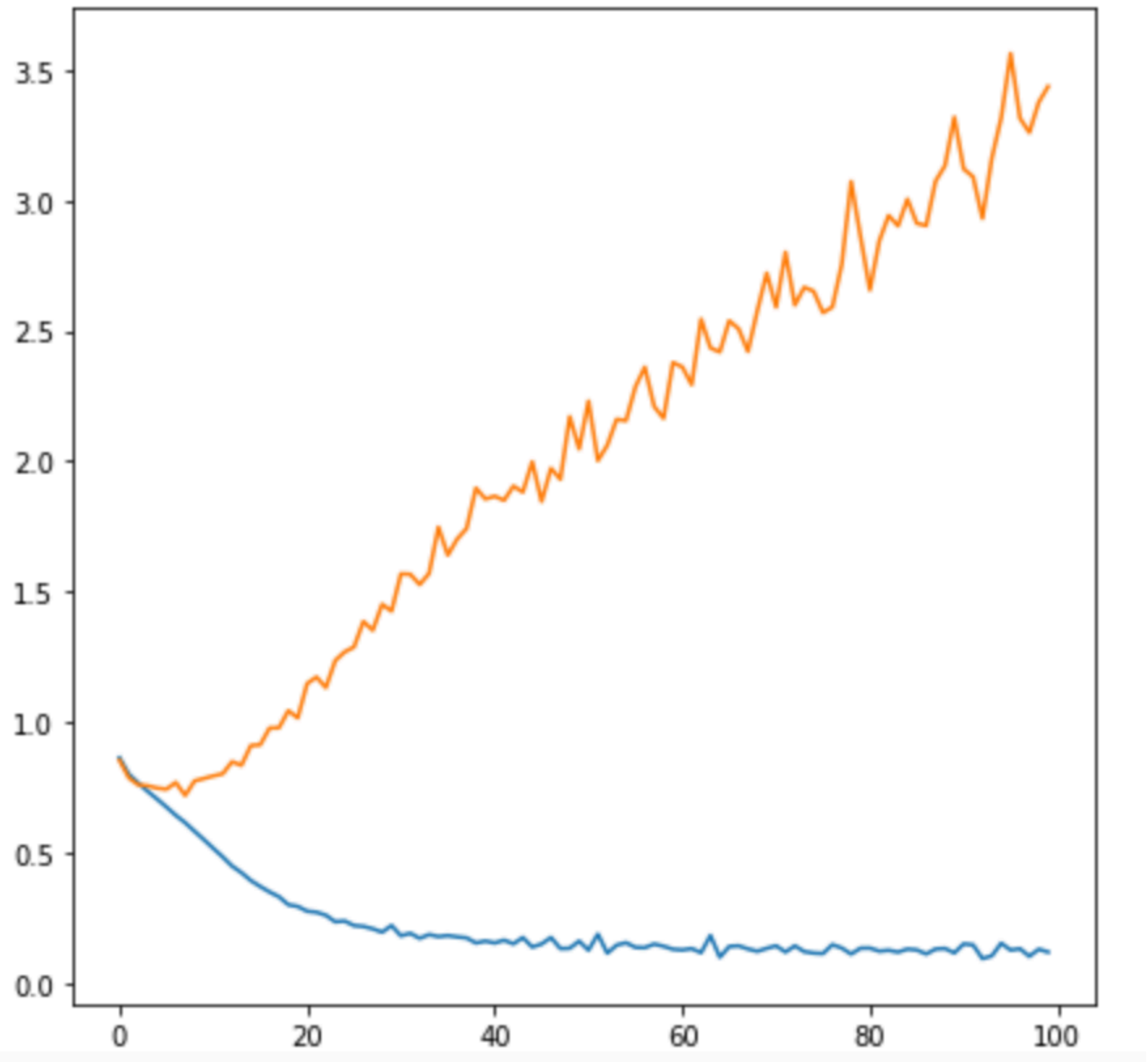


Data :Pavia University Dataset



Pavia University Dataset Overfitting

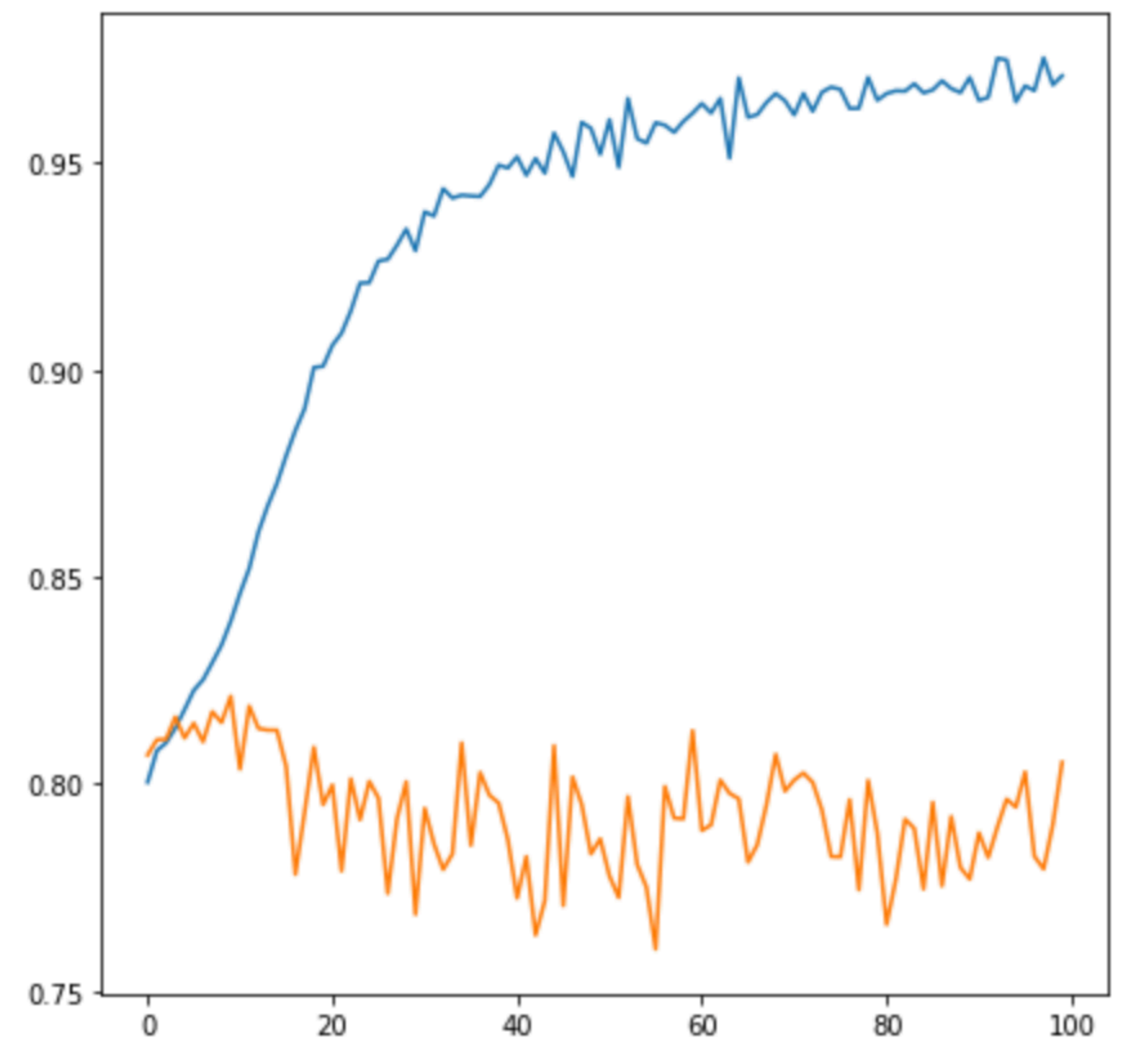
Loss



▬ Training Data set

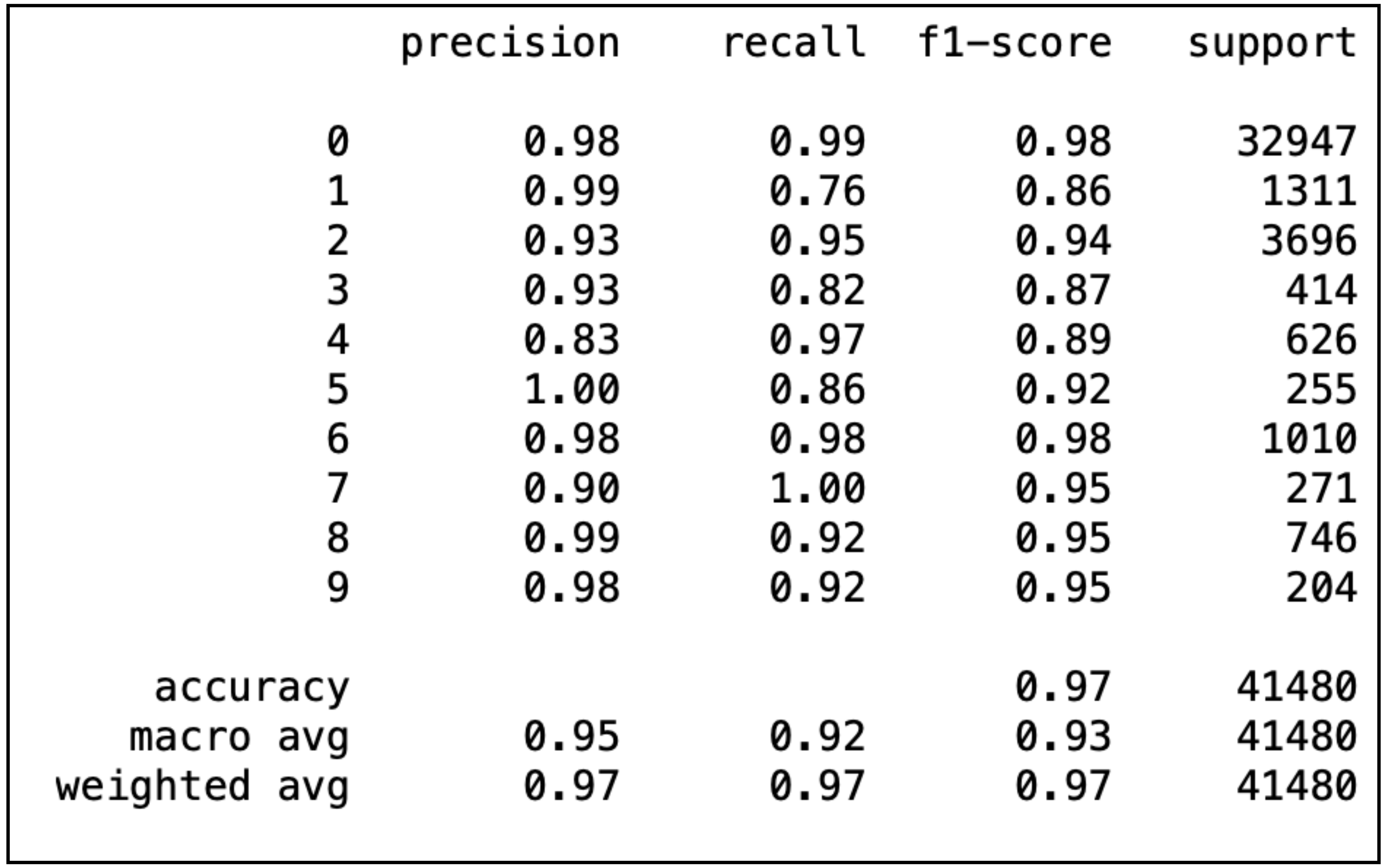
▬ Testing Data set

Accuracy



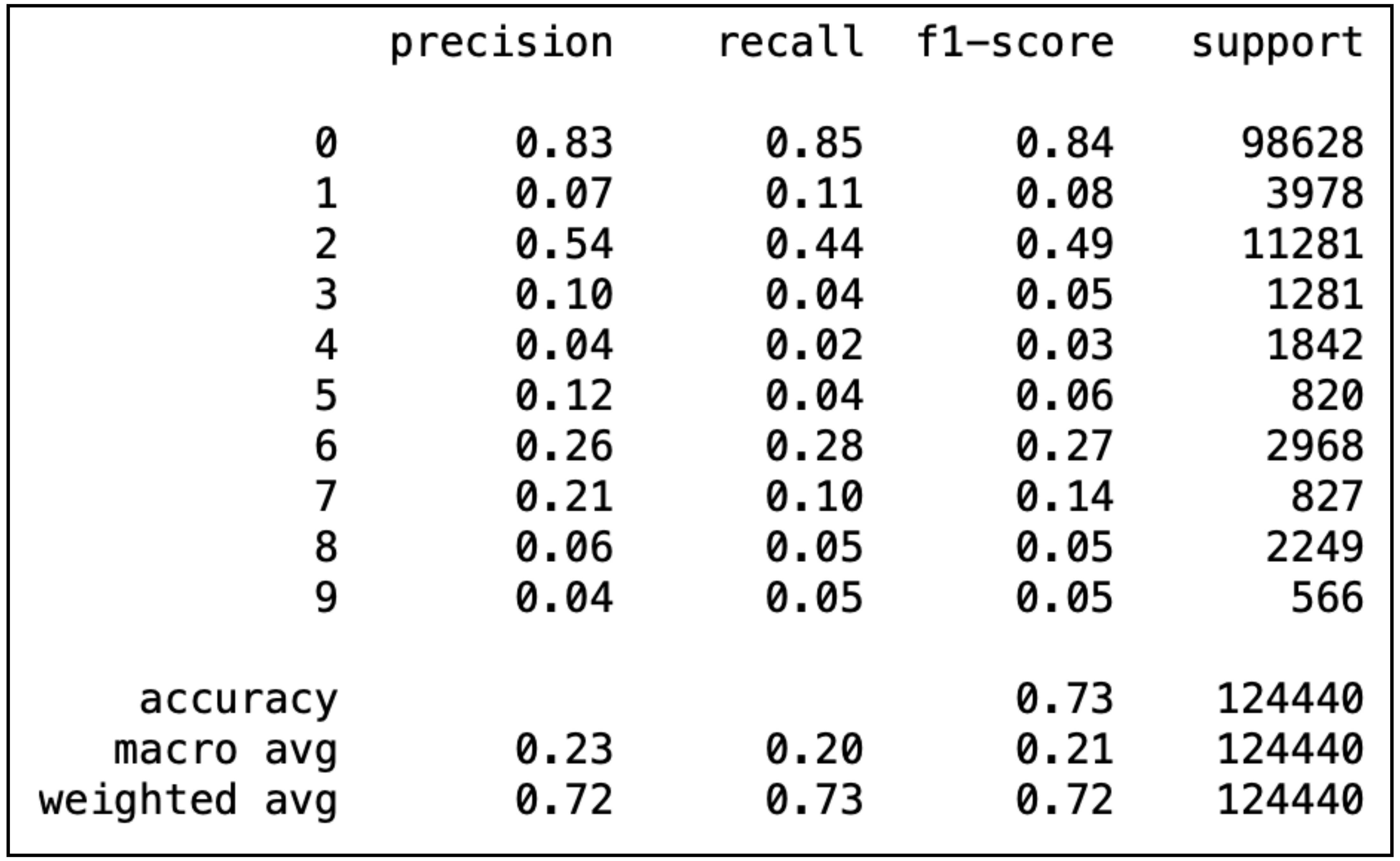
Pavia University Dataset Overfitting

Classification Report



```
              precision    recall  f1-score   support

           0       0.98      0.99      0.98     32947
           1       0.99      0.76      0.86      1311
           2       0.93      0.95      0.94      3696
           3       0.93      0.82      0.87       414
           4       0.83      0.97      0.89       626
           5       1.00      0.86      0.92       255
           6       0.98      0.98      0.98      1010
           7       0.90      1.00      0.95       271
           8       0.99      0.92      0.95       746
           9       0.98      0.92      0.95       204

    accuracy                           0.97     41480
   macro avg       0.95      0.92      0.93     41480
weighted avg       0.97      0.97      0.97     41480
```

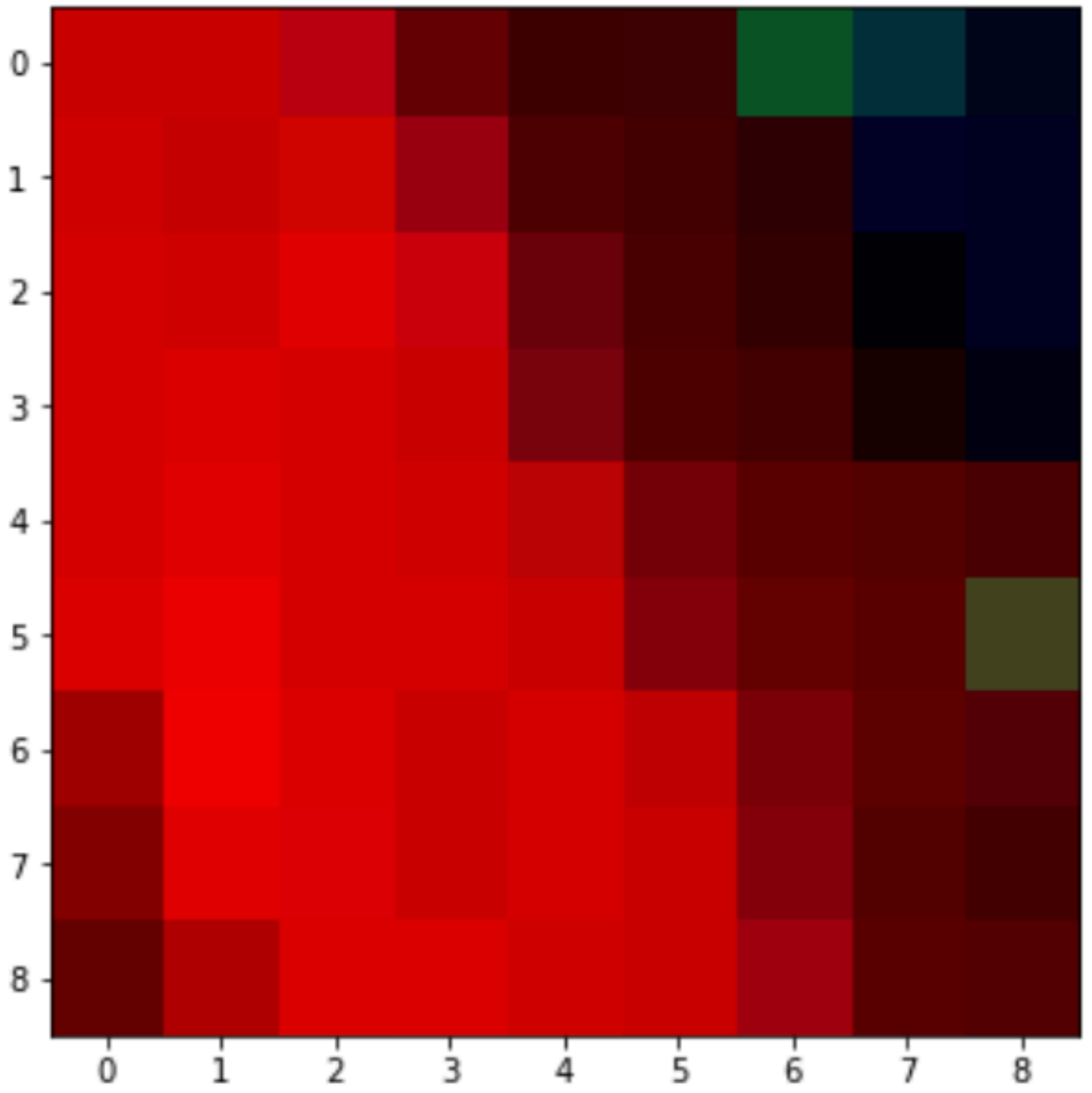
Training Data set



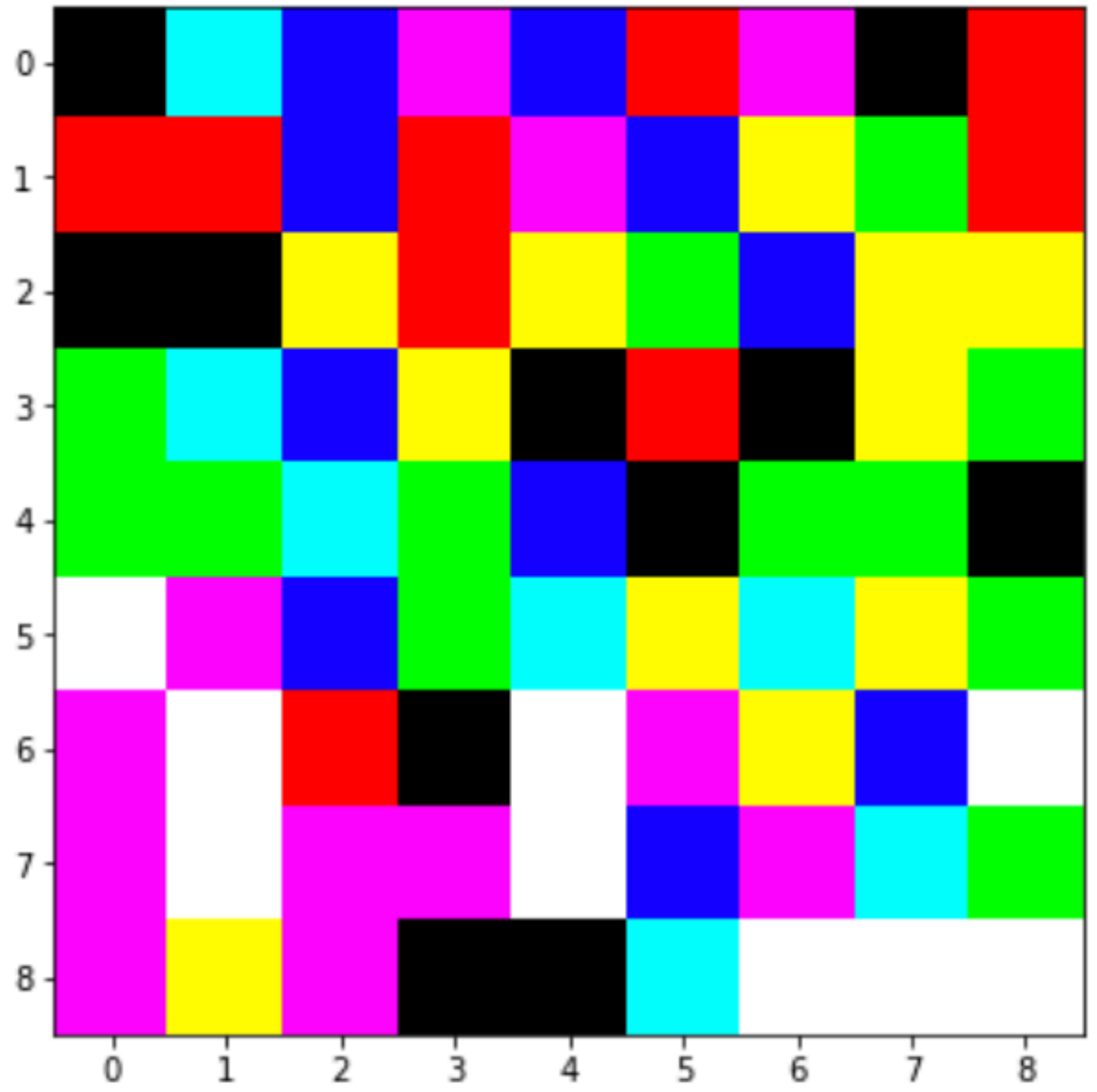
```
              precision    recall  f1-score   support

           0       0.83      0.85      0.84     98628
           1       0.07      0.11      0.08      3978
           2       0.54      0.44      0.49     11281
           3       0.10      0.04      0.05      1281
           4       0.04      0.02      0.03      1842
           5       0.12      0.04      0.06       820
           6       0.26      0.28      0.27      2968
           7       0.21      0.10      0.14       827
           8       0.06      0.05      0.05      2249
           9       0.04      0.05      0.05       566

    accuracy                           0.73    124440
   macro avg       0.23      0.20      0.21    124440
weighted avg       0.72      0.73      0.72    124440
```

Testing Data set

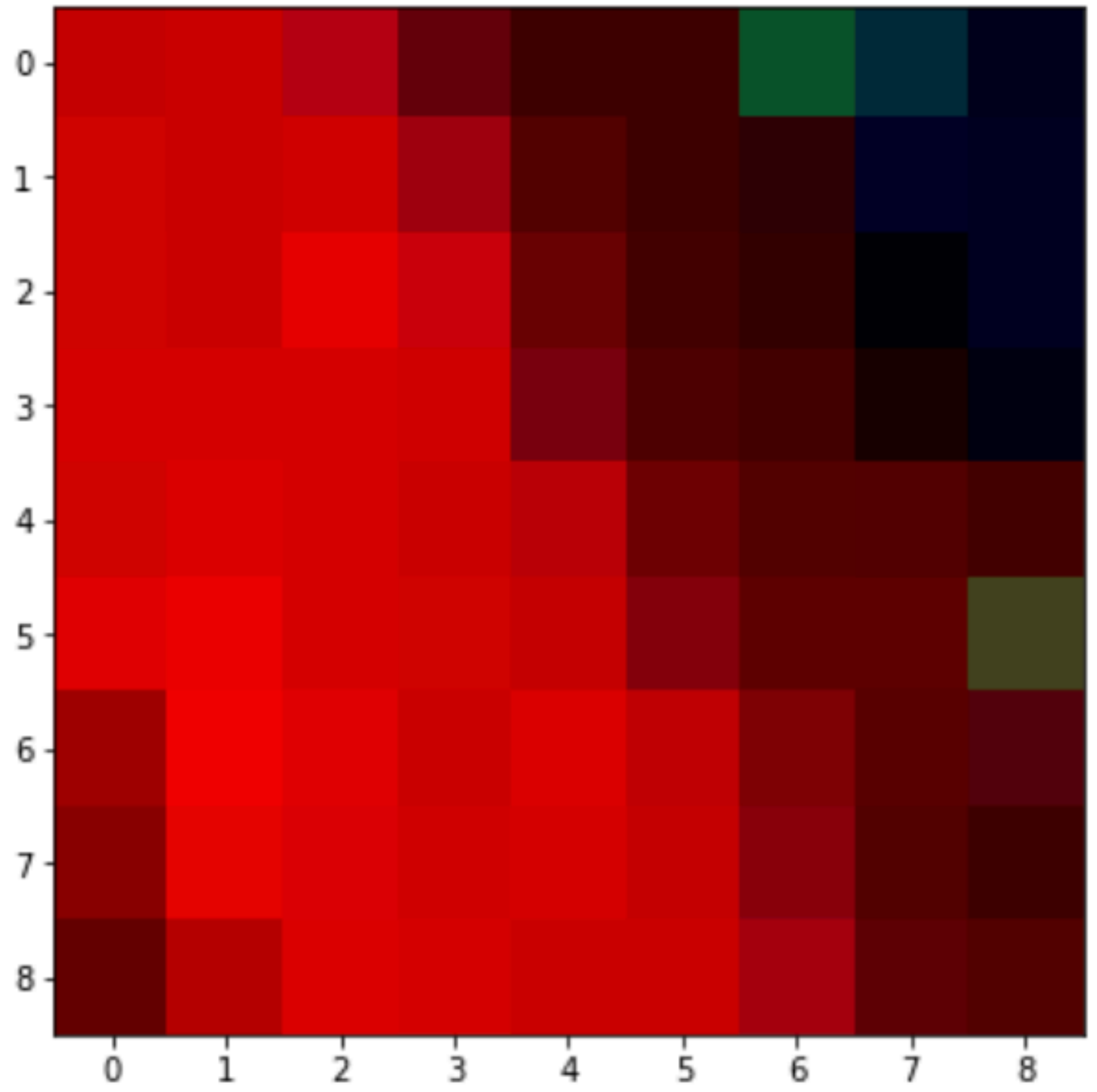
Add FGSM Adversarial Noise



+ 0.01 *

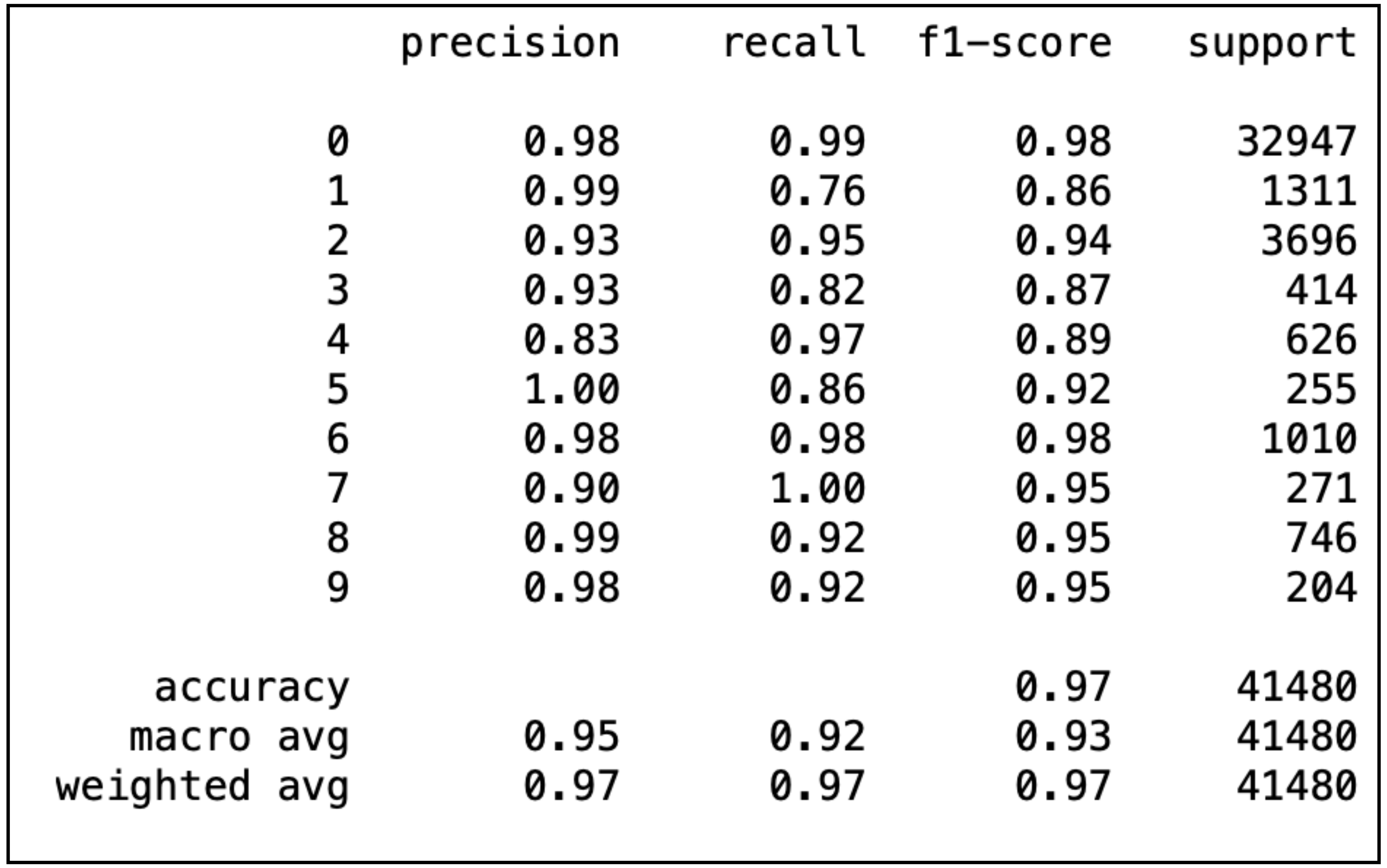


=



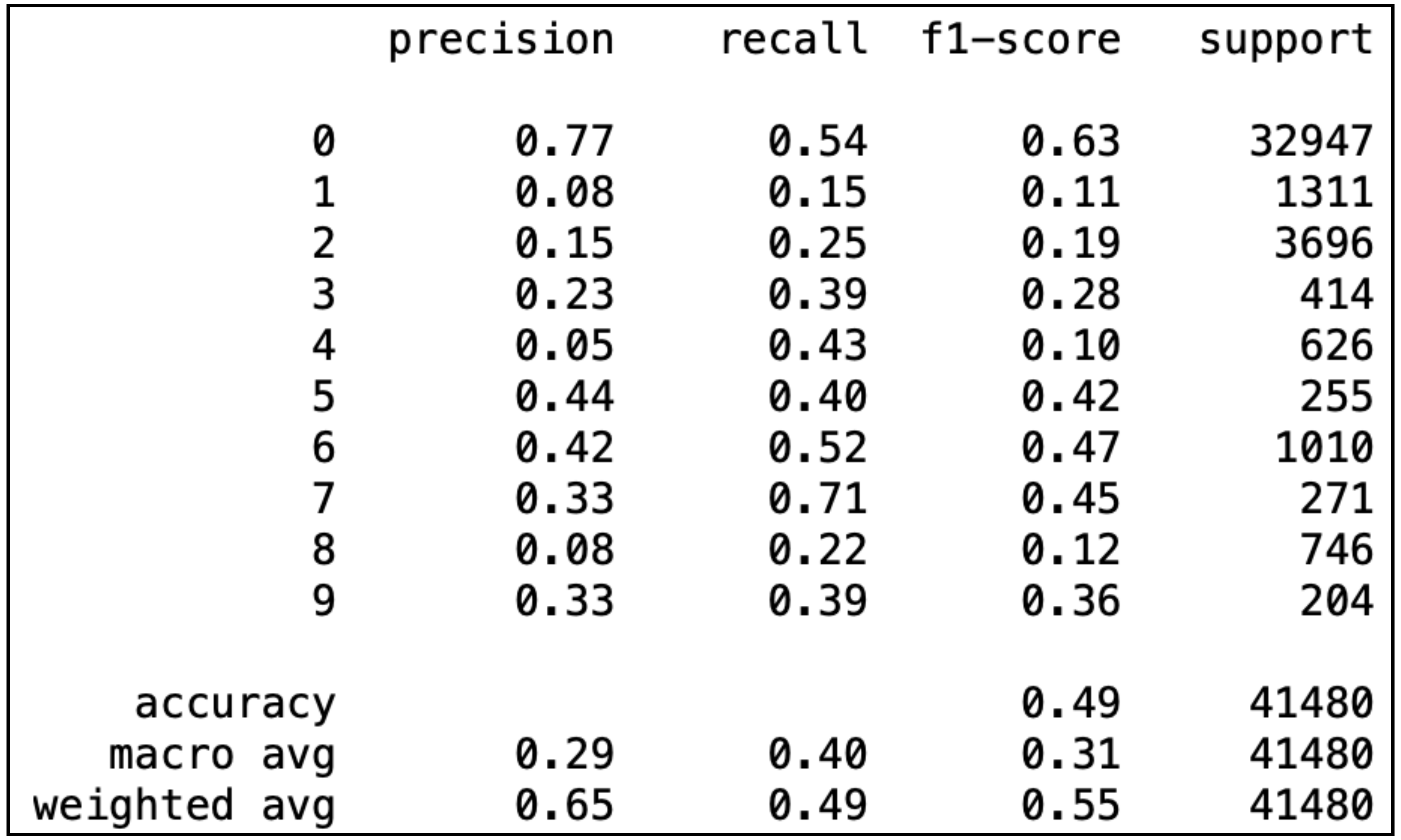
Result after add noise

Classification Report



```
              precision    recall  f1-score   support

           0       0.98      0.99      0.98     32947
           1       0.99      0.76      0.86      1311
           2       0.93      0.95      0.94      3696
           3       0.93      0.82      0.87       414
           4       0.83      0.97      0.89       626
           5       1.00      0.86      0.92       255
           6       0.98      0.98      0.98      1010
           7       0.90      1.00      0.95       271
           8       0.99      0.92      0.95       746
           9       0.98      0.92      0.95       204

    accuracy                           0.97     41480
   macro avg       0.95      0.92      0.93     41480
weighted avg       0.97      0.97      0.97     41480
```

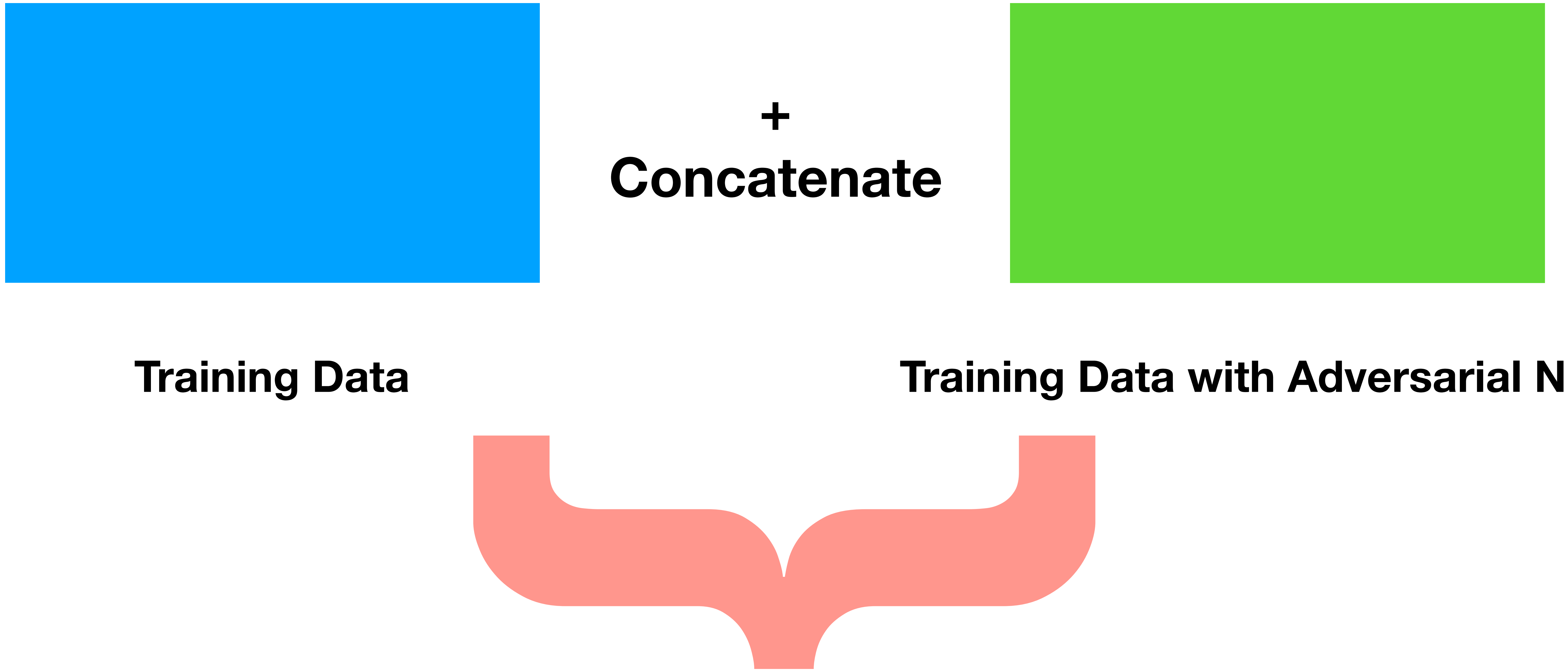
Training Data set



```
              precision    recall  f1-score   support

           0       0.77      0.54      0.63     32947
           1       0.08      0.15      0.11      1311
           2       0.15      0.25      0.19      3696
           3       0.23      0.39      0.28       414
           4       0.05      0.43      0.10       626
           5       0.44      0.40      0.42       255
           6       0.42      0.52      0.47      1010
           7       0.33      0.71      0.45       271
           8       0.08      0.22      0.12       746
           9       0.33      0.39      0.36       204

    accuracy                           0.49     41480
   macro avg       0.29      0.40      0.31     41480
weighted avg       0.65      0.49      0.55     41480
```

Training Data with Adversarial Noise

Add Noise data to train model



+

Concatenate

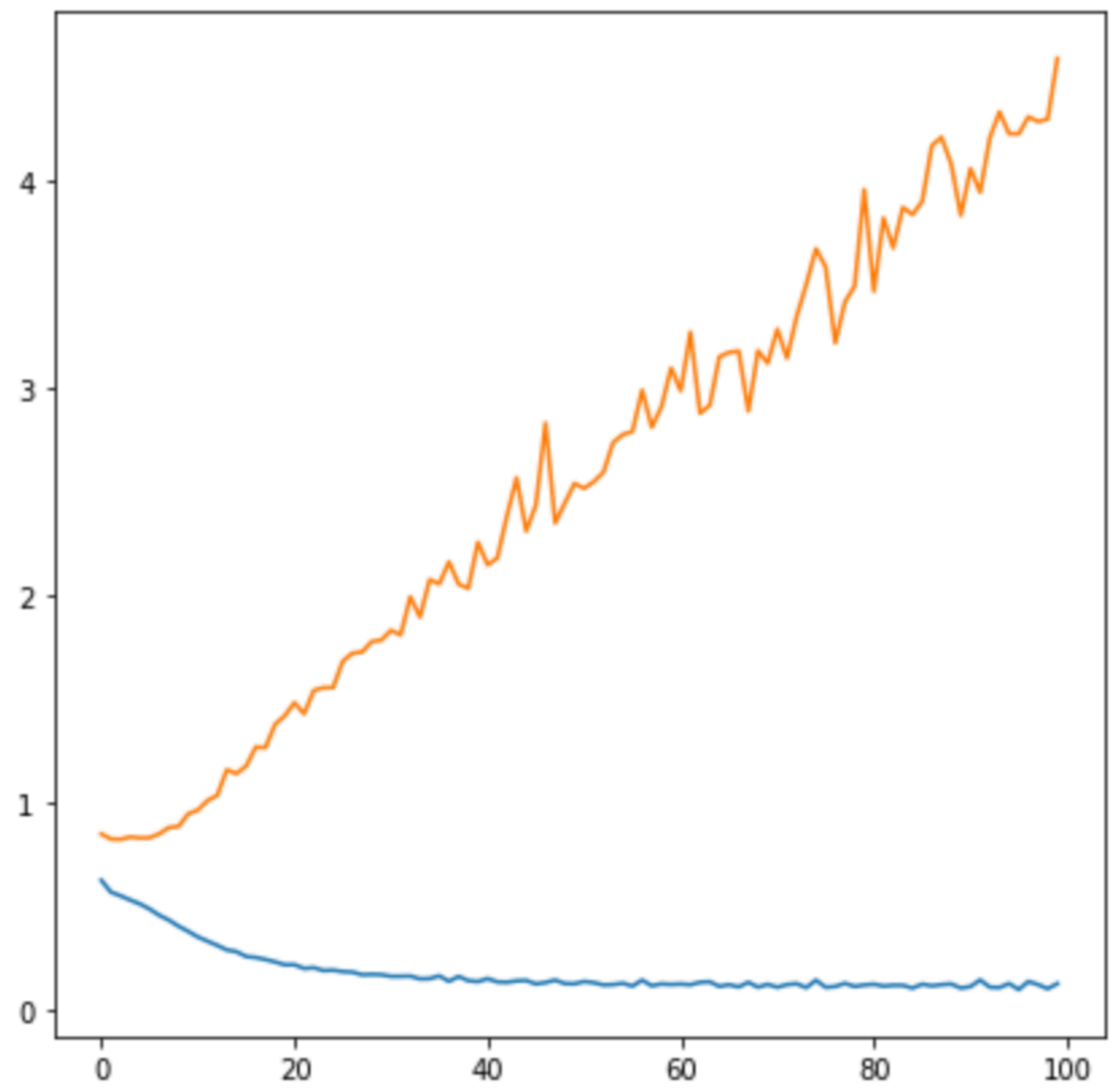
Training Data

Training Data with Adversarial Noise

Shuffle

Add Noise data to train model

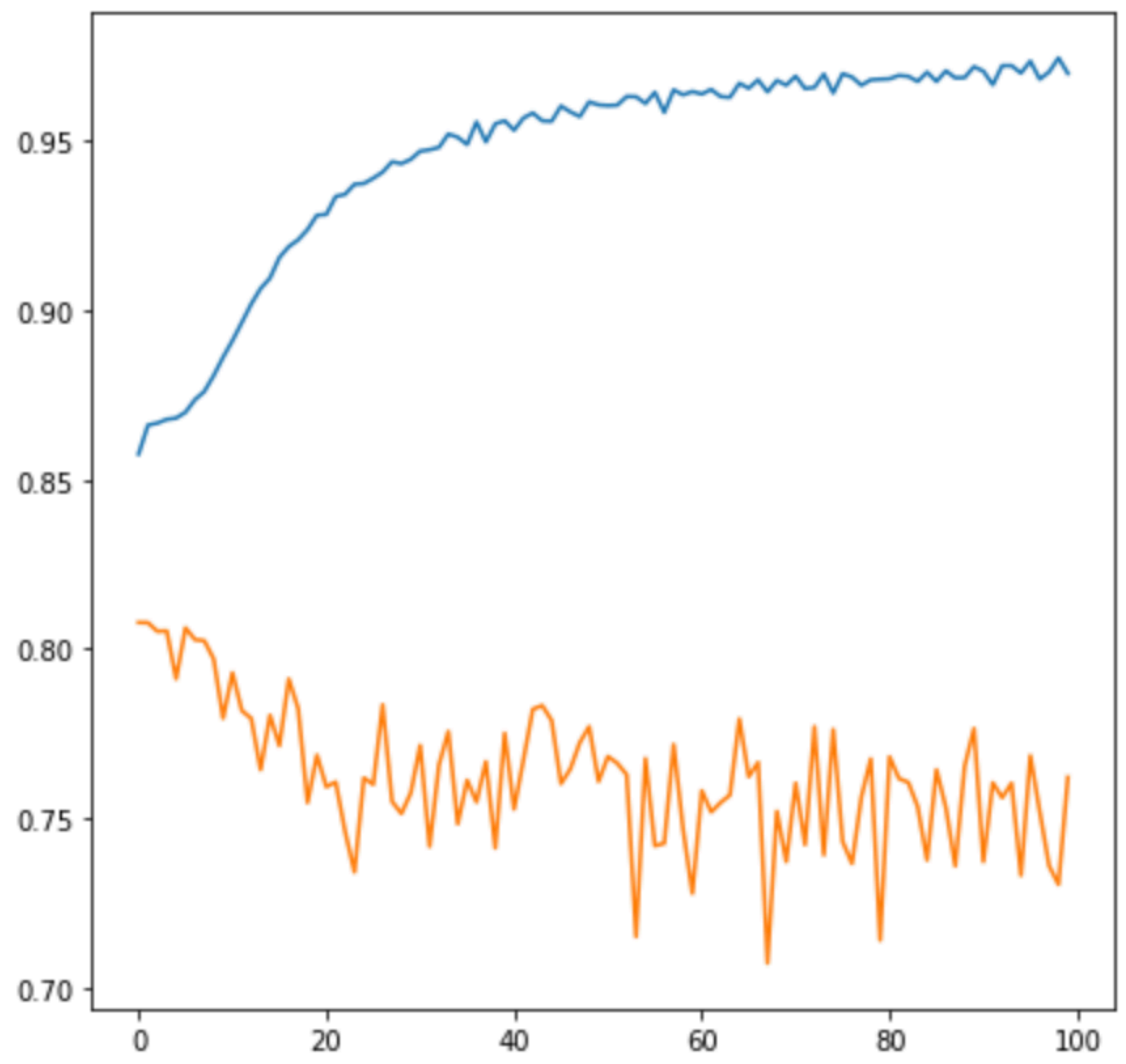
Loss



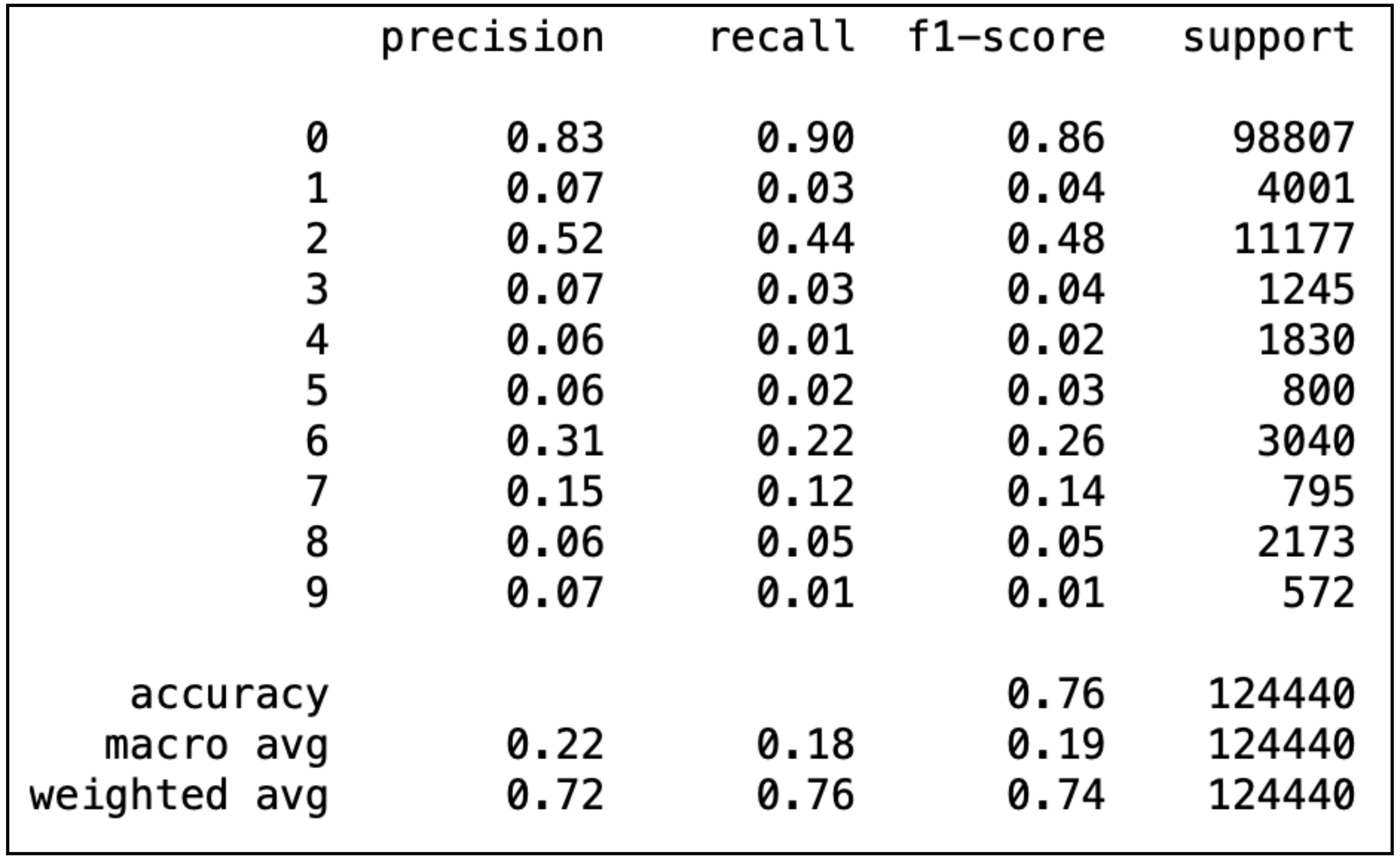
Training Data set

Testing Data set

Accuracy

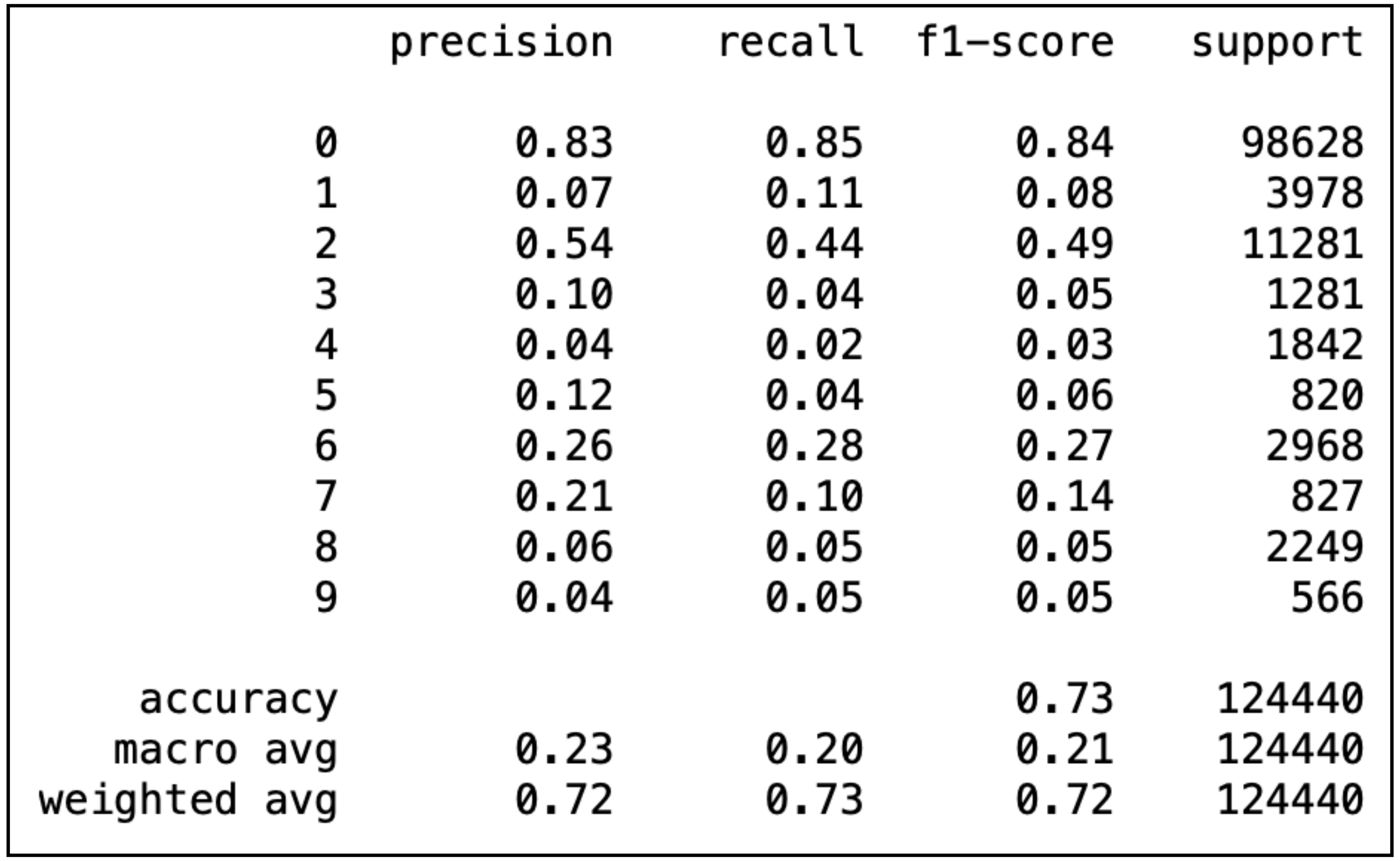


Add Noise data to train model



```
              precision    recall  f1-score   support

           0       0.83      0.90      0.86     98807
           1       0.07      0.03      0.04      4001
           2       0.52      0.44      0.48     11177
           3       0.07      0.03      0.04      1245
           4       0.06      0.01      0.02      1830
           5       0.06      0.02      0.03       800
           6       0.31      0.22      0.26      3040
           7       0.15      0.12      0.14       795
           8       0.06      0.05      0.05      2173
           9       0.07      0.01      0.01       572

    accuracy                           0.76    124440
   macro avg       0.22      0.18      0.19    124440
weighted avg       0.72      0.76      0.74    124440
```

Testing Data set



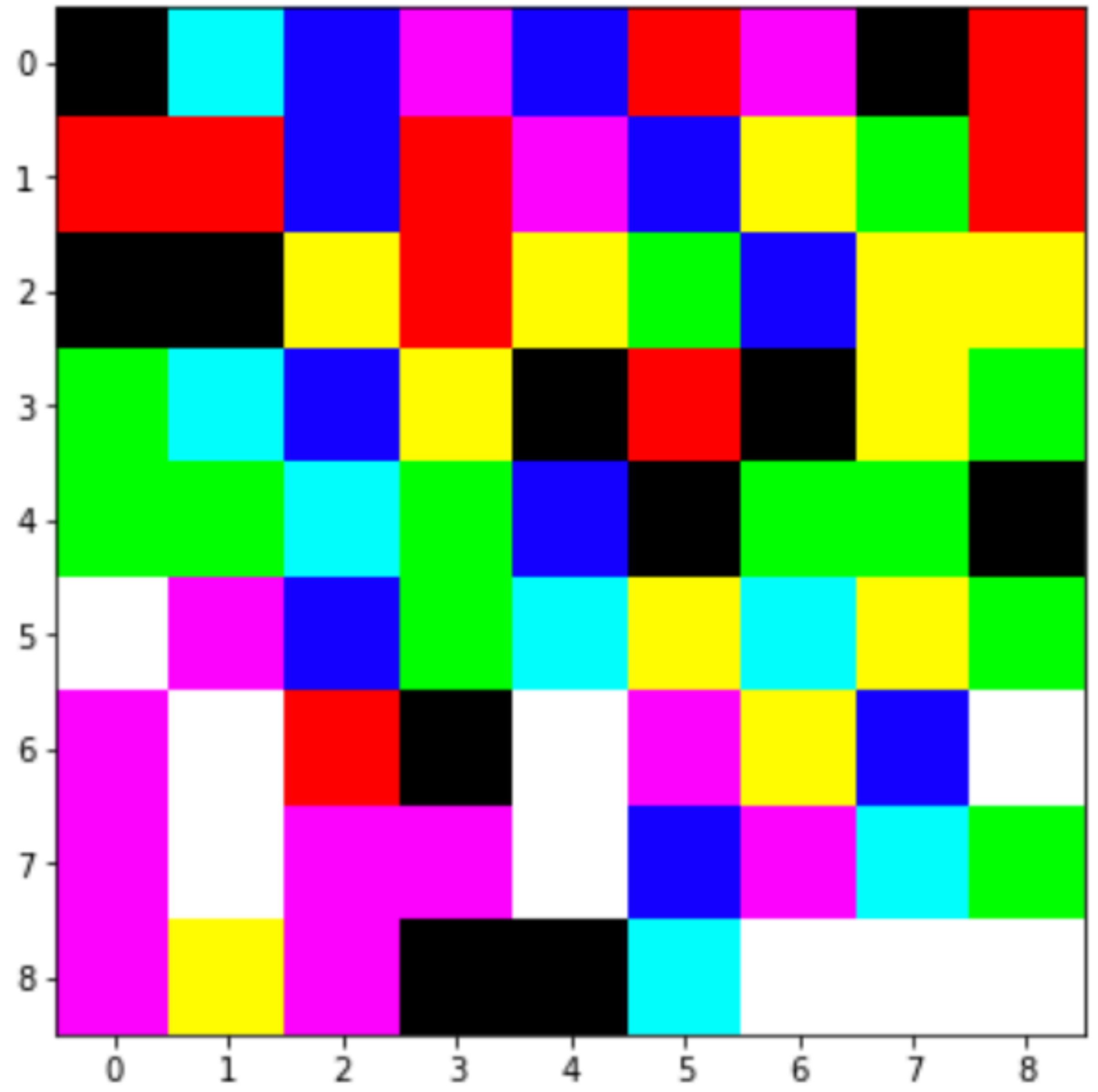
```
              precision    recall  f1-score   support

           0       0.83      0.85      0.84     98628
           1       0.07      0.11      0.08      3978
           2       0.54      0.44      0.49     11281
           3       0.10      0.04      0.05      1281
           4       0.04      0.02      0.03      1842
           5       0.12      0.04      0.06       820
           6       0.26      0.28      0.27      2968
           7       0.21      0.10      0.14       827
           8       0.06      0.05      0.05      2249
           9       0.04      0.05      0.05       566

    accuracy                           0.73    124440
   macro avg       0.23      0.20      0.21    124440
weighted avg       0.72      0.73      0.72    124440
```

Testing Data set
(Normal Model)

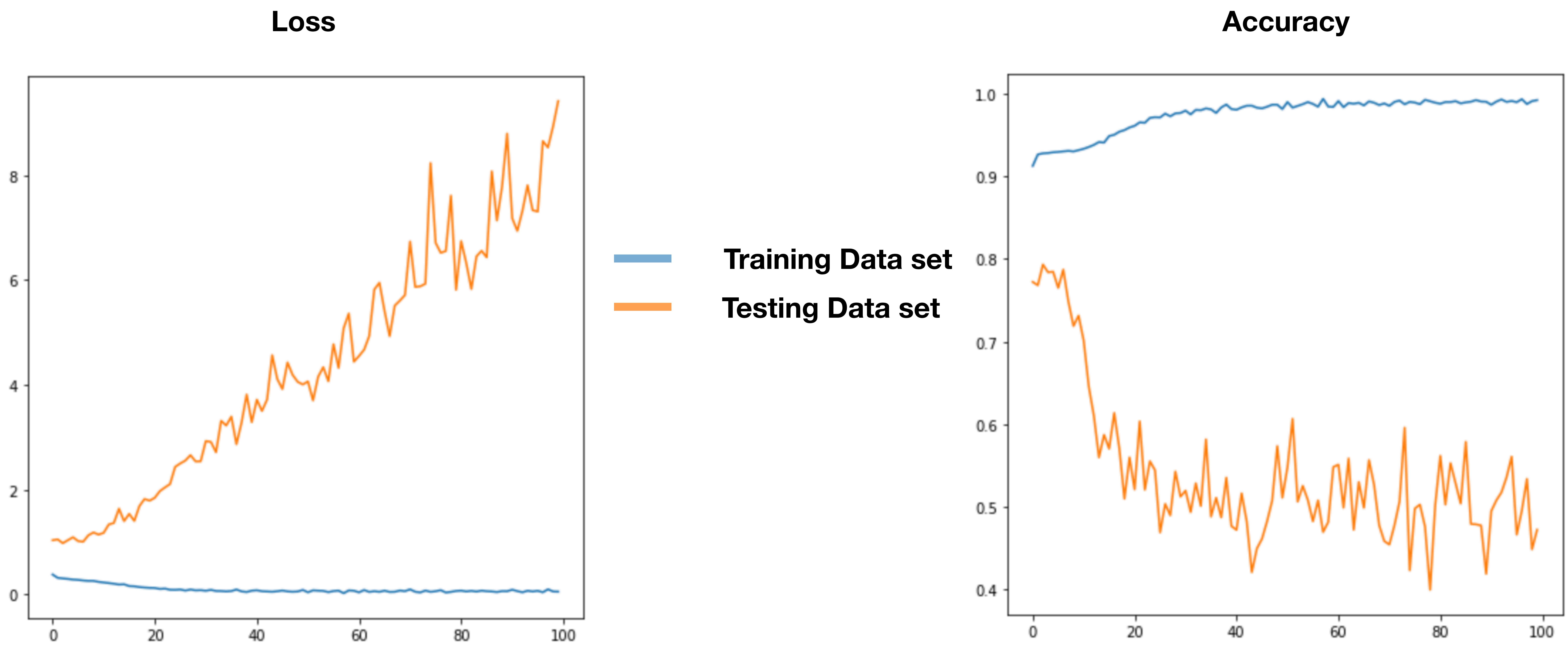
Train with noise data

Training Data

+ 0.01 *



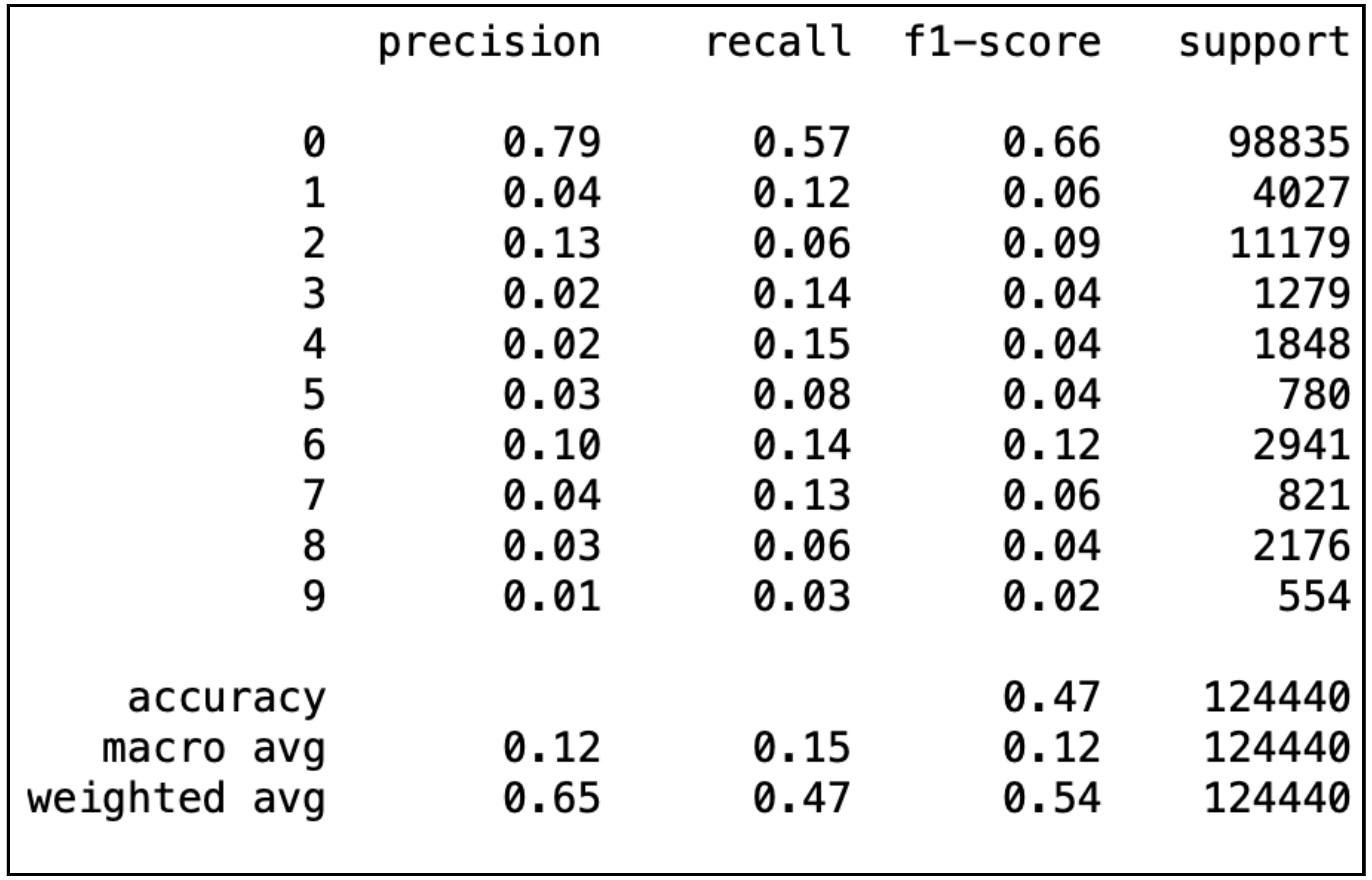
Train with noise data



Loss

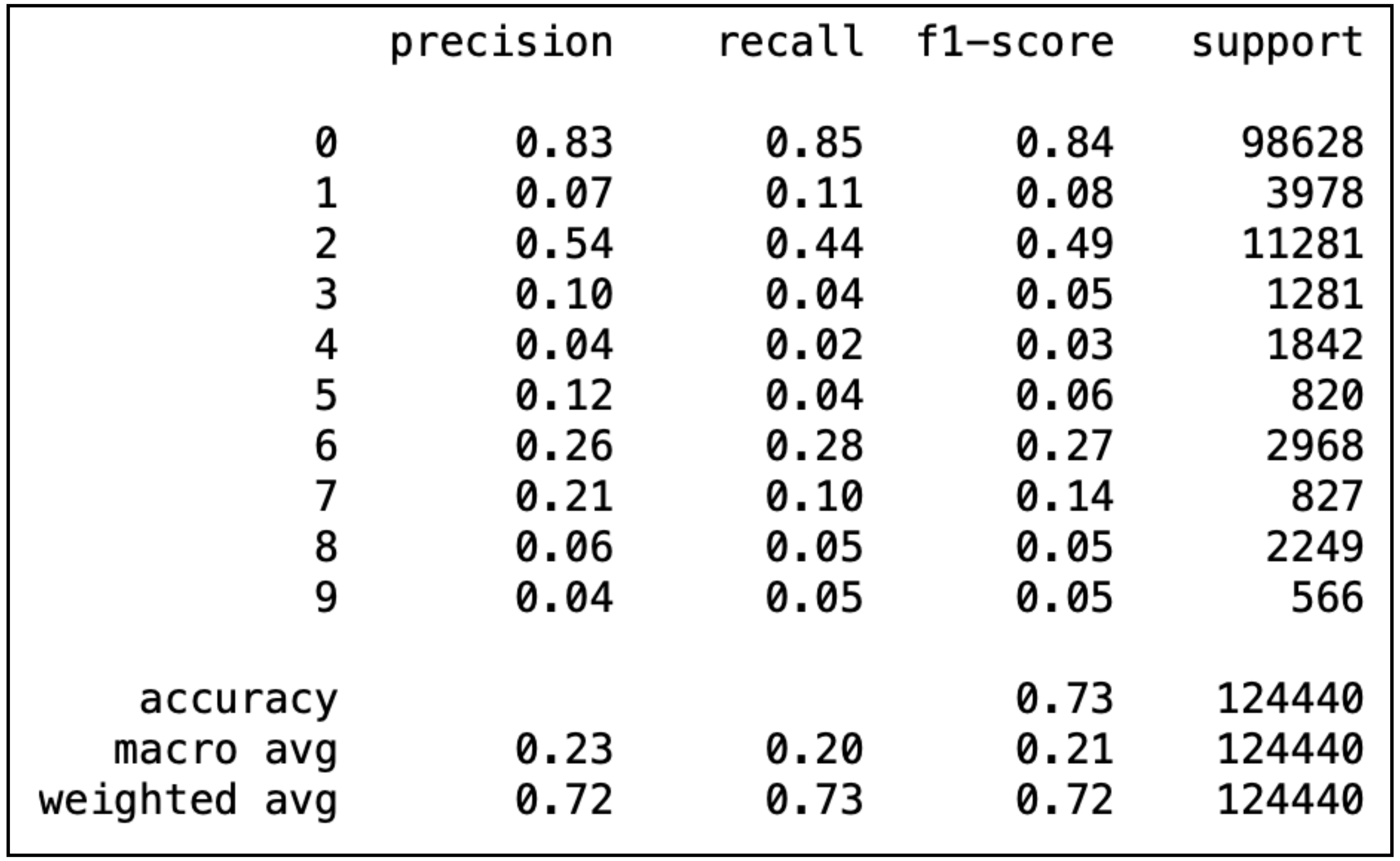
Accuracy

Train with noise data



```
              precision    recall  f1-score   support

           0       0.79      0.57      0.66     98835
           1       0.04      0.12      0.06      4027
           2       0.13      0.06      0.09     11179
           3       0.02      0.14      0.04      1279
           4       0.02      0.15      0.04      1848
           5       0.03      0.08      0.04       780
           6       0.10      0.14      0.12      2941
           7       0.04      0.13      0.06       821
           8       0.03      0.06      0.04      2176
           9       0.01      0.03      0.02       554

    accuracy                           0.47    124440
   macro avg       0.12      0.15      0.12    124440
weighted avg       0.65      0.47      0.54    124440
```

Testing Data set



```
              precision    recall  f1-score   support

           0       0.83      0.85      0.84     98628
           1       0.07      0.11      0.08      3978
           2       0.54      0.44      0.49     11281
           3       0.10      0.04      0.05      1281
           4       0.04      0.02      0.03      1842
           5       0.12      0.04      0.06       820
           6       0.26      0.28      0.27      2968
           7       0.21      0.10      0.14       827
           8       0.06      0.05      0.05      2249
           9       0.04      0.05      0.05       566

    accuracy                           0.73    124440
   macro avg       0.23      0.20      0.21    124440
weighted avg       0.72      0.73      0.72    124440
```

Testing Data set
(Normal Model)

End