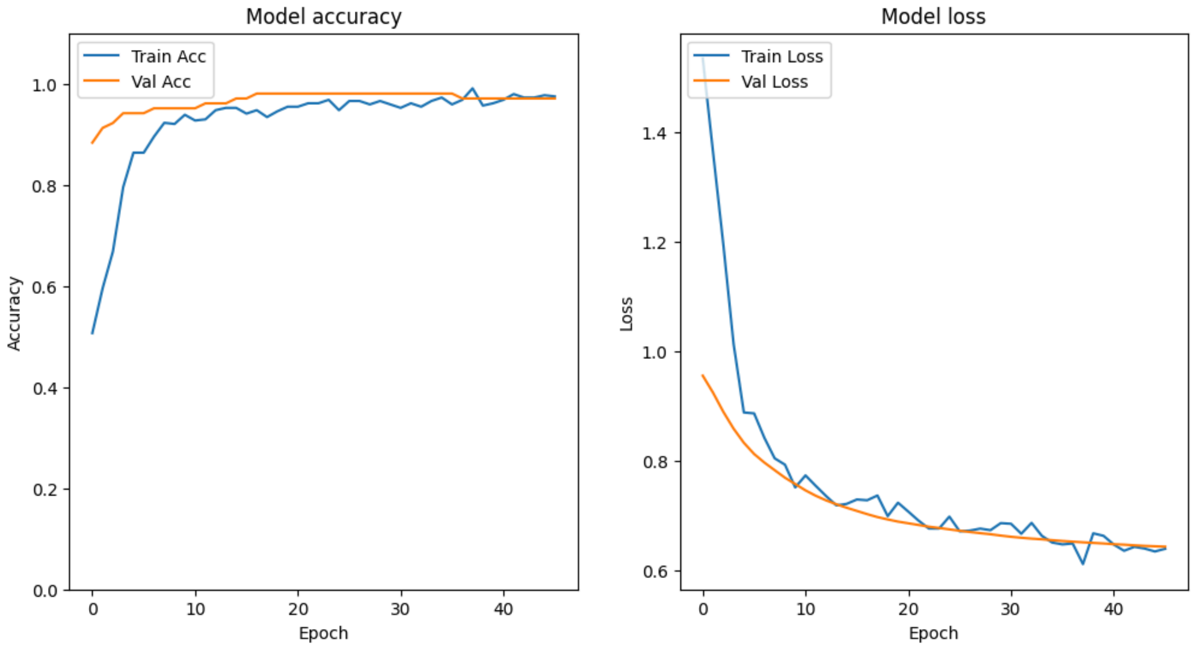
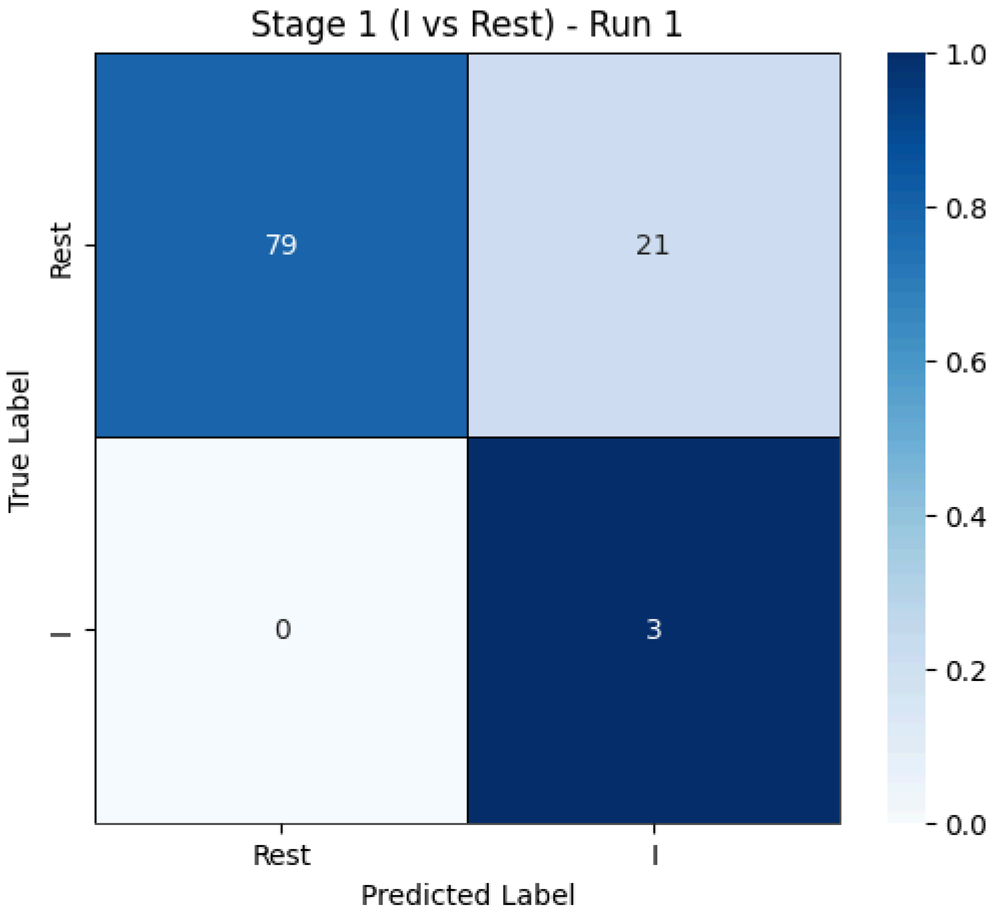
```
HIERARCHICAL
==============================================================
RUN 1: Using Split A for validation, Split B for testing
==============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```

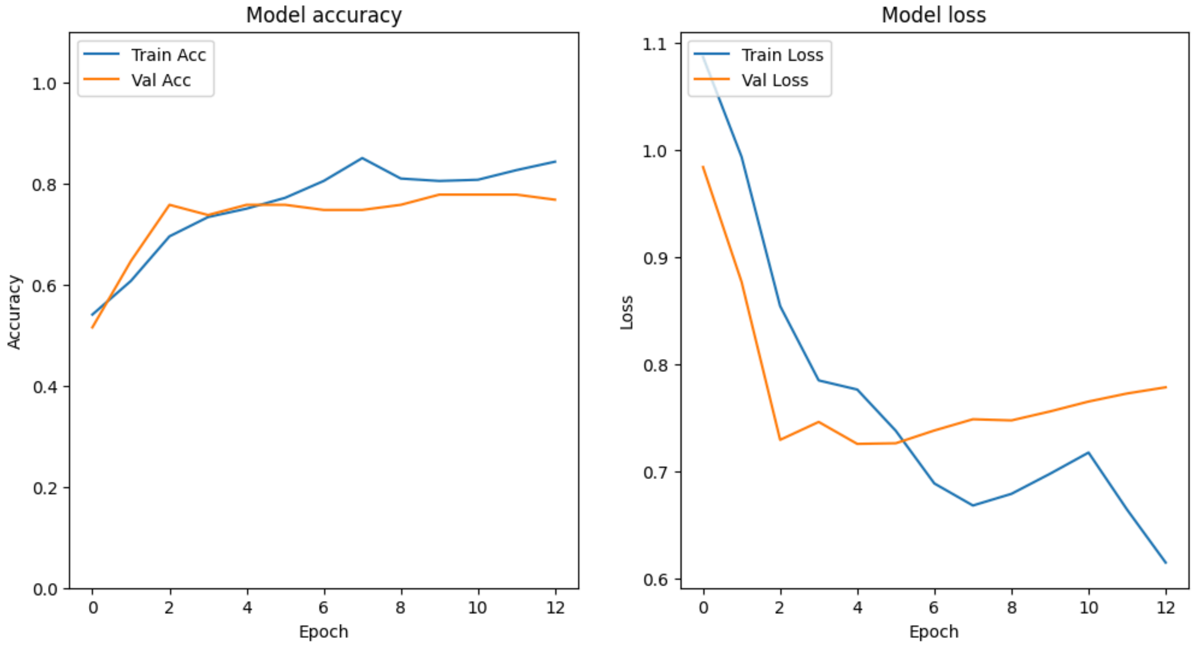


```
[Threshold Optimization] Best balanced_accuracy: 0.8889 at threshold=0.030
Optimal threshold (Stage 1 (I vs Rest)): 0.030
              precision    recall  f1-score   support

           0     1.0000    0.7900    0.8827       100
           1     0.1250    1.0000    0.2222         3

    accuracy                         0.7961       103
   macro avg     0.5625    0.8950    0.5525       103
weighted avg     0.9745    0.7961    0.8634       103

Balanced Accuracy: 0.895
```



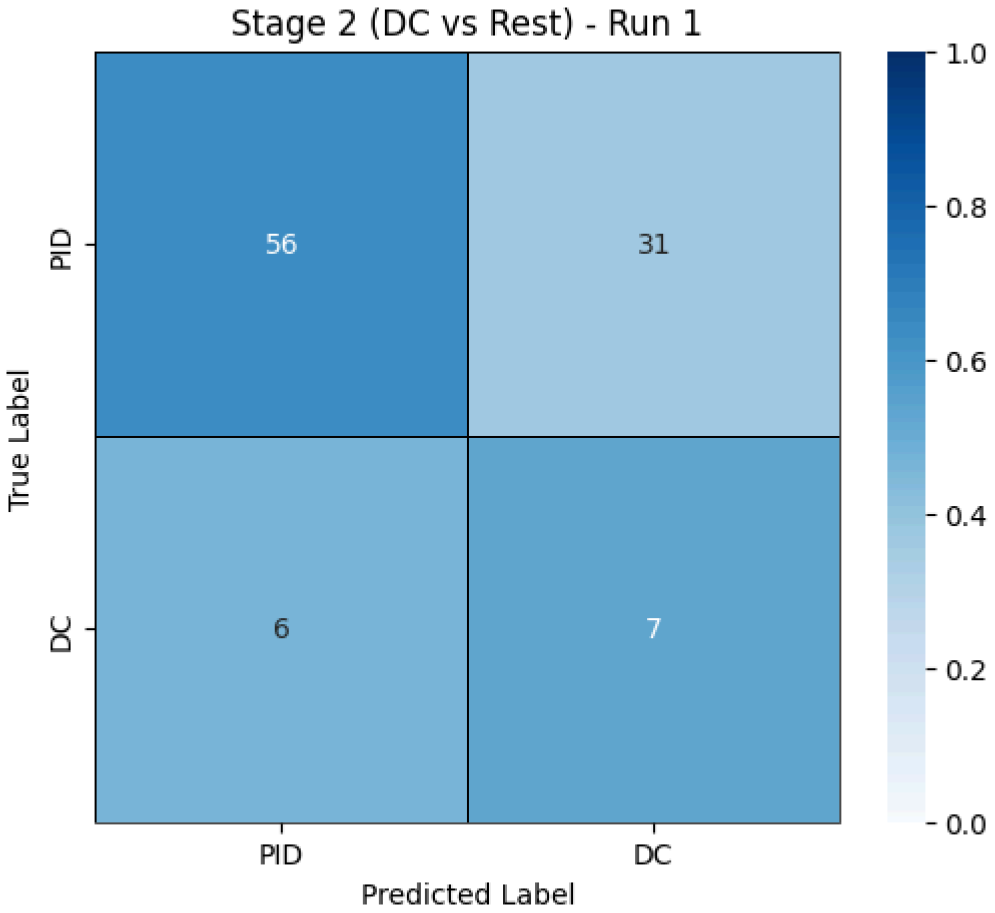
Stage 1 (I vs Rest) - Run 1

```
-- Stage 2 (DC vs Rest) --
```

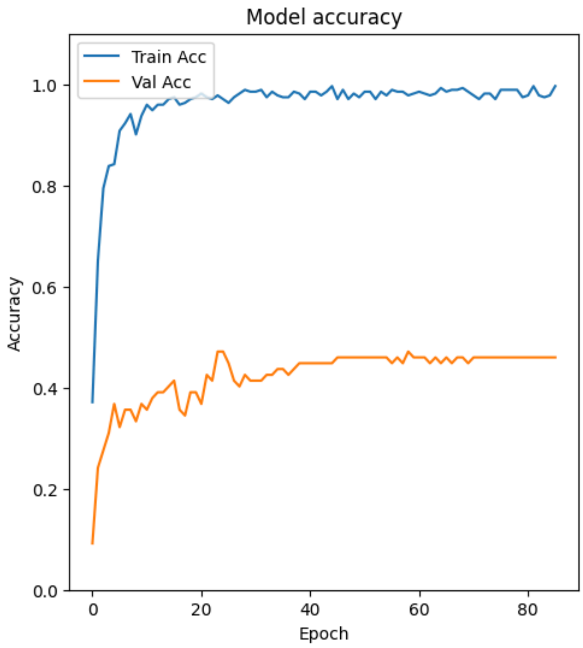


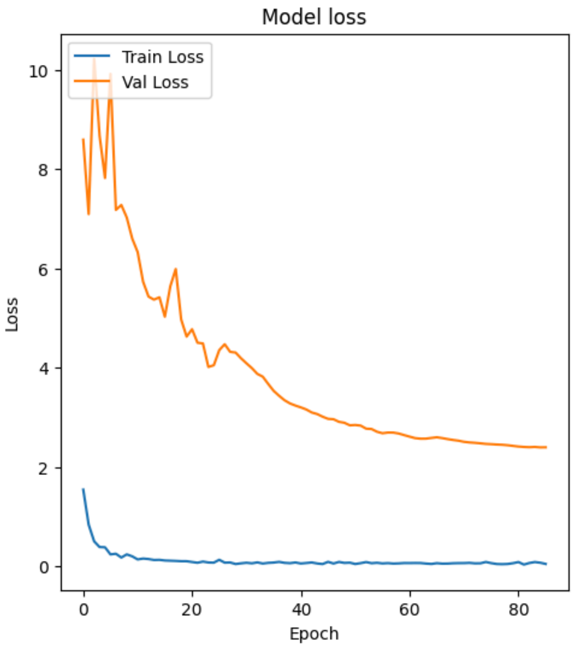
```
[Threshold Optimization] Best balanced_accuracy: 0.6451 at threshold=0.126
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

```
           0     0.9032    0.6437    0.7517        87
           1     0.1842    0.5385    0.2745        13

    accuracy                         0.6300       100
   macro avg     0.5437    0.5911    0.5131       100
weighted avg     0.8098    0.6300    0.6896       100

Balanced Accuracy: 0.5910698496905393
```

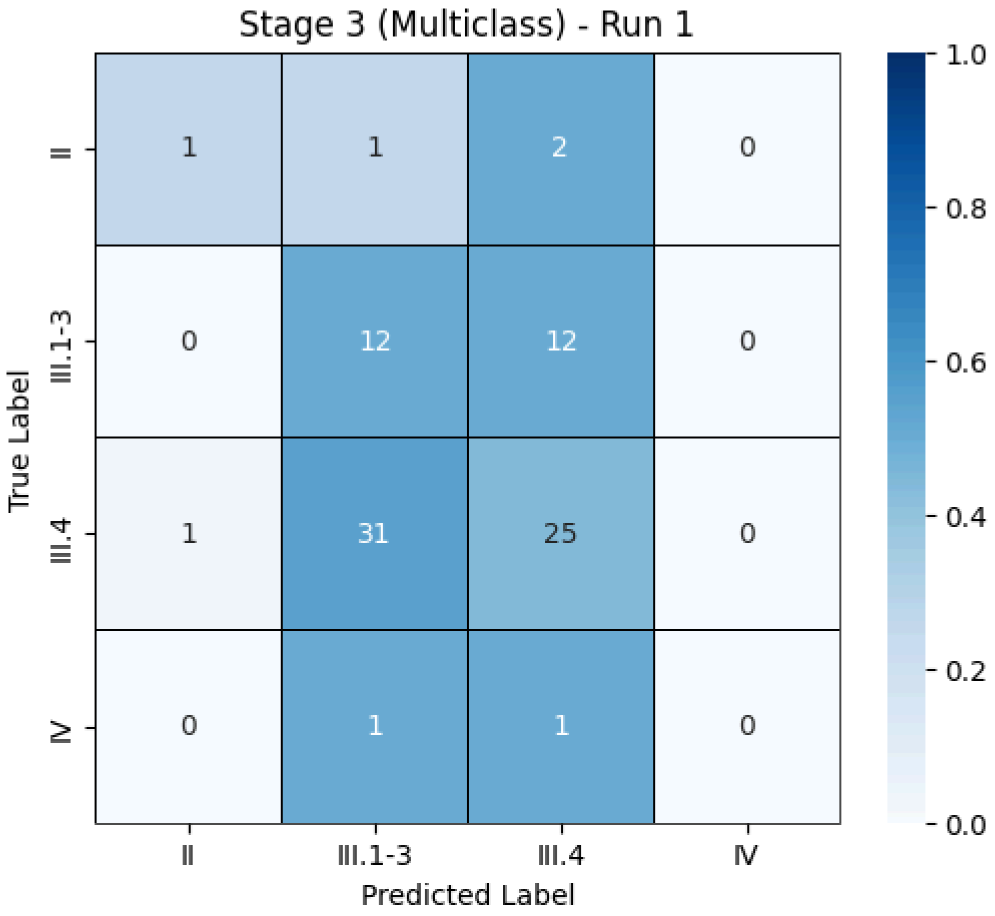


```
-- Stage 3 (Multiclass) --
```

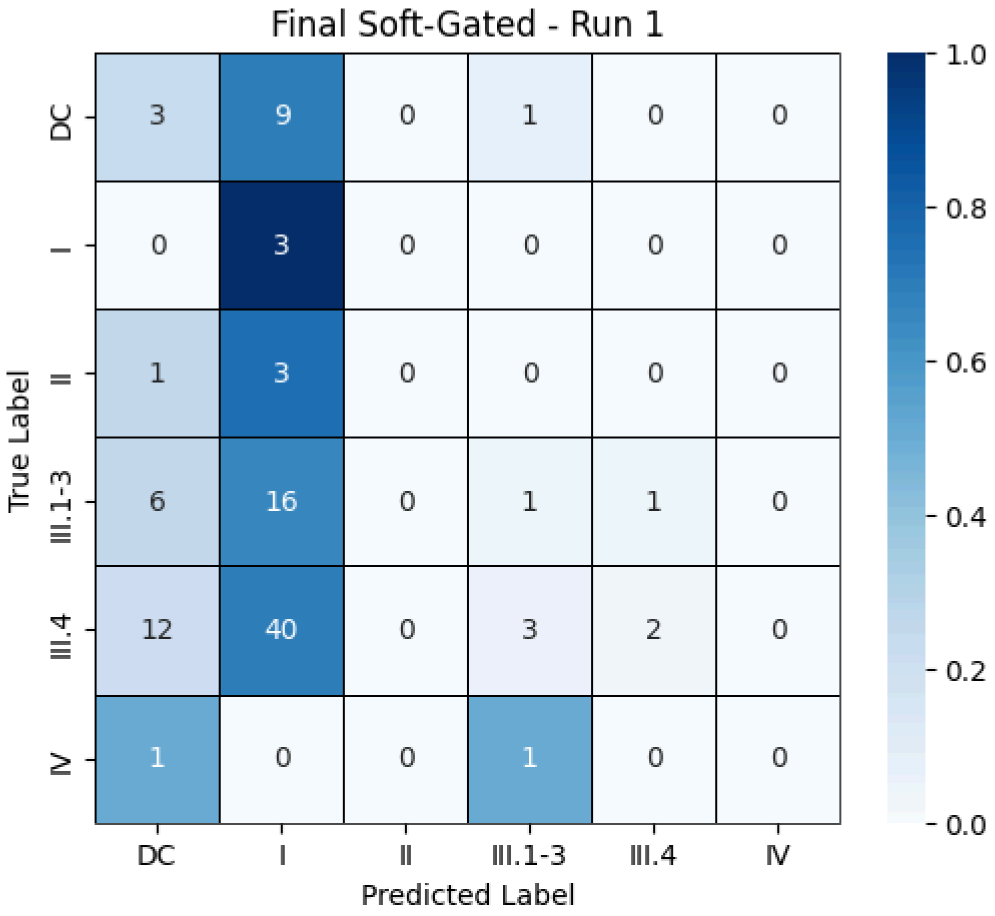




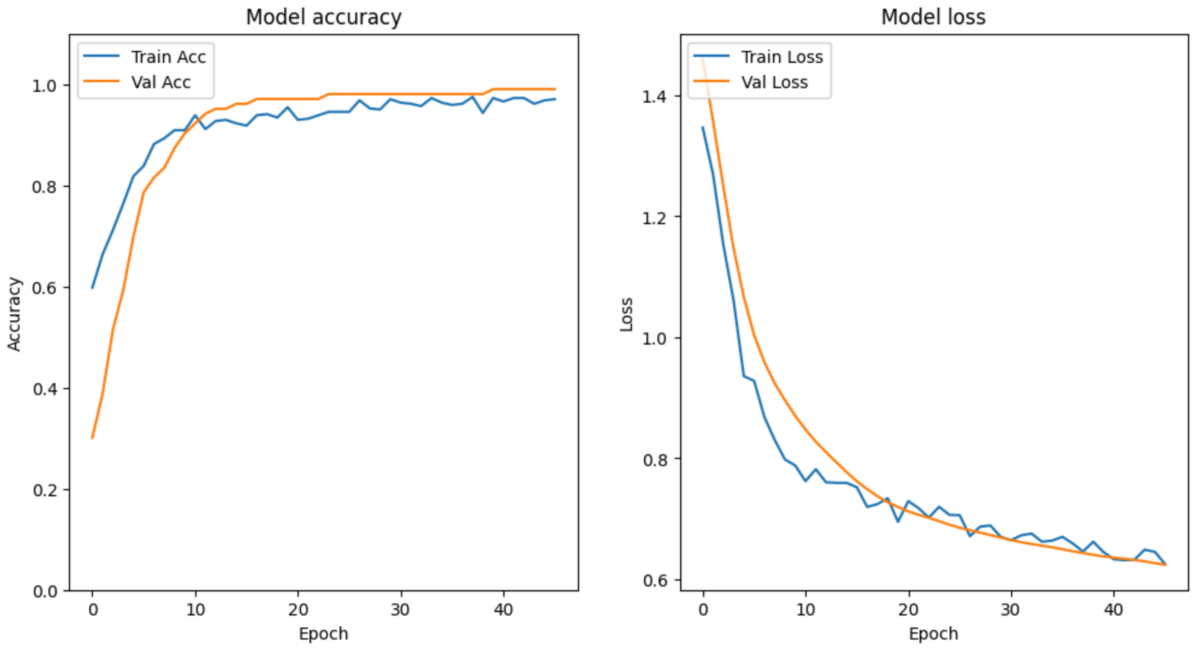
```
              precision    recall  f1-score   support

           0     0.5000    0.2500    0.3333         4
           1     0.2667    0.5000    0.3478        24
           2     0.6250    0.4386    0.5155        57
           3     0.0000    0.0000    0.0000         2

    accuracy                         0.4368        87
   macro avg     0.3479    0.2971    0.2992        87
weighted avg     0.5060    0.4368    0.4490        87

Balanced Accuracy: 0.29714912280701755
```



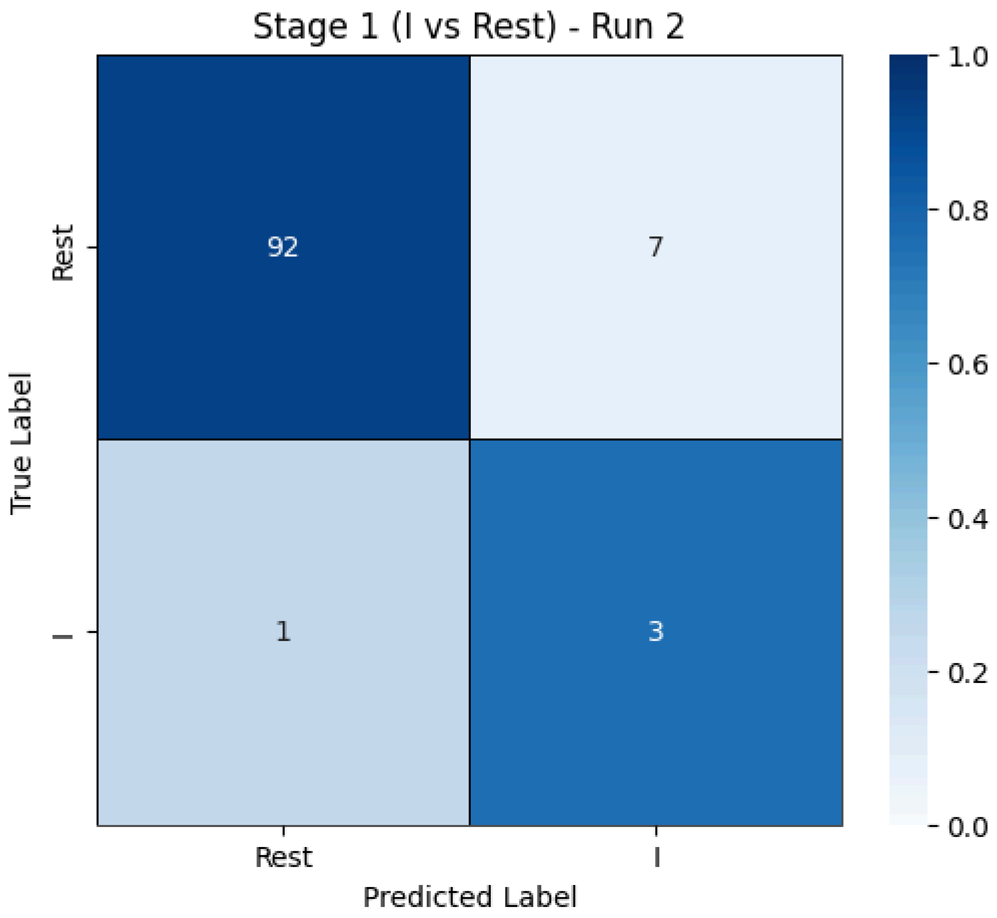
```
== Soft-Gated Overall (Test Set) - Run 1 ==
              precision    recall  f1-score   support

          DC     0.1304    0.2308    0.1667        13
           I     0.0423    1.0000    0.0811         3
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.1667    0.0417    0.0667        24
       III.4     0.6667    0.0351    0.0667        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.0874       103
   macro avg     0.1677    0.2179    0.0635       103
weighted avg     0.4255    0.0874    0.0758       103

Balanced Accuracy: 0.21792060278902384
```



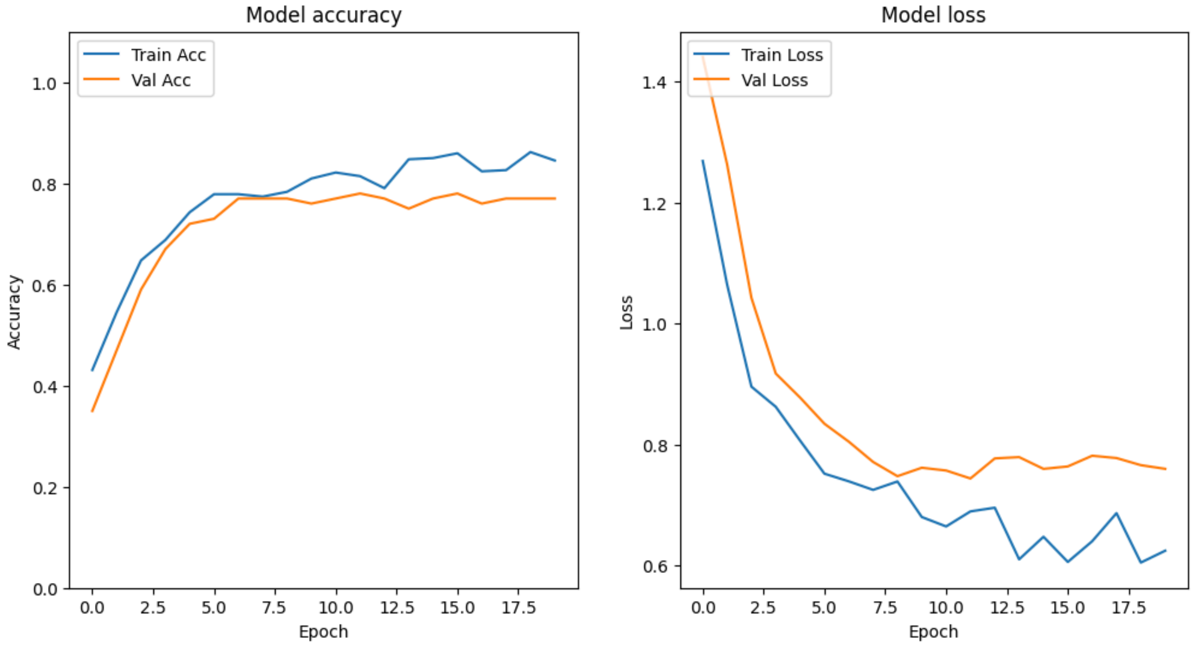
```
============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.9950 at threshold=0.392
Optimal threshold (Stage 1 (I vs Rest)): 0.392
              precision    recall  f1-score   support

           0     0.9892    0.9293    0.9583        99
           1     0.3000    0.7500    0.4286         4

    accuracy                         0.9223       103
   macro avg     0.6446    0.8396    0.6935       103
weighted avg     0.9625    0.9223    0.9378       103

Balanced Accuracy: 0.8396464646464646
```

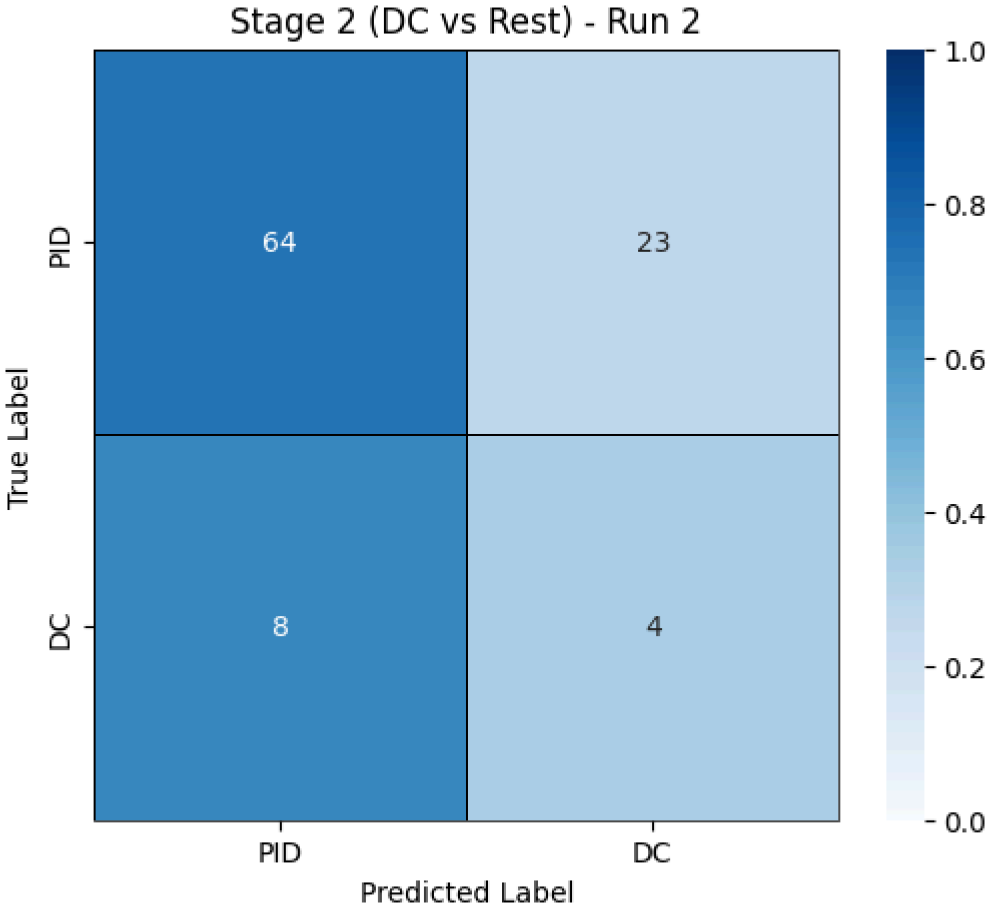


```
-- Stage 2 (DC vs Rest) --
```

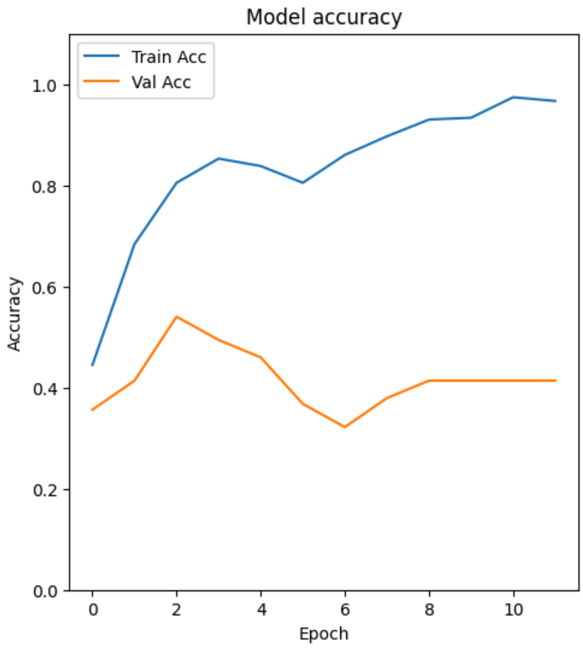


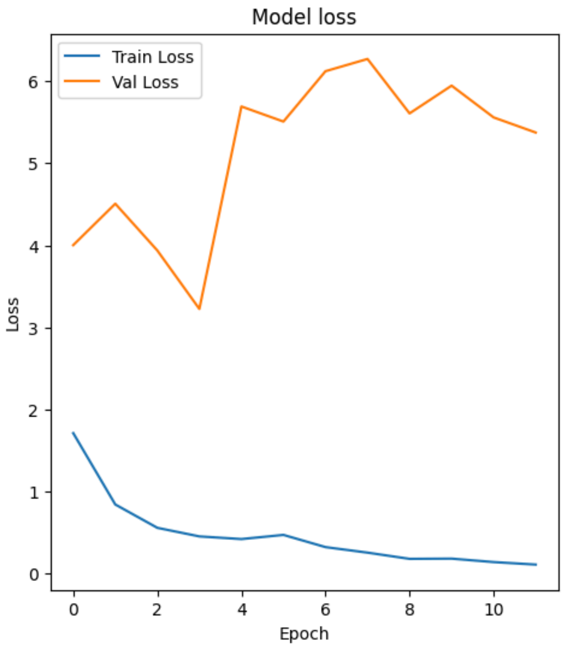
```
[Threshold Optimization] Best balanced_accuracy: 0.7679 at threshold=0.281
Optimal threshold (Stage 2 (DC vs Rest)): 0.281
              precision    recall  f1-score   support
```

```
           0     0.8889    0.7356    0.8050        87
           1     0.1481    0.3333    0.2051        12

    accuracy                         0.6869        99
   macro avg     0.5185    0.5345    0.5051        99
weighted avg     0.7991    0.6869    0.7323        99

Balanced Accuracy: 0.5344827586206896
```

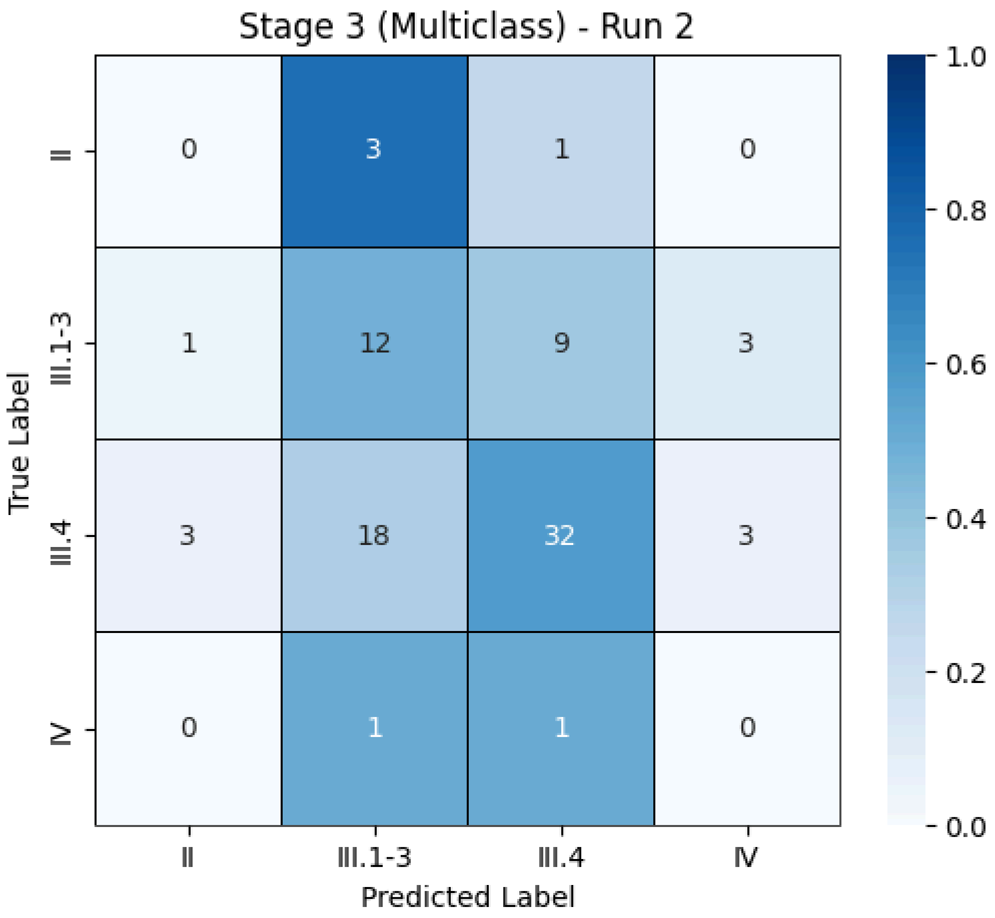


-- Stage 3 (Multiclass) --

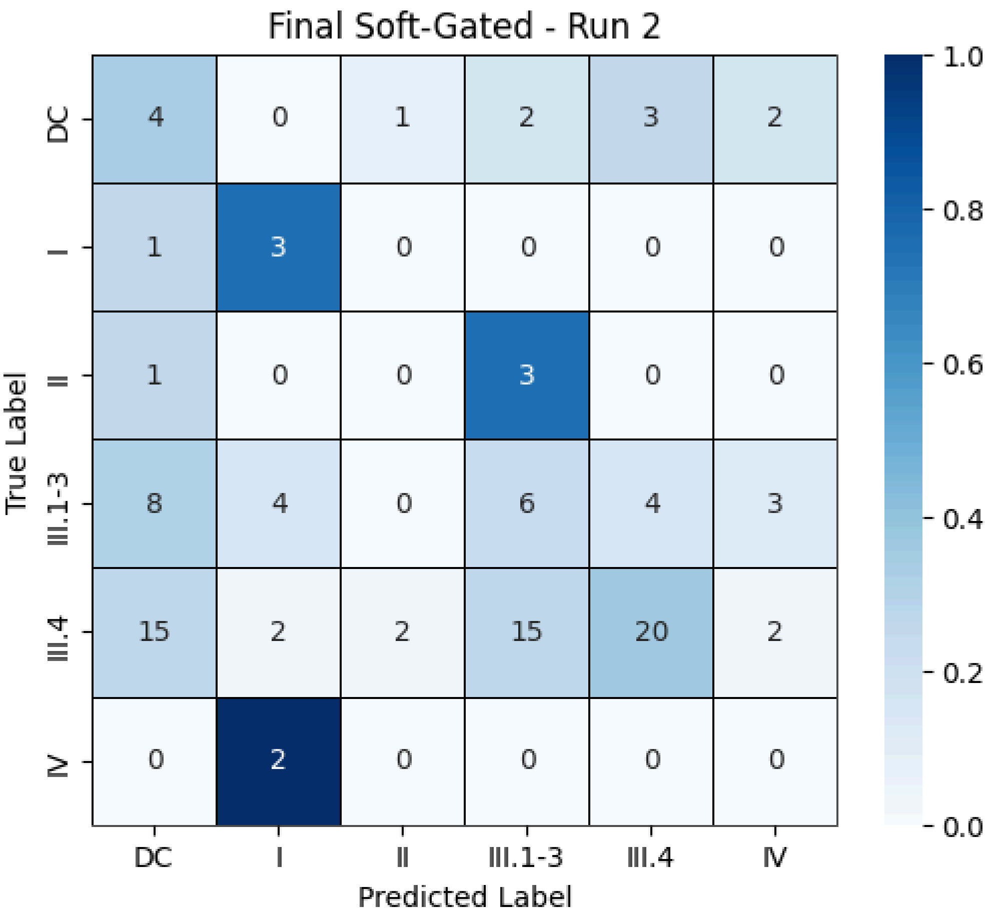




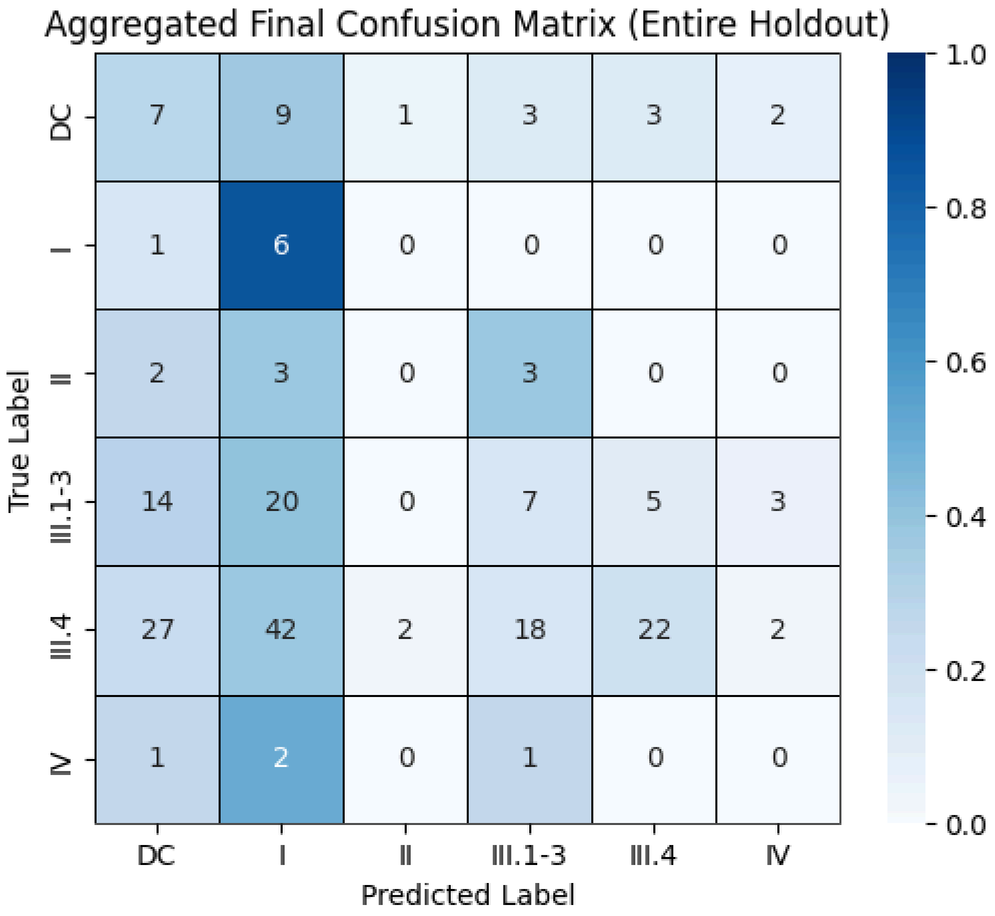
```
          precision    recall  f1-score   support

         0     0.0000    0.0000    0.0000         4
         1     0.3529    0.4800    0.4068        25
         2     0.7442    0.5714    0.6465        56
         3     0.0000    0.0000    0.0000         2

    accuracy                        0.5057        87
   macro avg     0.2743    0.2629    0.2633        87
weighted avg     0.5804    0.5057    0.5330        87

Balanced Accuracy: 0.26285714285714284
```



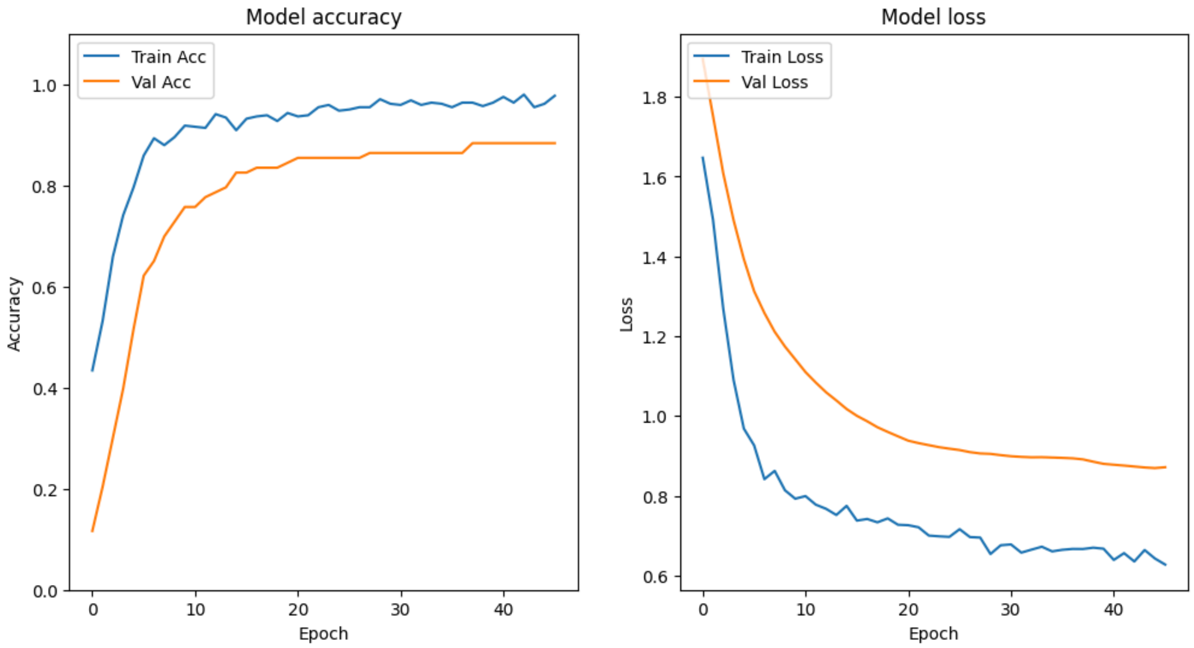
```
== Soft-Gated Overall (Test Set) - Run 2 ==
              precision    recall  f1-score   support

          DC     0.1379    0.3333    0.1951        12
           I     0.2727    0.7500    0.4000         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.2308    0.2400    0.2353        25
       III.4     0.7407    0.3571    0.4819        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.3204       103
   macro avg     0.2304    0.2801    0.2187       103
weighted avg     0.4854    0.3204    0.3574       103

Balanced Accuracy: 0.2800793650793651
```



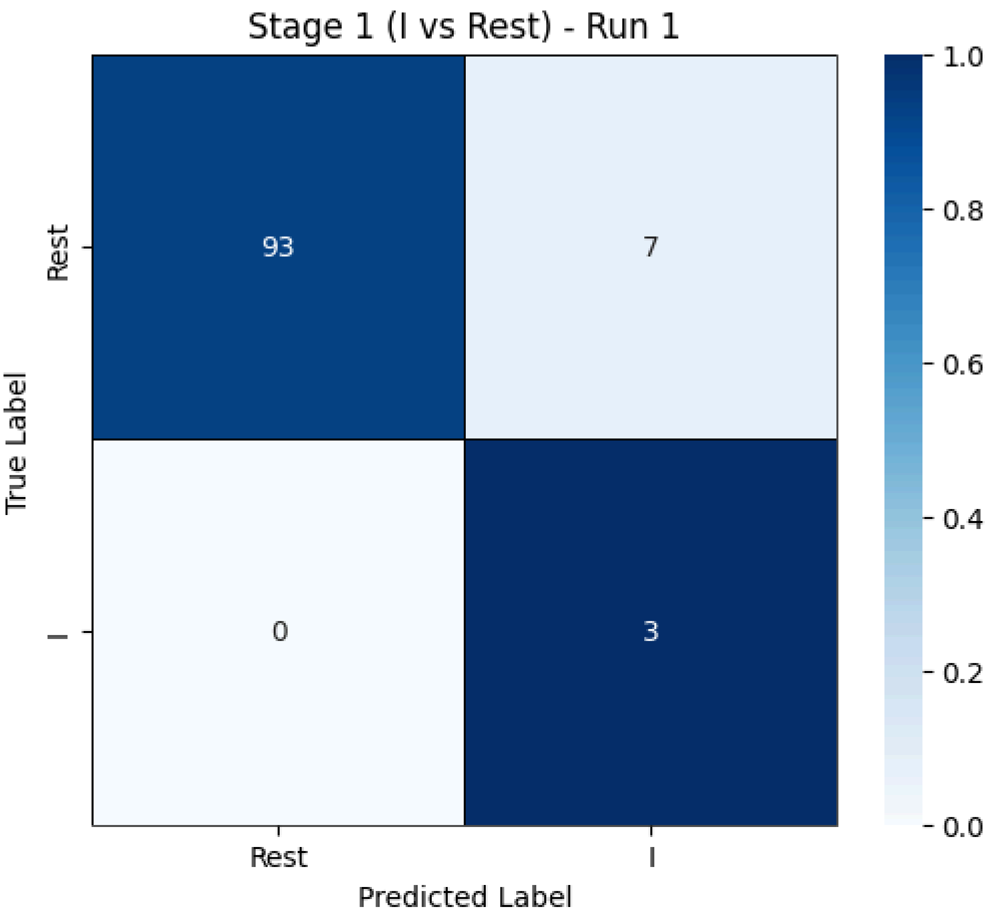
```
============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
============================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.1346    0.2800    0.1818        25
           I     0.0732    0.8571    0.1348         7
          II     0.0000    0.0000    0.0000         8
     III.1-3     0.2188    0.1429    0.1728        49
       III.4     0.7333    0.1947    0.3077       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.2039       206
   macro avg     0.1933    0.2458    0.1329       206
weighted avg     0.4731    0.2039    0.2365       206


Aggregated Balanced Accuracy: 0.2458
```



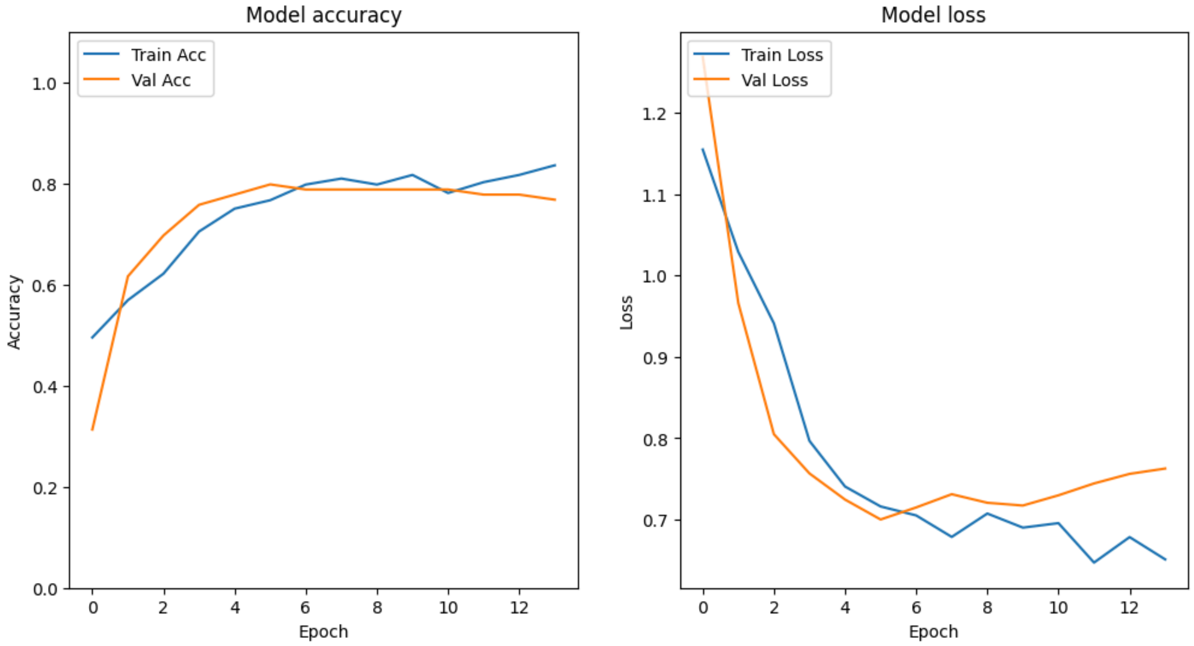
```
== Average Stage Balanced Accuracies ==
Stage 1 (I vs Rest): 0.8673
Stage 2 (DC vs Rest): 0.5628
Stage 3 (Multiclass): 0.2800
Final (Soft-Gated): 0.2490

============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.9444 at threshold=0.513
Optimal threshold (Stage 1 (I vs Rest)): 0.513
              precision    recall  f1-score   support

           0     1.0000    0.9300    0.9637       100
           1     0.3000    1.0000    0.4615         3

    accuracy                         0.9320       103
   macro avg     0.6500    0.9650    0.7126       103
weighted avg     0.9796    0.9320    0.9491       103

Balanced Accuracy: 0.9650000000000001
```

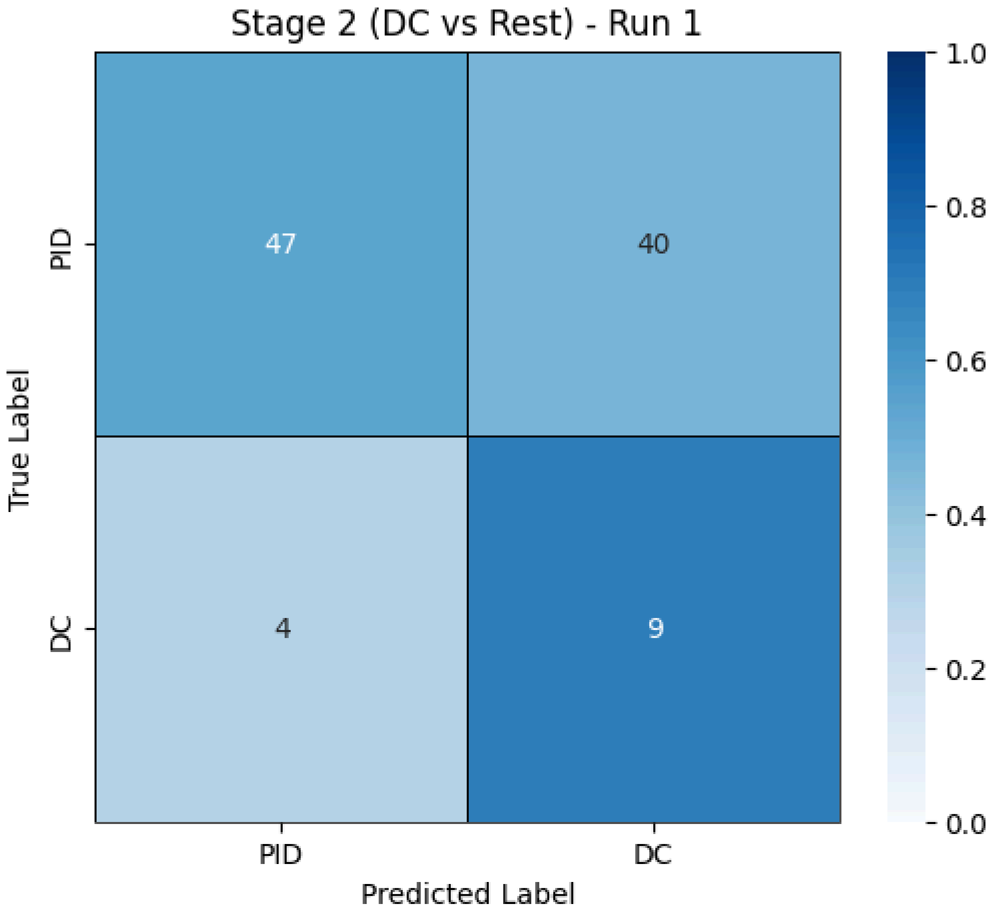


```
-- Stage 2 (DC vs Rest) --
```

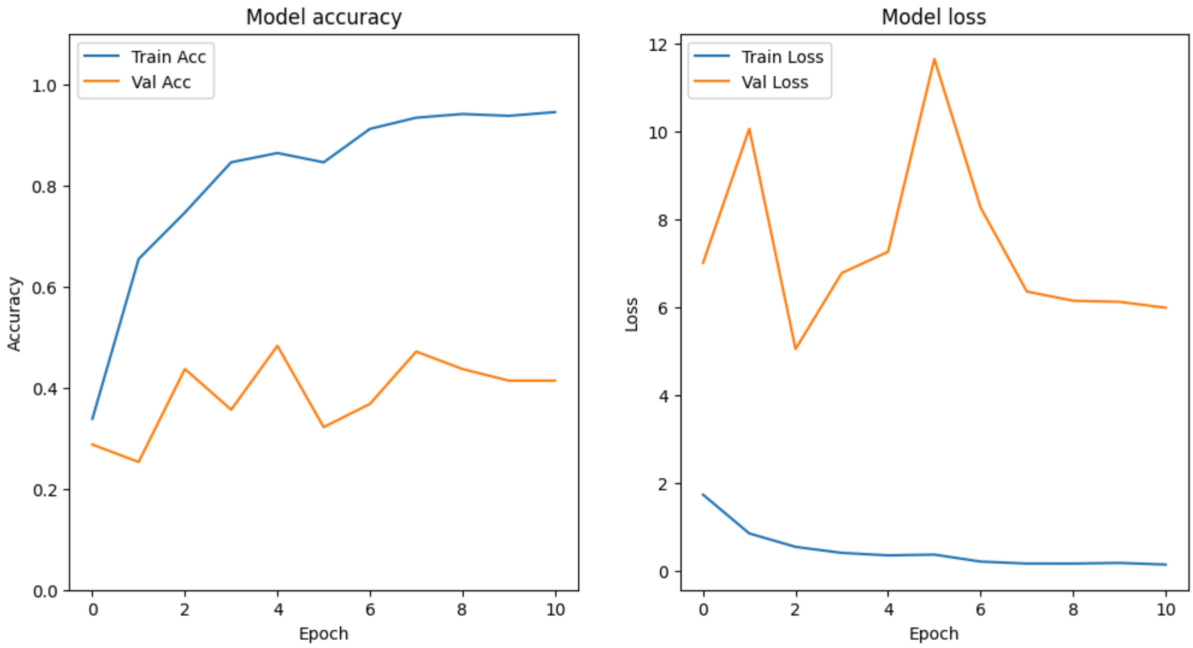


```
[Threshold Optimization] Best balanced_accuracy: 0.6422 at threshold=0.080
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

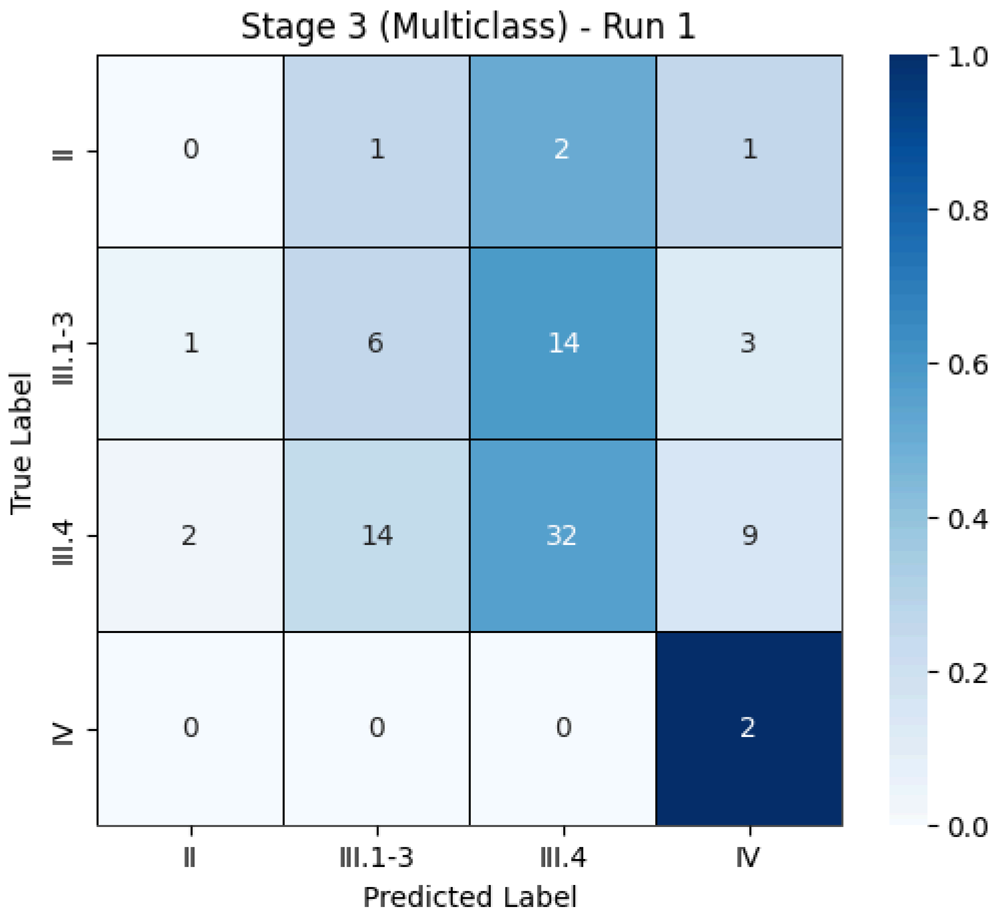
```
           0     0.9216    0.5402    0.6812        87
           1     0.1837    0.6923    0.2903        13

    accuracy                         0.5600       100
   macro avg     0.5526    0.6163    0.4857       100
weighted avg     0.8256    0.5600    0.6304       100

Balanced Accuracy: 0.6162687886825817
```



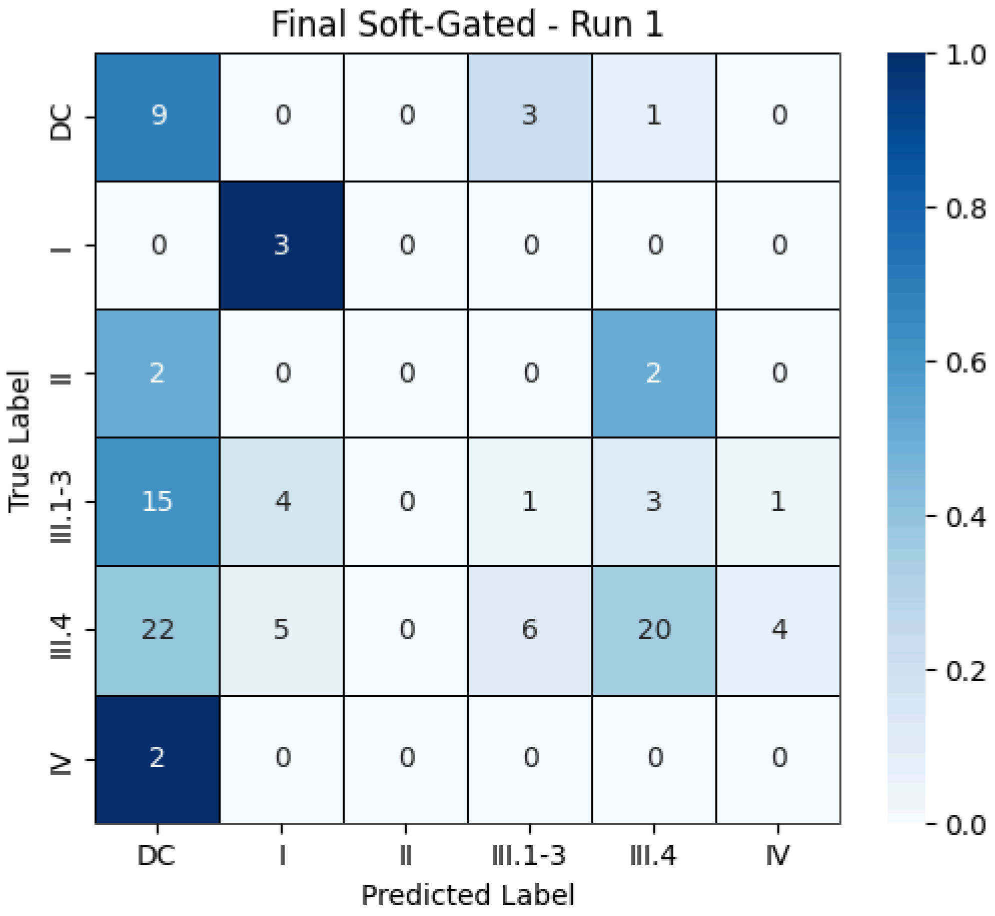
-- Stage 3 (Multiclass) --



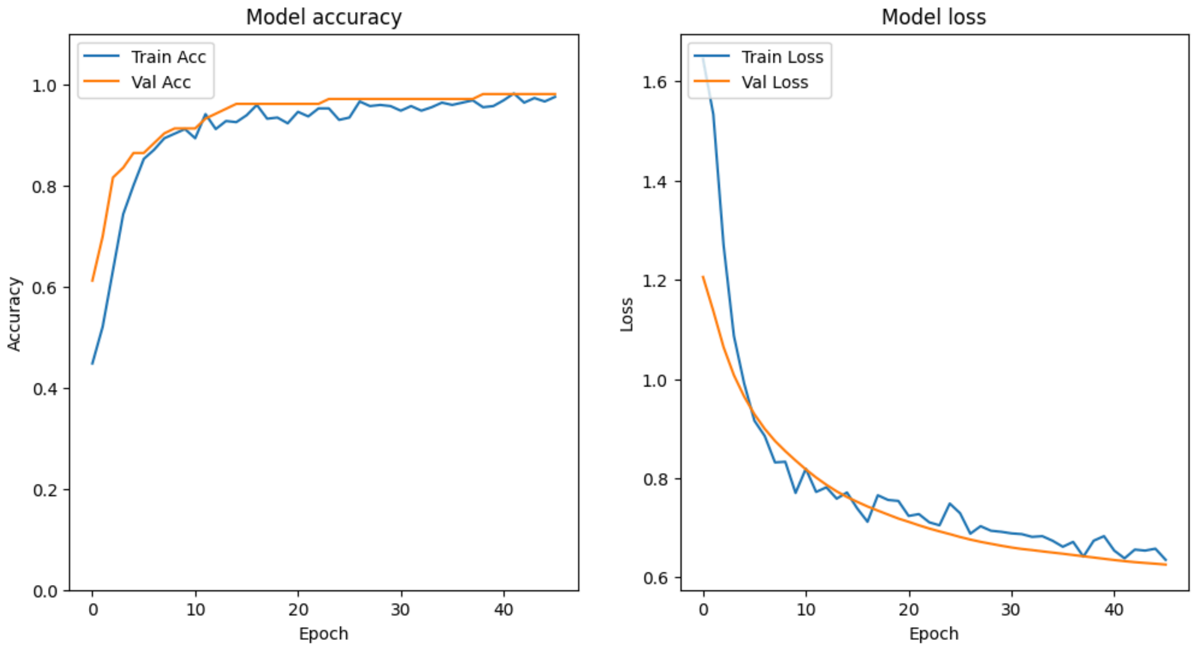
```
         precision    recall  f1-score   support

      0     0.0000    0.0000    0.0000         4
      1     0.2857    0.2500    0.2667        24
      2     0.6667    0.5614    0.6095        57
      3     0.1333    1.0000    0.2353         2

    accuracy                    0.4598        87
   macro avg     0.2714    0.4529    0.2779        87
weighted avg     0.5187    0.4598    0.4783        87

Balanced Accuracy: 0.45285087719298245
```

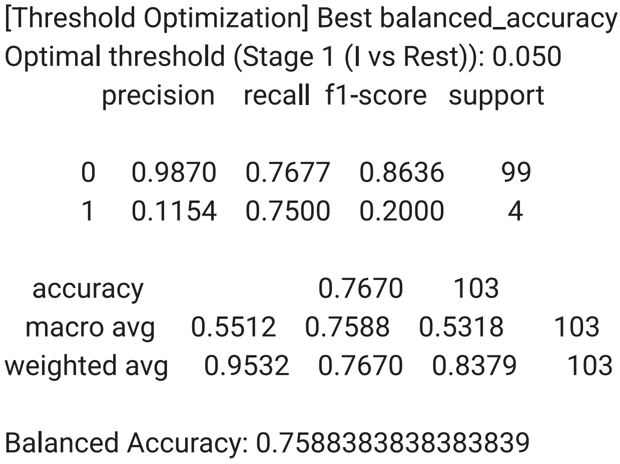


```
== Soft-Gated Overall (Test Set) - Run 1 ==
              precision    recall  f1-score   support

          DC     0.1800    0.6923    0.2857        13
           I     0.2500    1.0000    0.4000         3
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.1000    0.0417    0.0588        24
       III.4     0.7692    0.3509    0.4819        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.3204       103
   macro avg     0.2165    0.3475    0.2044       103
weighted avg     0.4790    0.3204    0.3281       103

Balanced Accuracy: 0.34747525865946916
```

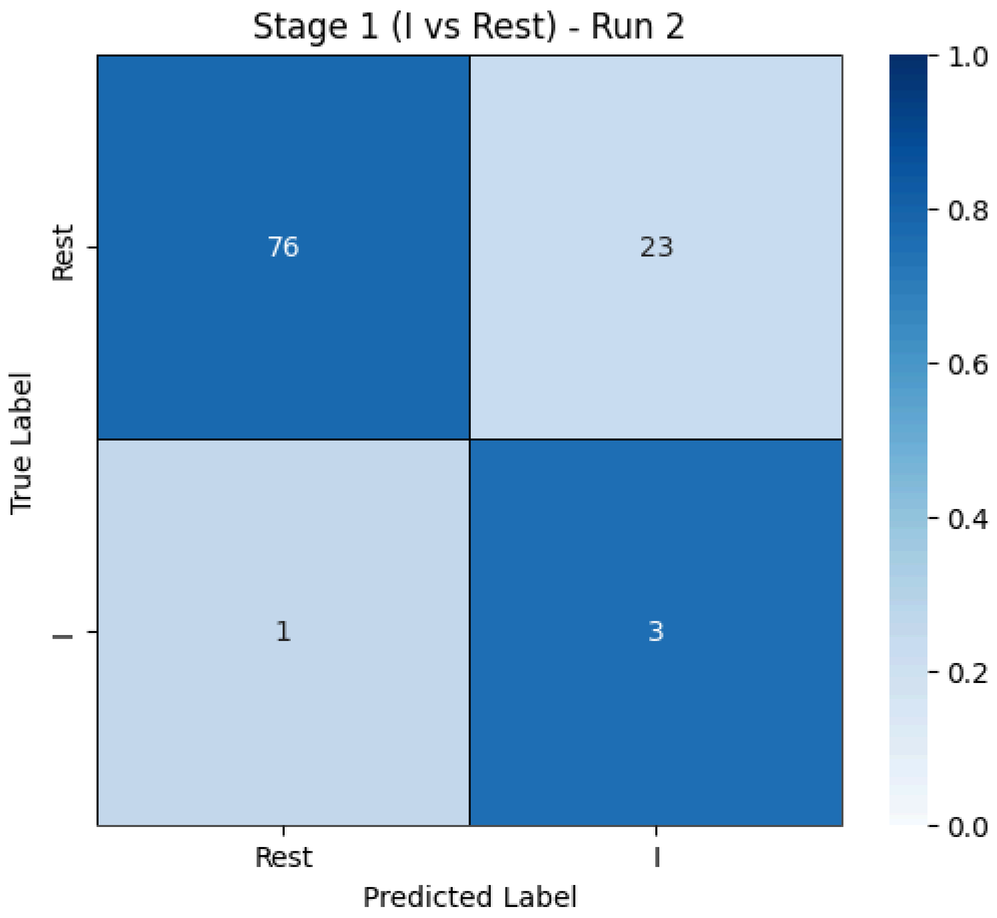


```
============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```

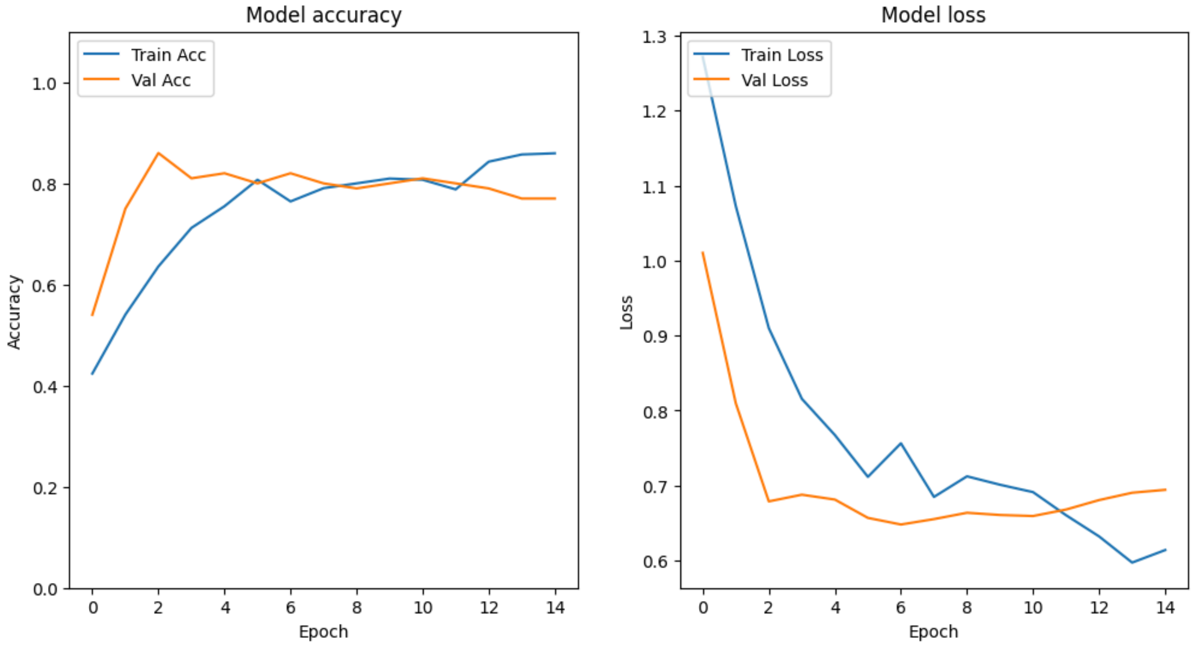




```
[Threshold Optimization] Best balanced_accuracy: 0.9050 at threshold=0.050
Optimal threshold (Stage 1 (I vs Rest)): 0.050
              precision    recall  f1-score   support

           0     0.9870    0.7677    0.8636        99
           1     0.1154    0.7500    0.2000         4

    accuracy                         0.7670       103
   macro avg     0.5512    0.7588    0.5318       103
weighted avg     0.9532    0.7670    0.8379       103

Balanced Accuracy: 0.7588383838383839
```

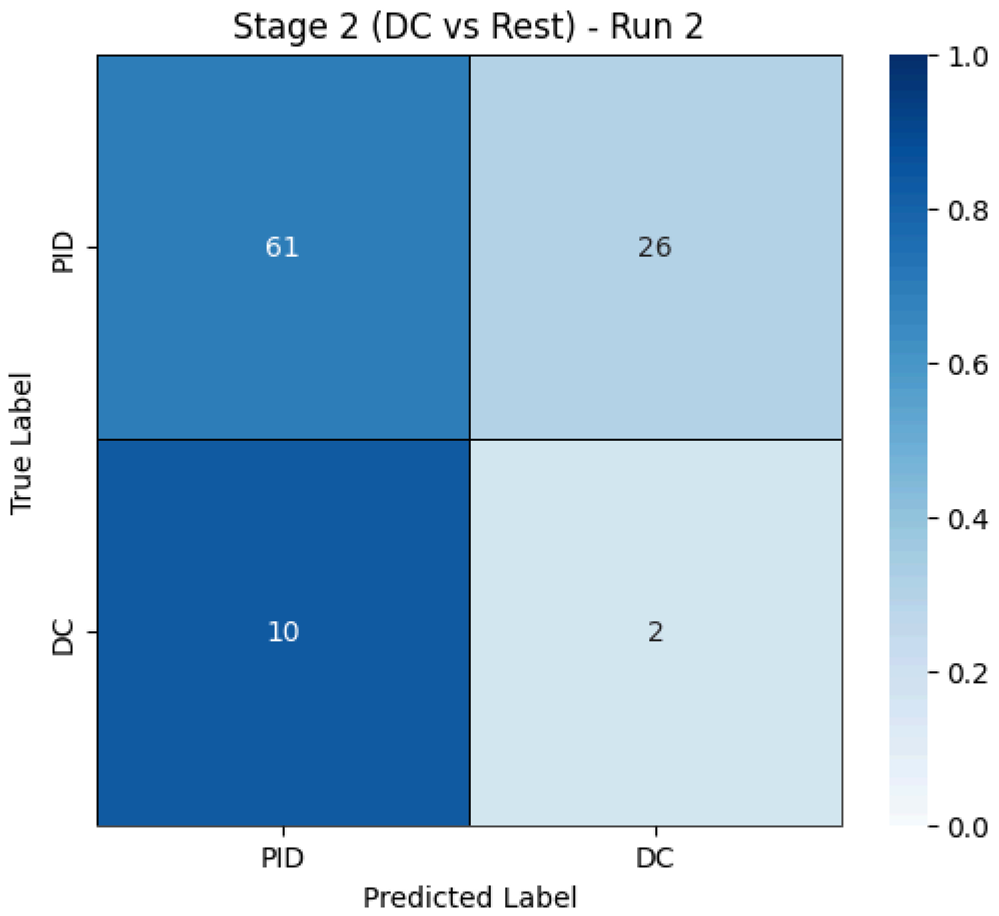


```
-- Stage 2 (DC vs Rest) --
```

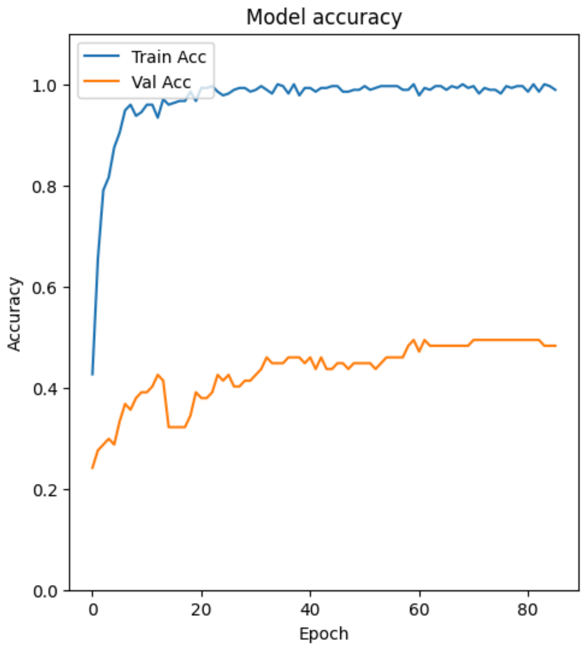


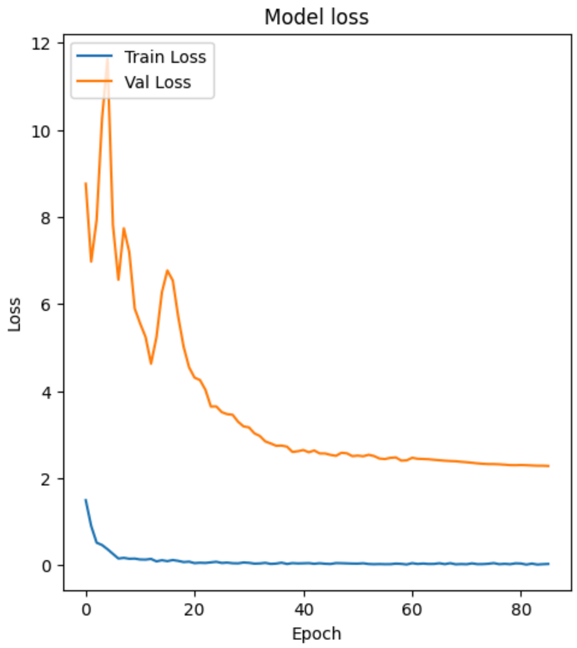
```
[Threshold Optimization] Best balanced_accuracy: 0.7104 at threshold=0.116
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

```
           0     0.8592    0.7011    0.7722        87
           1     0.0714    0.1667    0.1000        12

    accuracy                         0.6364        99
   macro avg     0.4653    0.4339    0.4361        99
weighted avg     0.7637    0.6364    0.6907        99

Balanced Accuracy: 0.4339080459770115
```

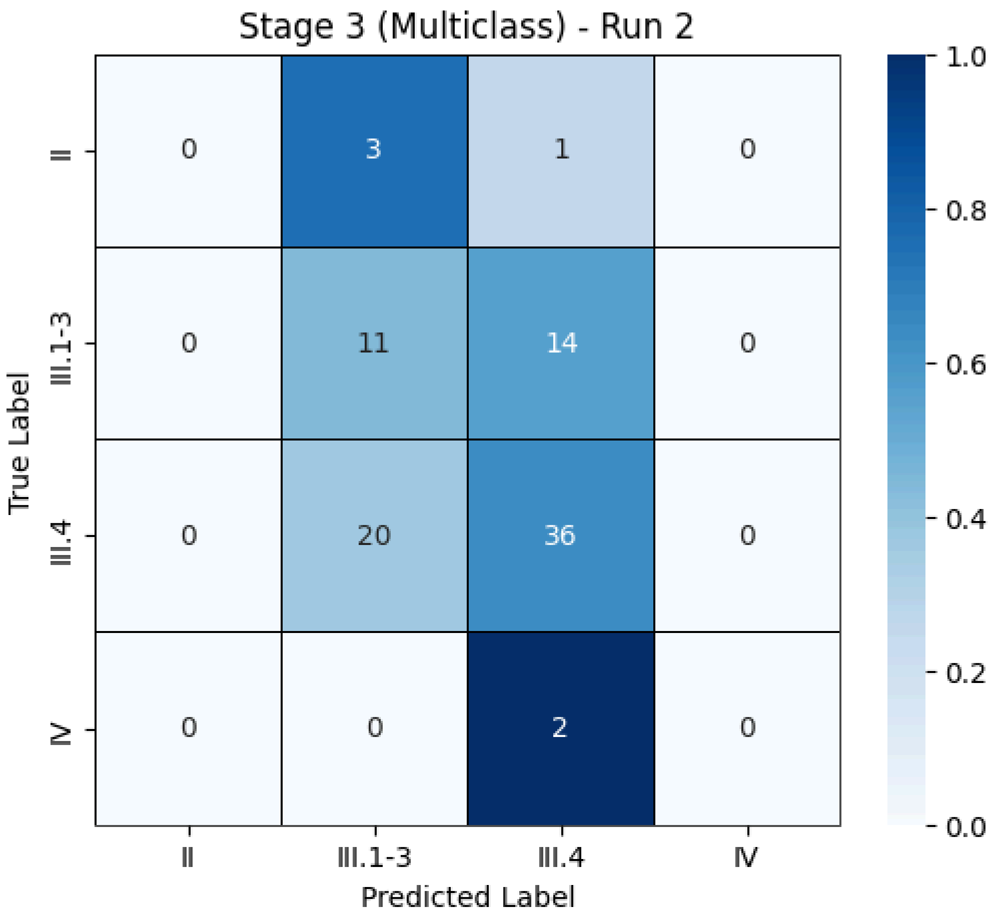


-- Stage 3 (Multiclass) --

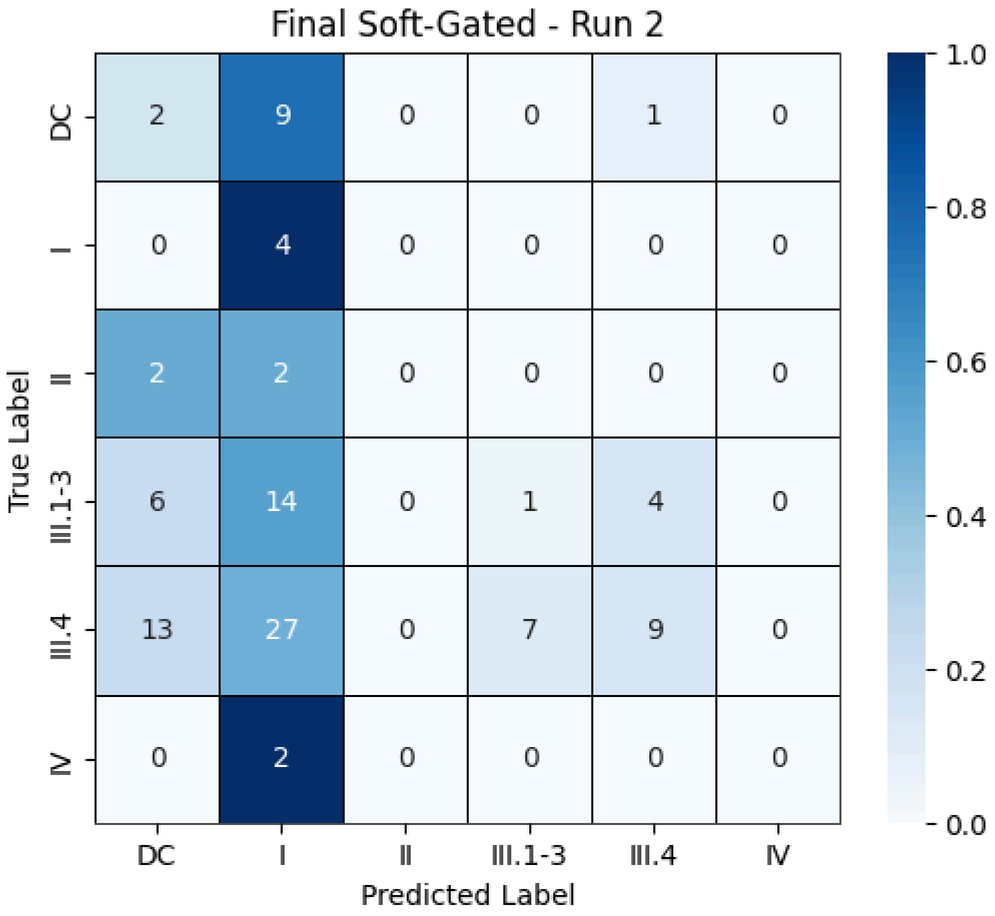




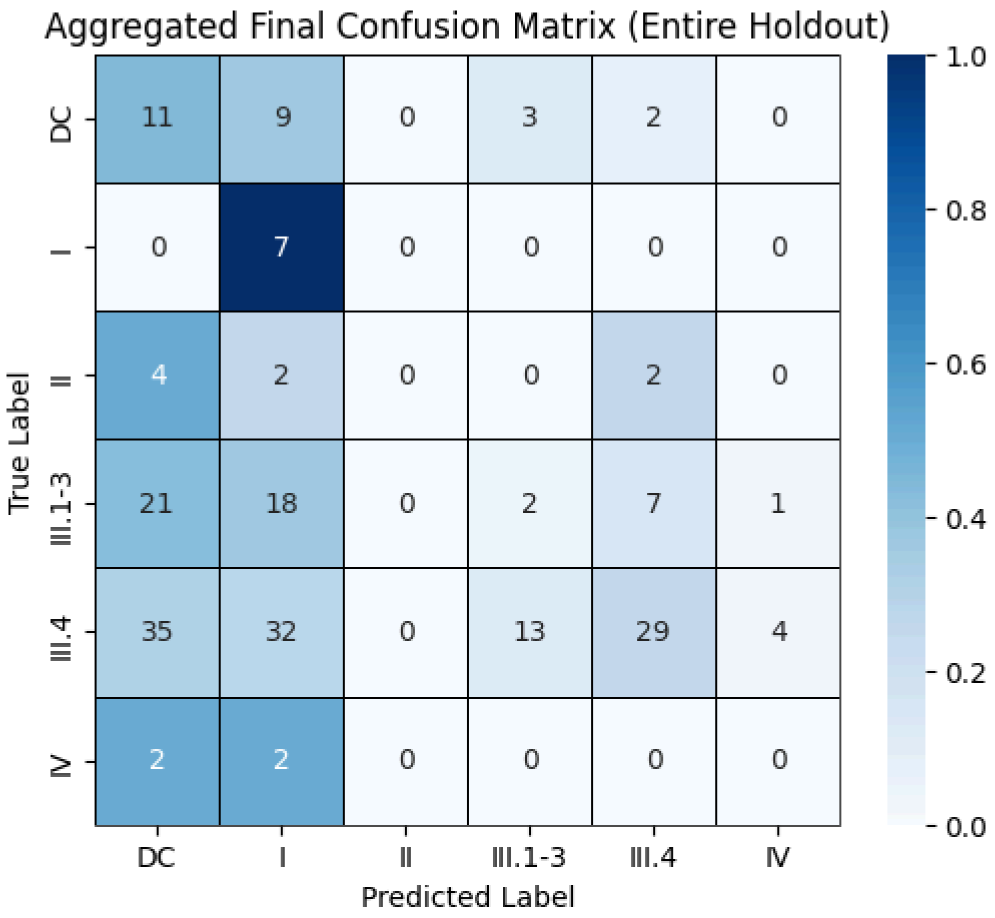
```
             precision    recall  f1-score   support

           0     0.0000    0.0000    0.0000         4
           1     0.3235    0.4400    0.3729        25
           2     0.6792    0.6429    0.6606        56
           3     0.0000    0.0000    0.0000         2

    accuracy                         0.5402        87
   macro avg     0.2507    0.2707    0.2584        87
weighted avg     0.5302    0.5402    0.5323        87

Balanced Accuracy: 0.27071428571428574
```



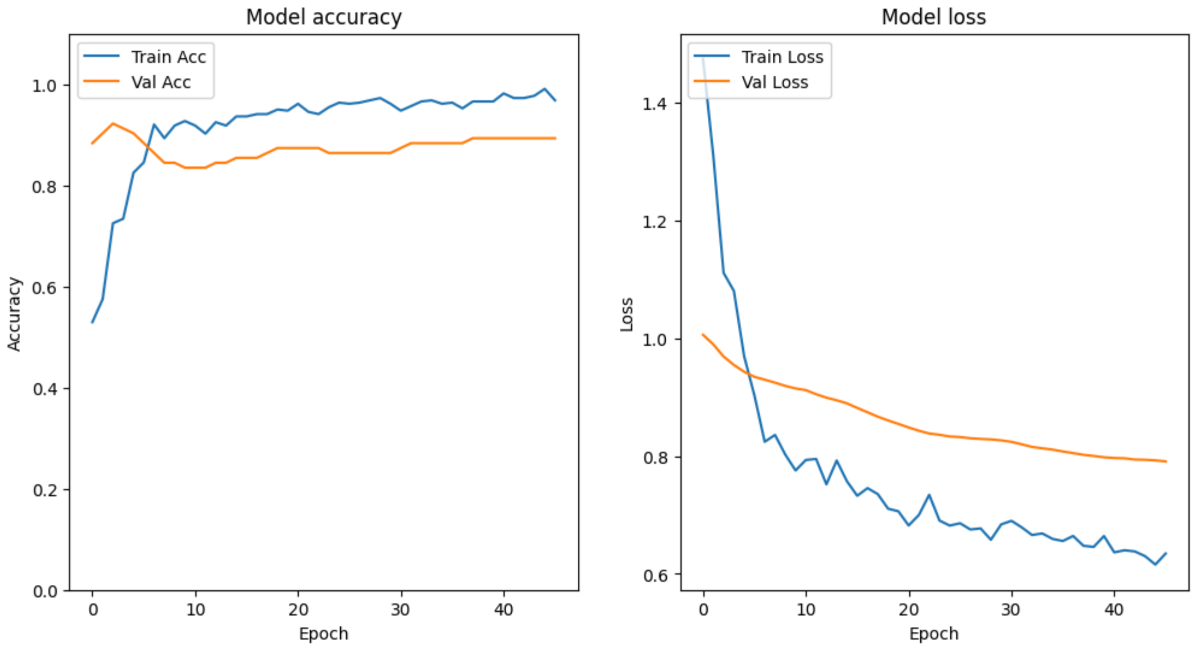
```
== Soft-Gated Overall (Test Set) - Run 2 ==
              precision    recall  f1-score   support

          DC     0.0870    0.1667    0.1143        12
           I     0.0690    1.0000    0.1290         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.1250    0.0400    0.0606        25
       III.4     0.6429    0.1607    0.2571        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.1553       103
   macro avg     0.1540    0.2279    0.0935       103
weighted avg     0.3927    0.1553    0.1728       103

Balanced Accuracy: 0.22789682539682543
```



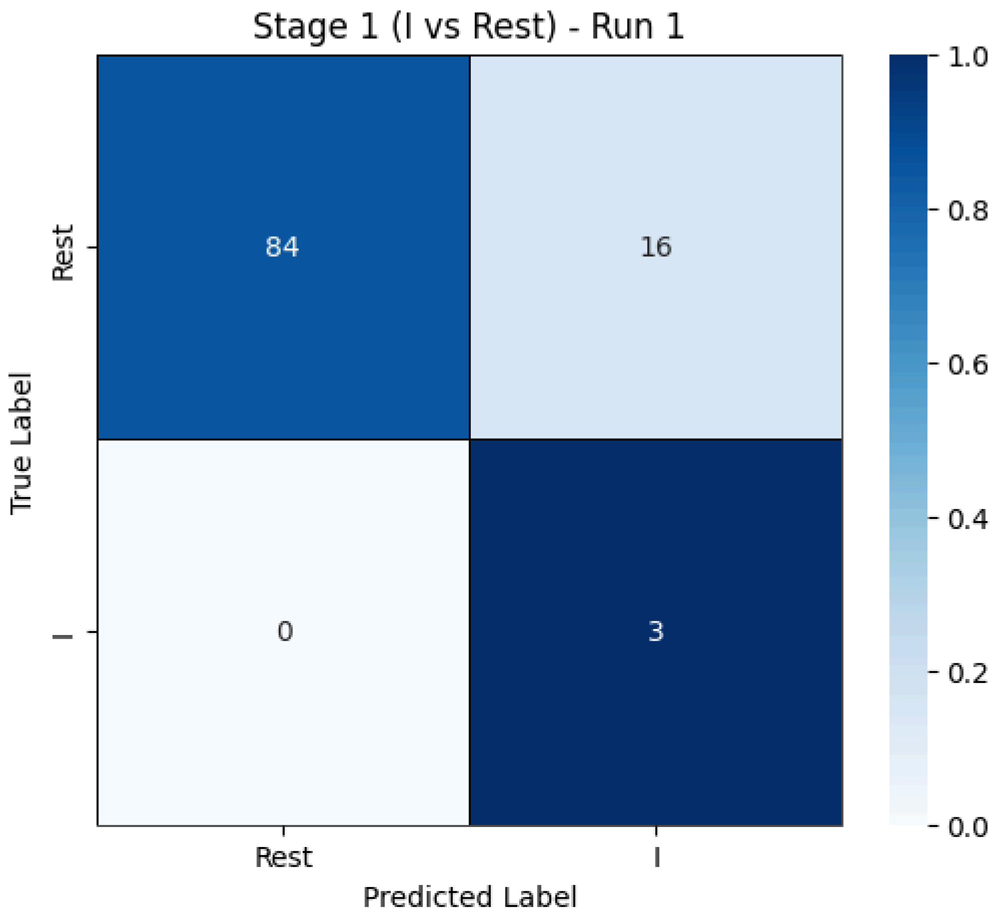
```
============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
============================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.1507    0.4400    0.2245        25
           I     0.1000    1.0000    0.1818         7
          II     0.0000    0.0000    0.0000         8
     III.1-3     0.1111    0.0408    0.0597        49
       III.4     0.7250    0.2566    0.3791       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.2379       206
   macro avg     0.1811    0.2896    0.1408       206
weighted avg     0.4458    0.2379    0.2556       206


Aggregated Balanced Accuracy: 0.2896
```



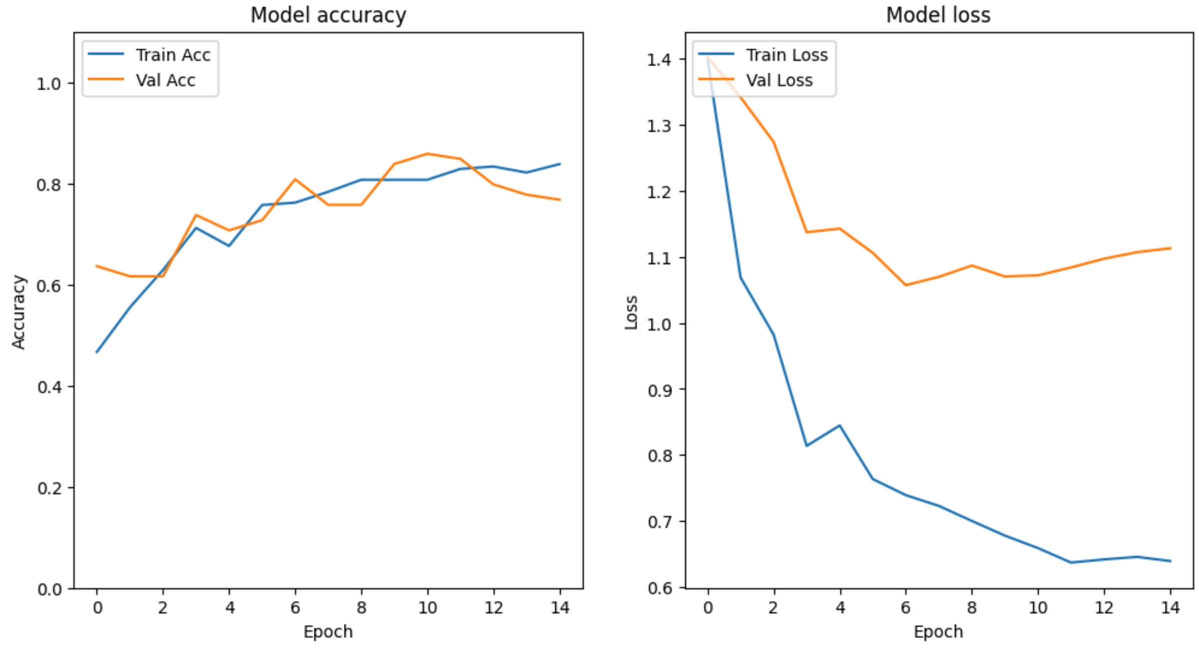
```
== Average Stage Balanced Accuracies ==
Stage 1 (I vs Rest): 0.8619
Stage 2 (DC vs Rest): 0.5251
Stage 3 (Multiclass): 0.3618
Final (Soft-Gated): 0.2877

============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.9091 at threshold=0.291
Optimal threshold (Stage 1 (I vs Rest)): 0.291
              precision    recall  f1-score   support

           0     1.0000    0.8400    0.9130       100
           1     0.1579    1.0000    0.2727         3

    accuracy                         0.8447       103
   macro avg     0.5789    0.9200    0.5929       103
weighted avg     0.9755    0.8447    0.8944       103

Balanced Accuracy: 0.9199999999999999
```

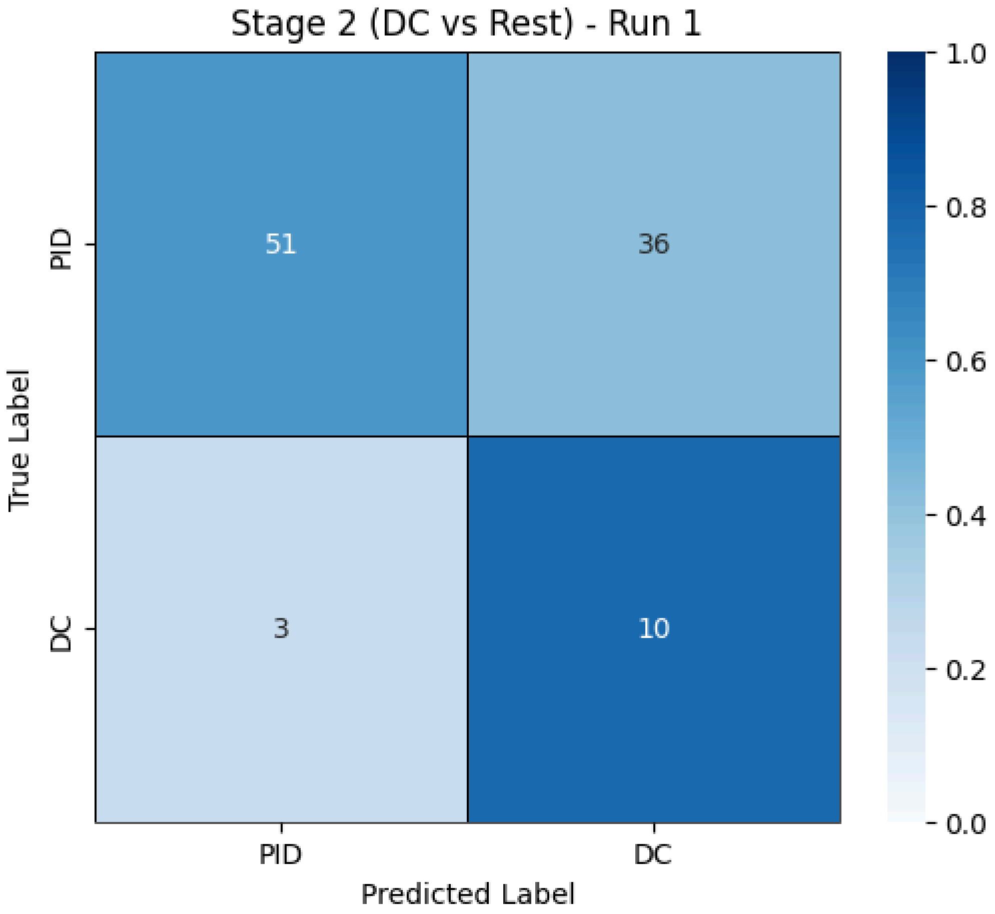


```
-- Stage 2 (DC vs Rest) --
```

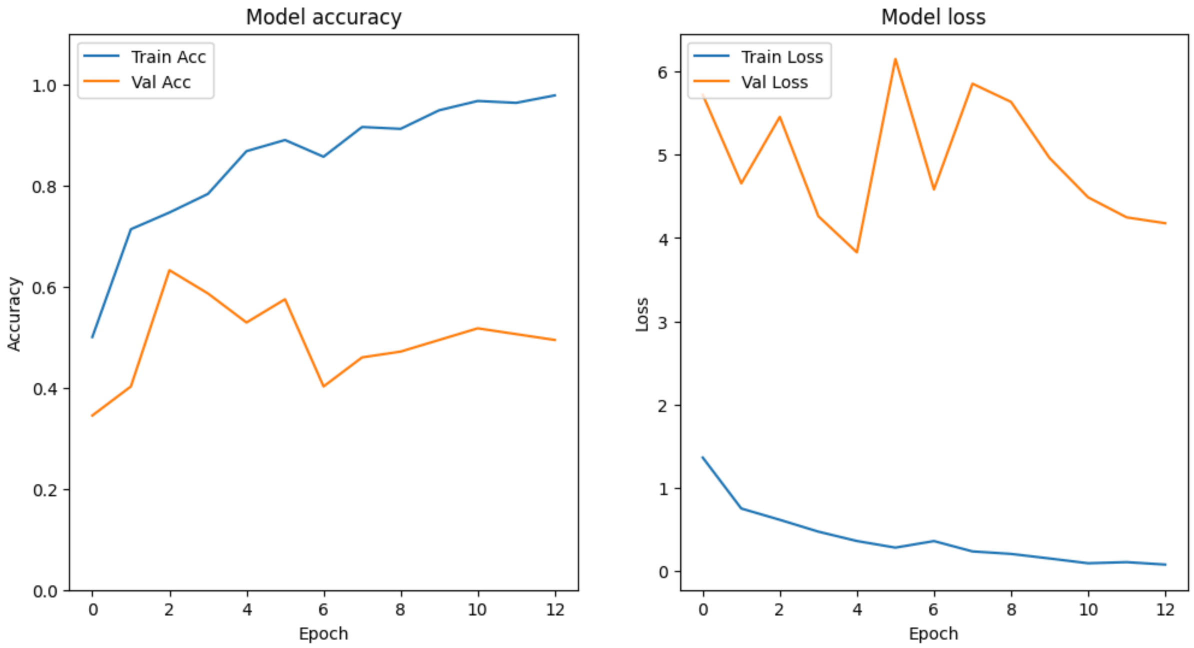


```
[Threshold Optimization] Best balanced_accuracy: 0.6279 at threshold=0.090
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

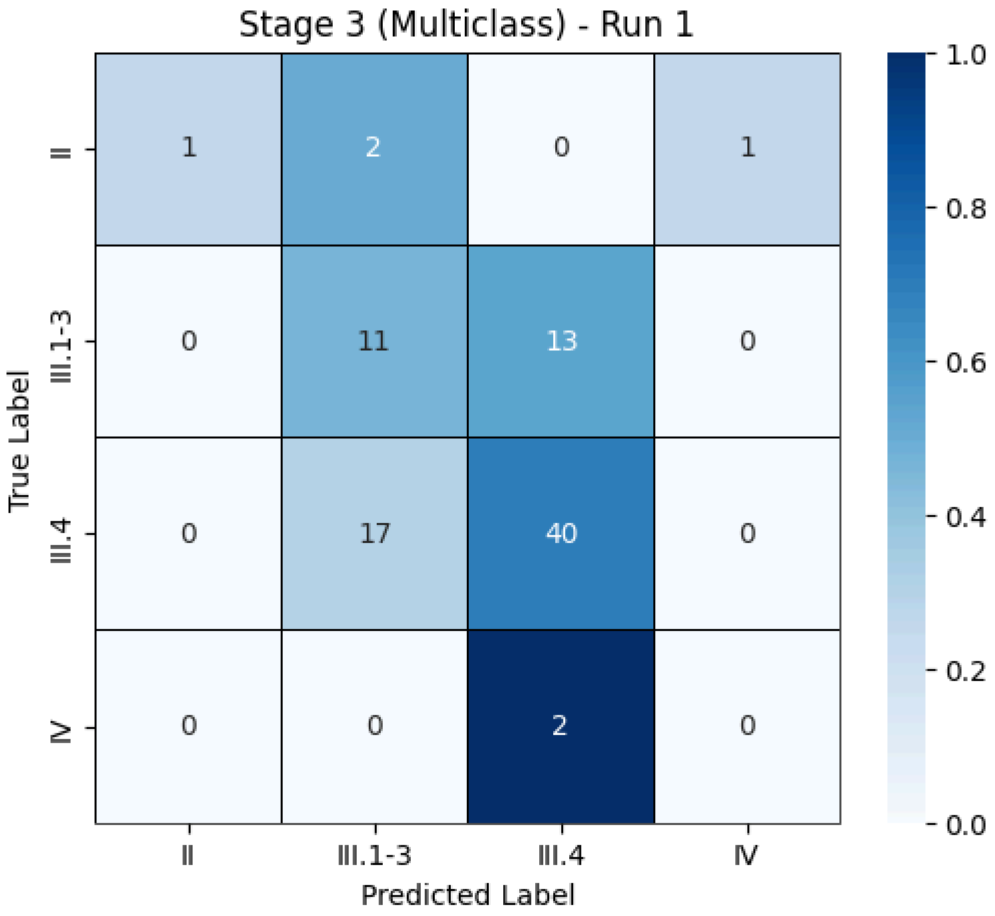
```
           0     0.9444    0.5862    0.7234        87
           1     0.2174    0.7692    0.3390        13

    accuracy                         0.6100       100
   macro avg     0.5809    0.6777    0.5312       100
weighted avg     0.8499    0.6100    0.6734       100

Balanced Accuracy: 0.6777188328912467
```



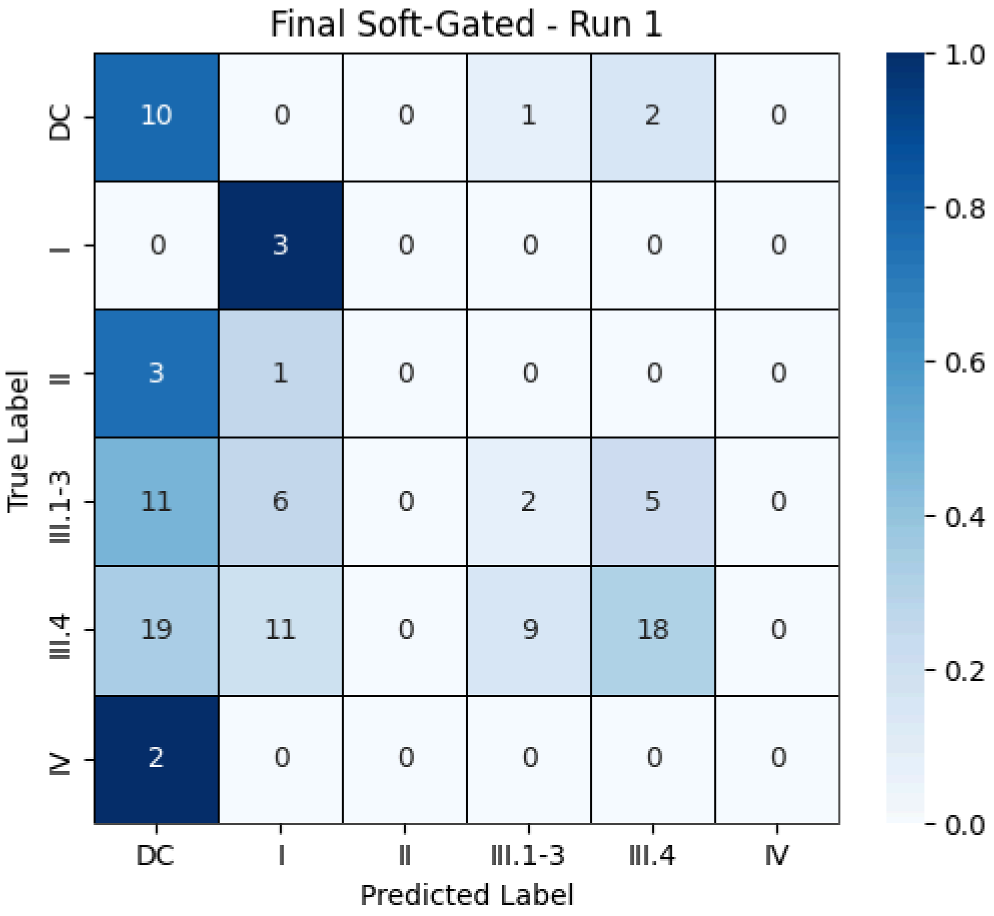
```
-- Stage 3 (Multiclass) --
```



```
              precision    recall  f1-score   support

           0     1.0000    0.2500    0.4000         4
           1     0.3667    0.4583    0.4074        24
           2     0.7273    0.7018    0.7143        57
           3     0.0000    0.0000    0.0000         2

    accuracy                         0.5977        87
   macro avg     0.5235    0.3525    0.3804        87
weighted avg     0.6236    0.5977    0.5988        87

Balanced Accuracy: 0.3525219298245614
```

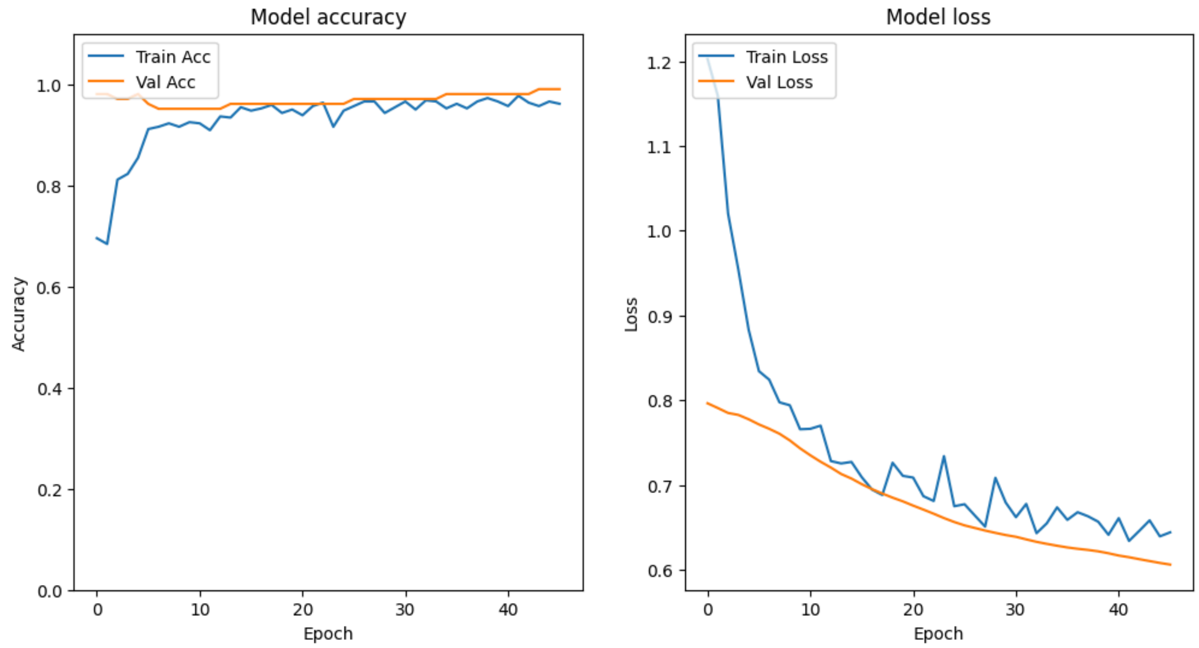


```
== Soft-Gated Overall (Test Set) - Run 1 ==
          precision    recall  f1-score   support

       DC     0.2222    0.7692    0.3448        13
        I     0.1429    1.0000    0.2500         3
       II     0.0000    0.0000    0.0000         4
  III.1-3     0.1667    0.0833    0.1111        24
    III.4     0.7200    0.3158    0.4390        57
       IV     0.0000    0.0000    0.0000         2

 accuracy                         0.3204       103
macro avg     0.2086    0.3614    0.1908       103
weighted avg  0.4695    0.3204    0.3196       103

Balanced Accuracy: 0.36139226270805214
```

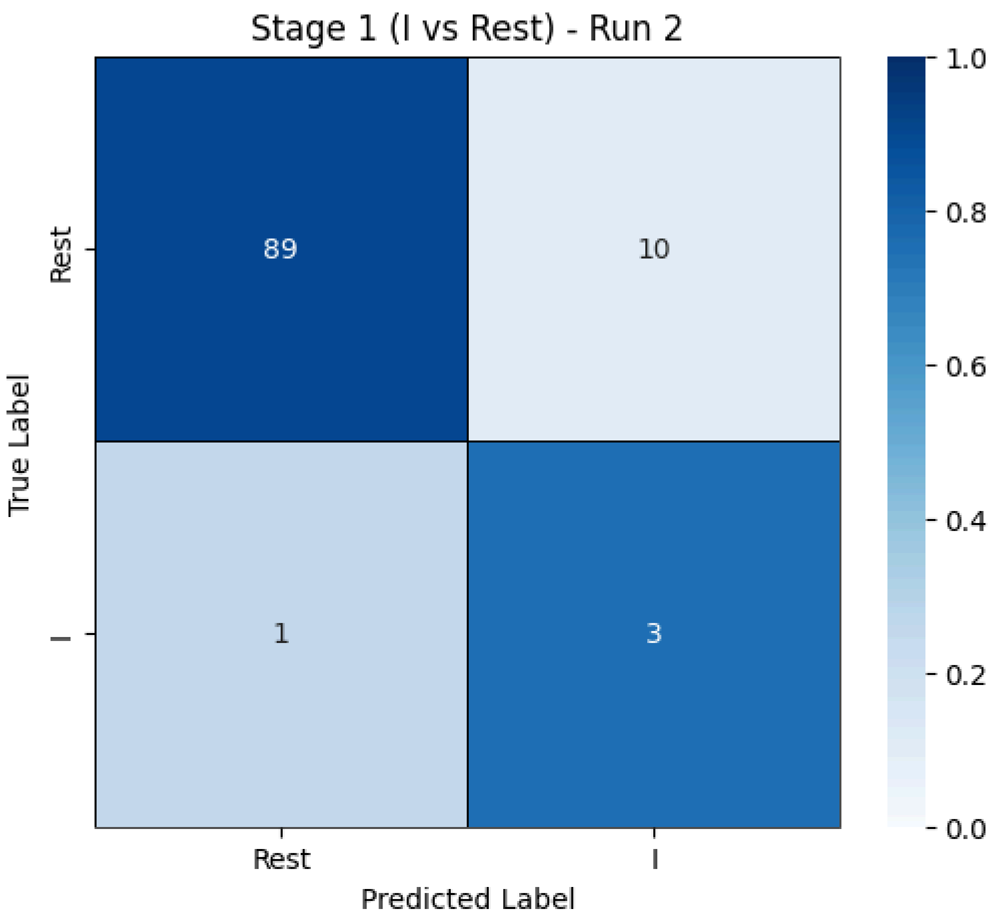


	DC	I	II	III.1-3	III.4	IV
DC	10	0	0	1	2	0
I	0	3	0	0	0	0
II	3	1	0	0	0	0
III.1-3	11	6	0	2	5	0
III.4	19	11	0	9	18	0
IV	2	0	0	0	0	0

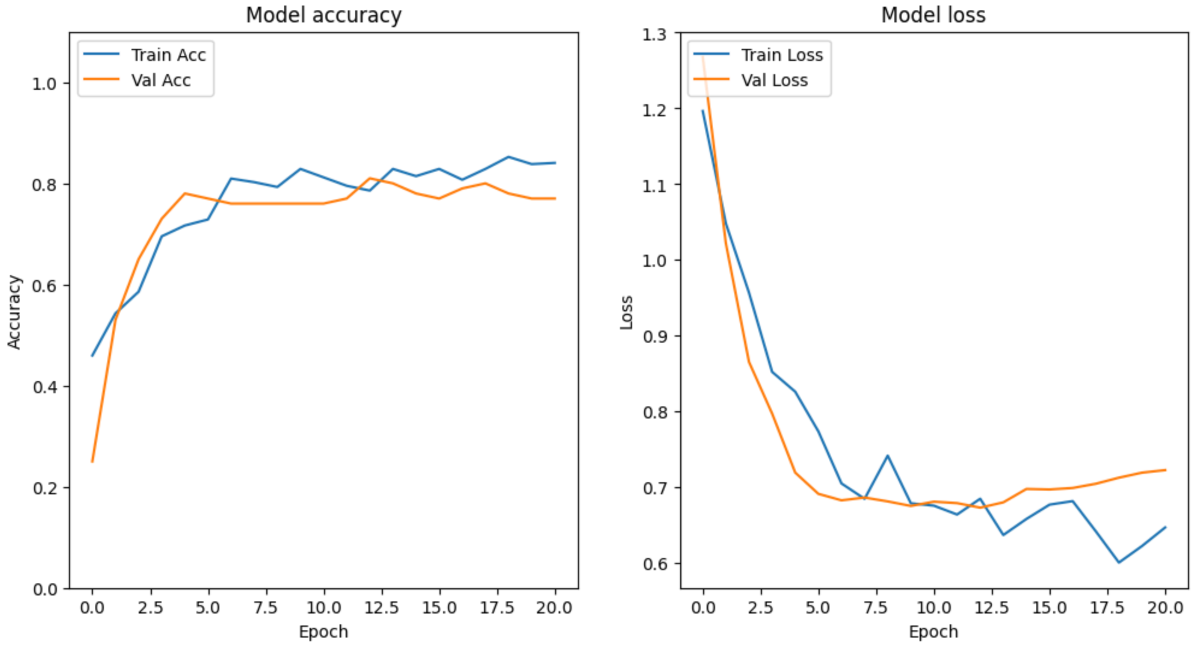
```
============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.9850 at threshold=0.281
Optimal threshold (Stage 1 (I vs Rest)): 0.281
              precision    recall  f1-score   support

           0     0.9889    0.8990    0.9418        99
           1     0.2308    0.7500    0.3529         4

    accuracy                         0.8932       103
   macro avg     0.6098    0.8245    0.6474       103
weighted avg     0.9594    0.8932    0.9189       103

Balanced Accuracy: 0.8244949494949495
```

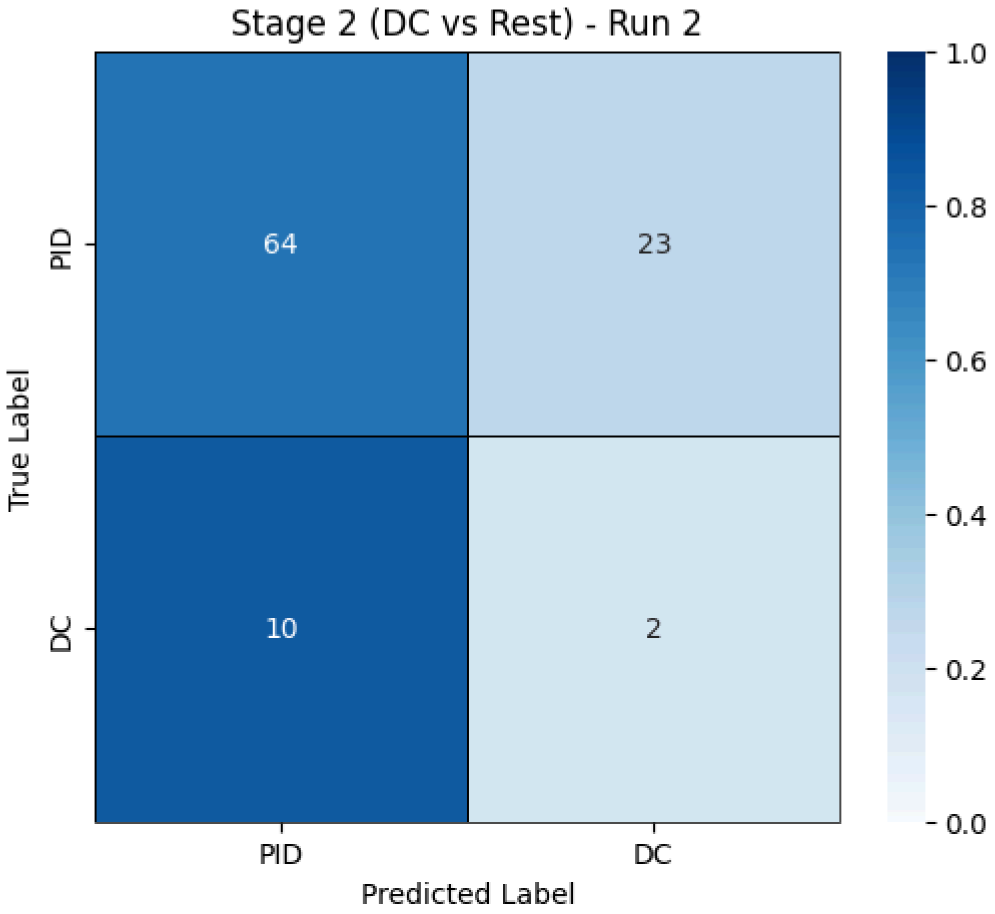


```
-- Stage 2 (DC vs Rest) --
```

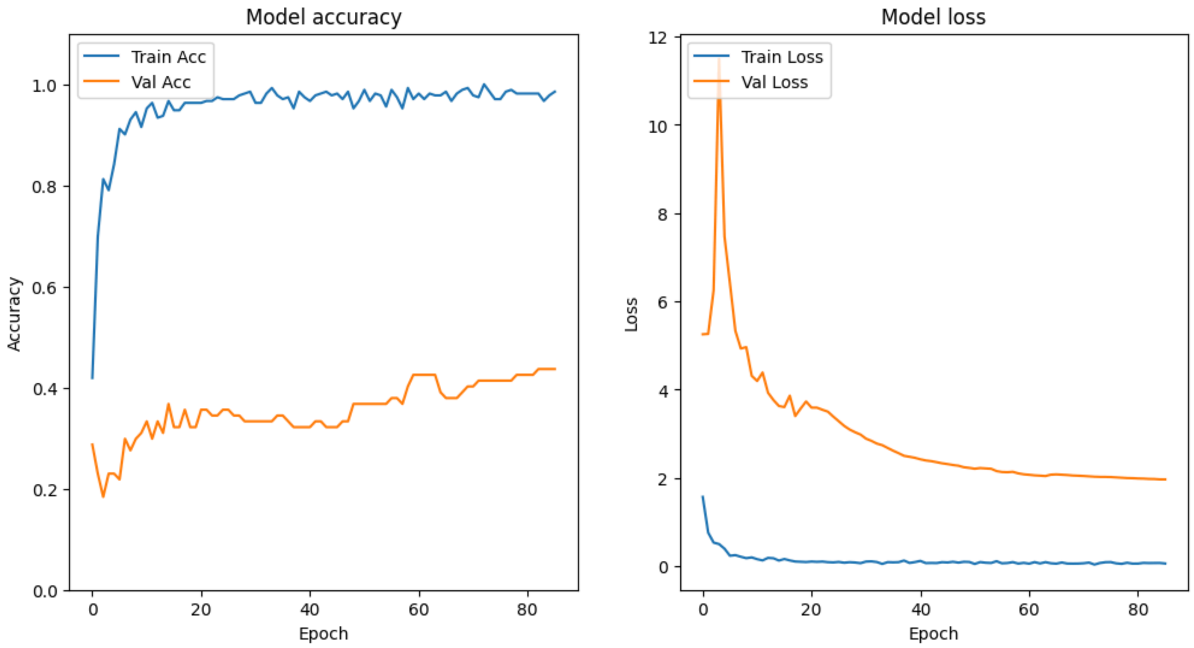


```
[Threshold Optimization] Best balanced_accuracy: 0.7259 at threshold=0.075
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

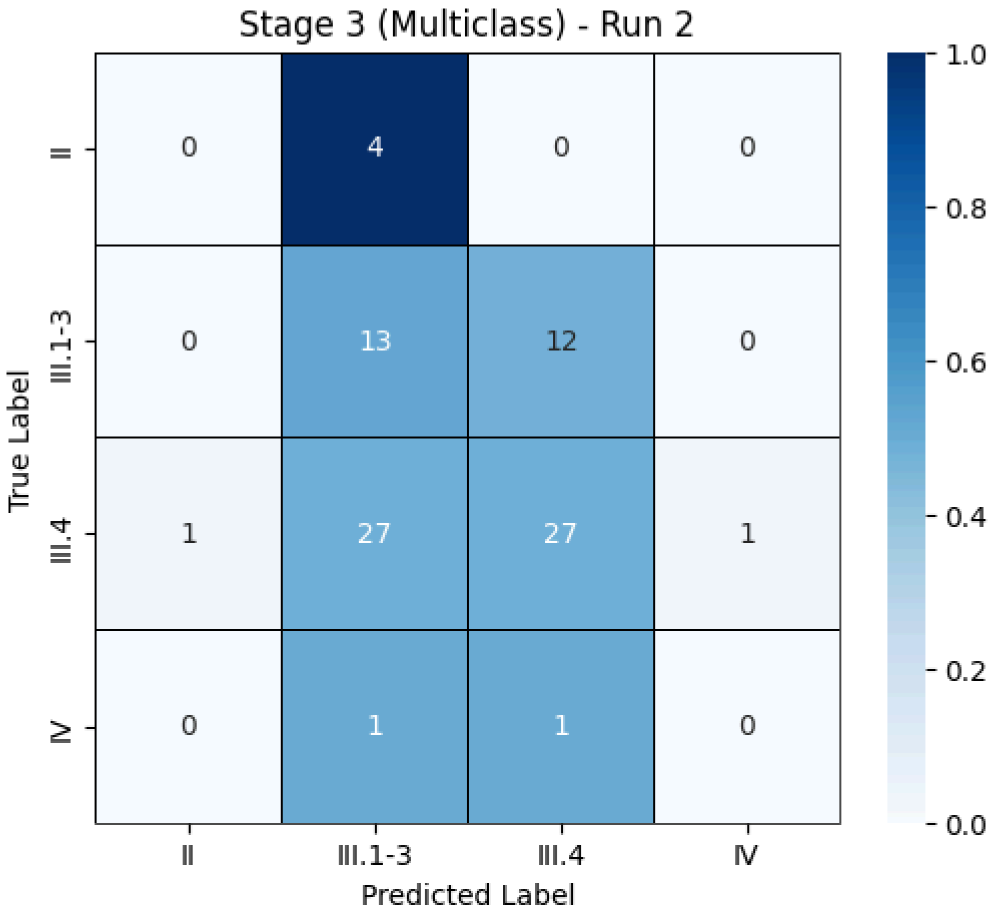
```
           0     0.8649    0.7356    0.7950        87
           1     0.0800    0.1667    0.1081        12

    accuracy                         0.6667        99
   macro avg     0.4724    0.4511    0.4516        99
weighted avg     0.7697    0.6667    0.7118        99

Balanced Accuracy: 0.4511494252873563
```



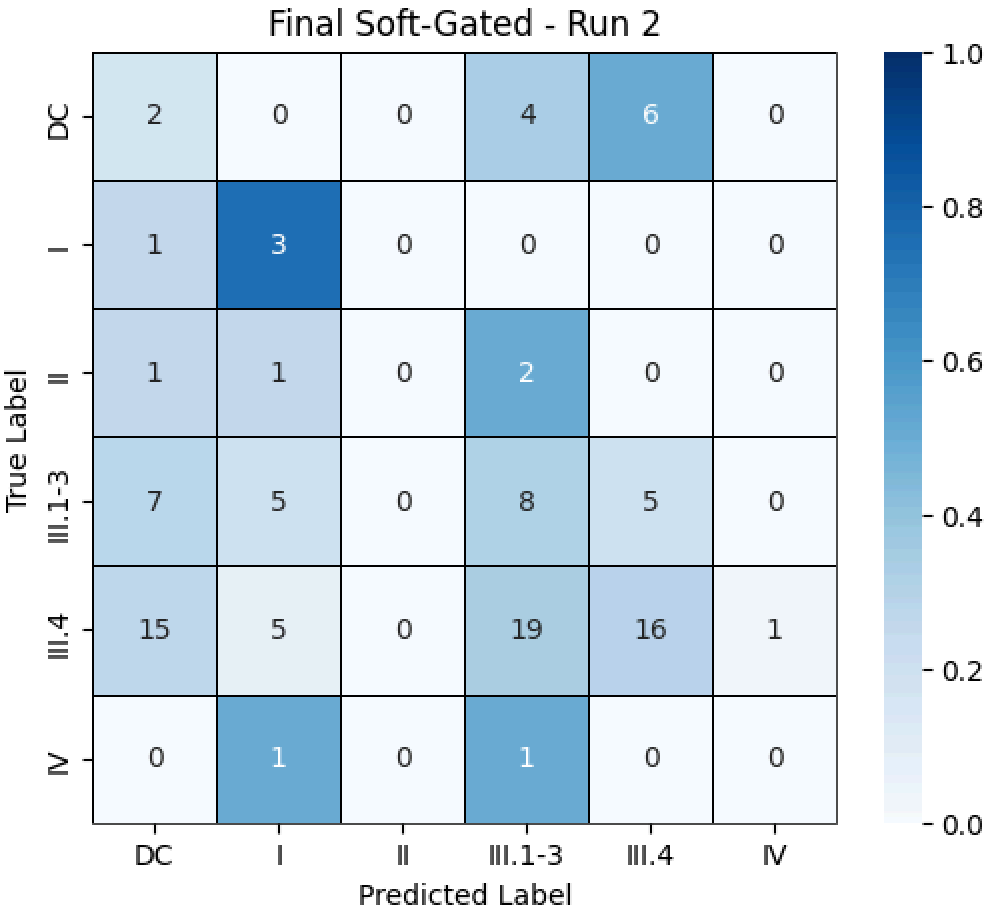
```
-- Stage 3 (Multiclass) --
```



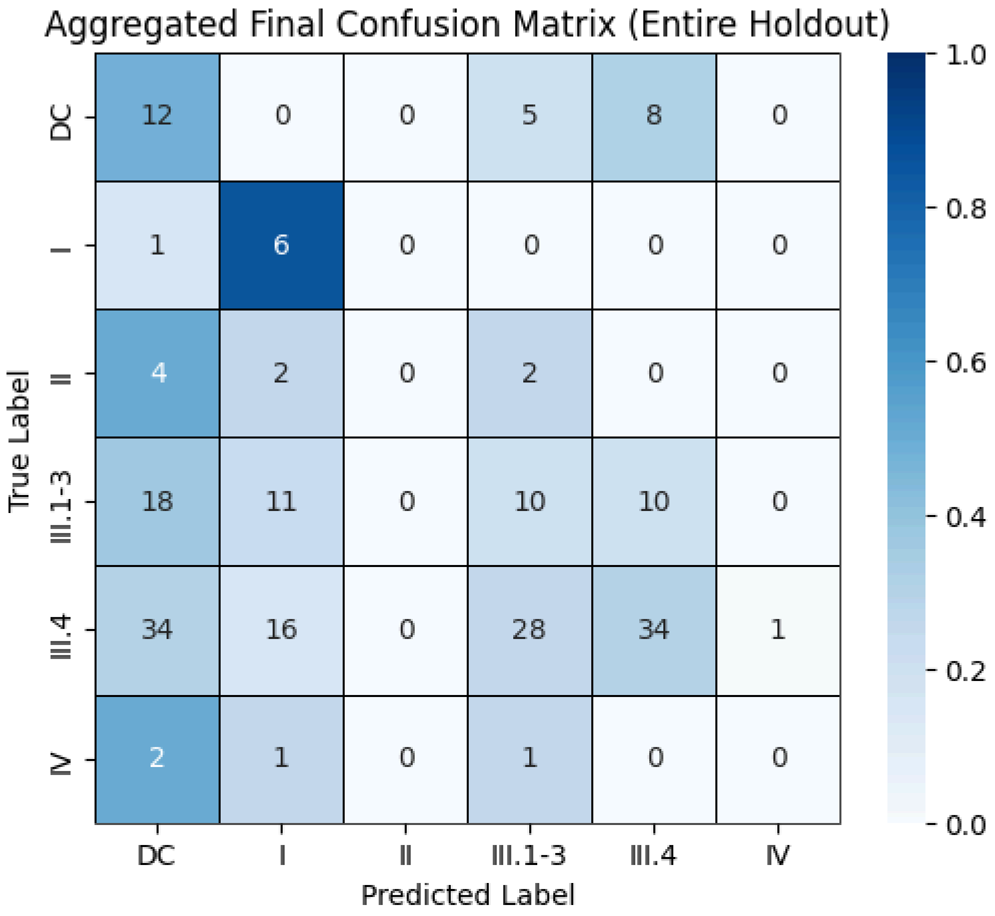
```
           precision    recall  f1-score   support

         0     0.0000    0.0000    0.0000         4
         1     0.2889    0.5200    0.3714        25
         2     0.6750    0.4821    0.5625        56
         3     0.0000    0.0000    0.0000         2

  accuracy                         0.4598        87
 macro avg     0.2410    0.2505    0.2335        87
weighted avg     0.5175    0.4598    0.4688        87

Balanced Accuracy: 0.2505357142857143
```



```
== Soft-Gated Overall (Test Set) - Run 2 ==
              precision    recall  f1-score   support

          DC     0.0769    0.1667    0.1053        12
           I     0.2000    0.7500    0.3158         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.2353    0.3200    0.2712        25
       III.4     0.5926    0.2857    0.3855        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.2816       103
   macro avg     0.1841    0.2537    0.1796       103
weighted avg     0.3960    0.2816    0.3000       103

Balanced Accuracy: 0.25373015873015875
```

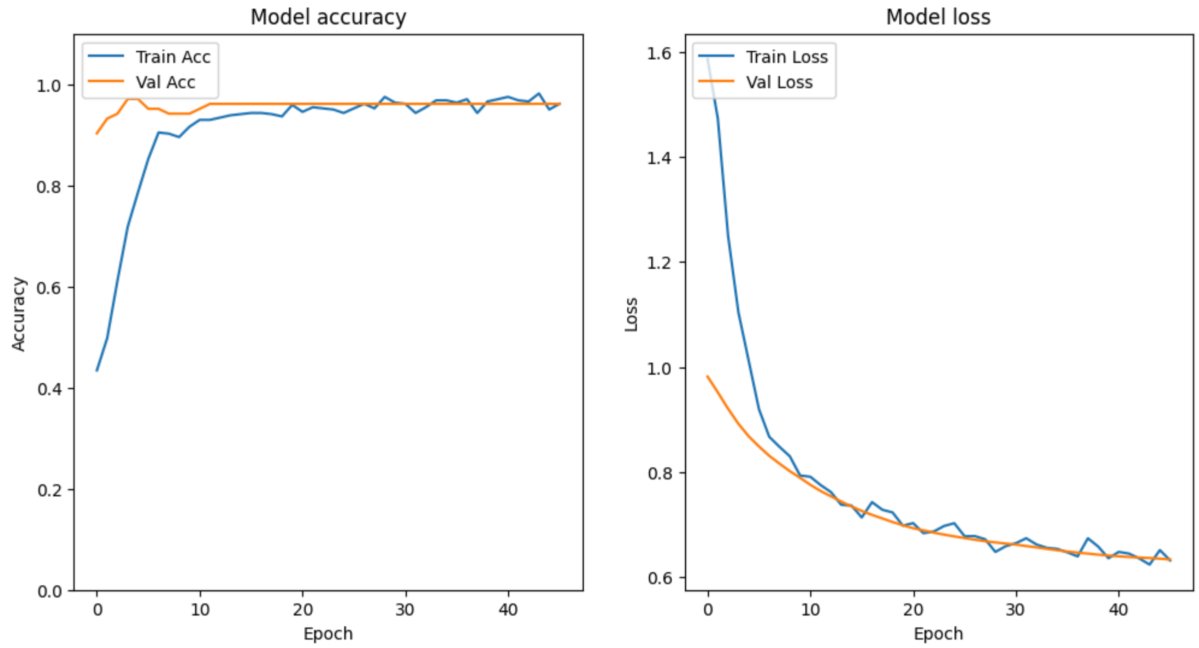


```
================================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
================================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.1690    0.4800    0.2500        25
           I     0.1667    0.8571    0.2791         7
          II     0.0000    0.0000    0.0000         8
     III.1-3     0.2174    0.2041    0.2105        49
       III.4     0.6538    0.3009    0.4121       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.3010       206
   macro avg     0.2012    0.3070    0.1920       206
weighted avg     0.4365    0.3010    0.3160       206


Aggregated Balanced Accuracy: 0.3070
```

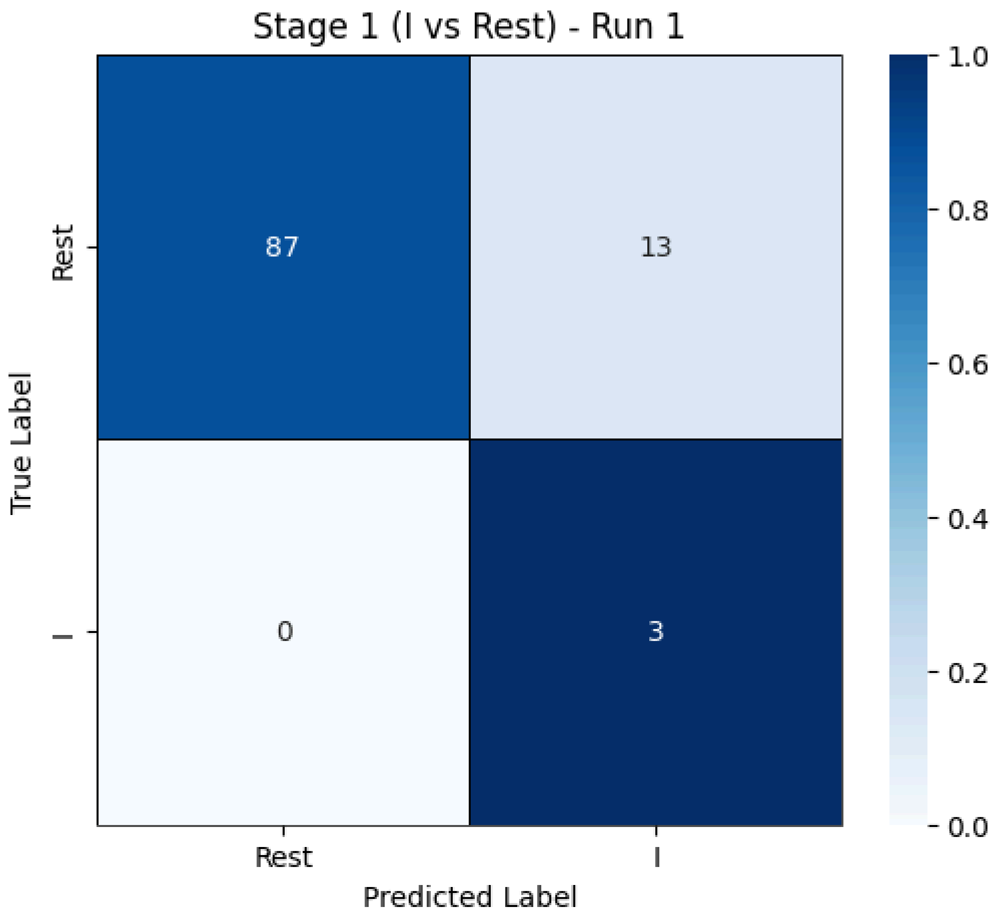


	DC	I	II	III.1-3	III.4	IV
DC	12	0	0	5	8	0
I	1	6	0	0	0	0
II	4	2	0	2	0	0
III.1-3	18	11	0	10	10	0
III.4	34	16	0	28	34	1
IV	2	1	0	1	0	0

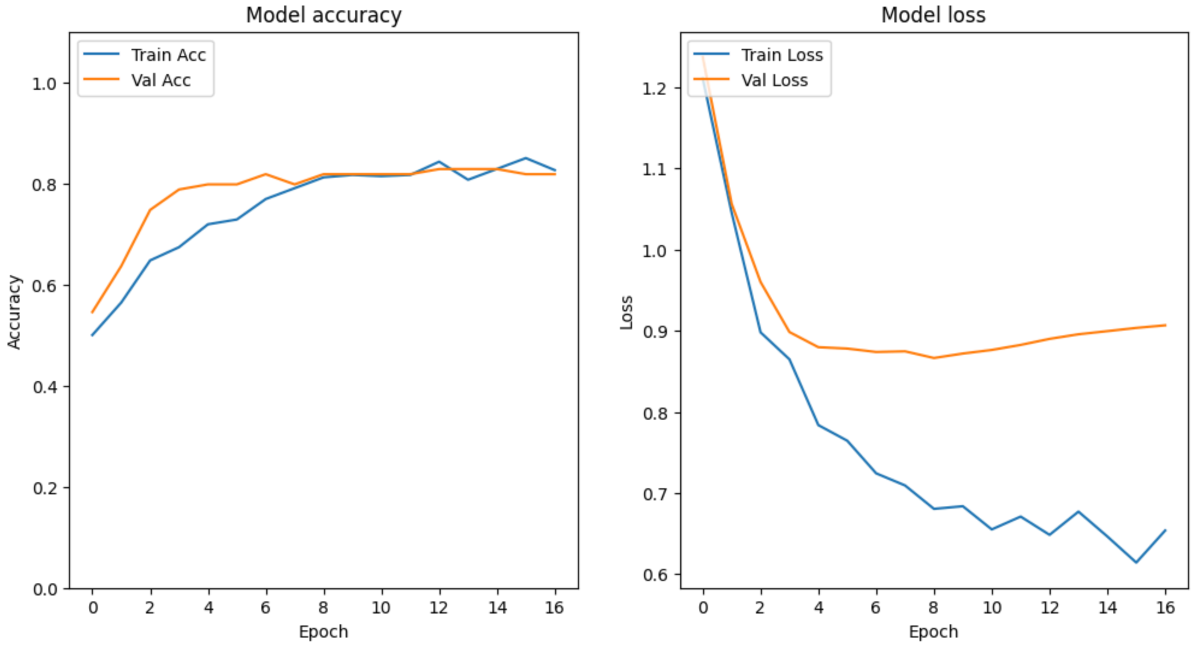
```
== Average Stage Balanced Accuracies ==
Stage 1 (I vs Rest): 0.8722
Stage 2 (DC vs Rest): 0.5644
Stage 3 (Multiclass): 0.3015
Final (Soft-Gated): 0.3076

============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.8838 at threshold=0.050
Optimal threshold (Stage 1 (I vs Rest)): 0.050
              precision    recall  f1-score   support

           0     1.0000    0.8700    0.9305       100
           1     0.1875    1.0000    0.3158         3

    accuracy                         0.8738       103
   macro avg     0.5938    0.9350    0.6231       103
weighted avg     0.9763    0.8738    0.9126       103

Balanced Accuracy: 0.935
```

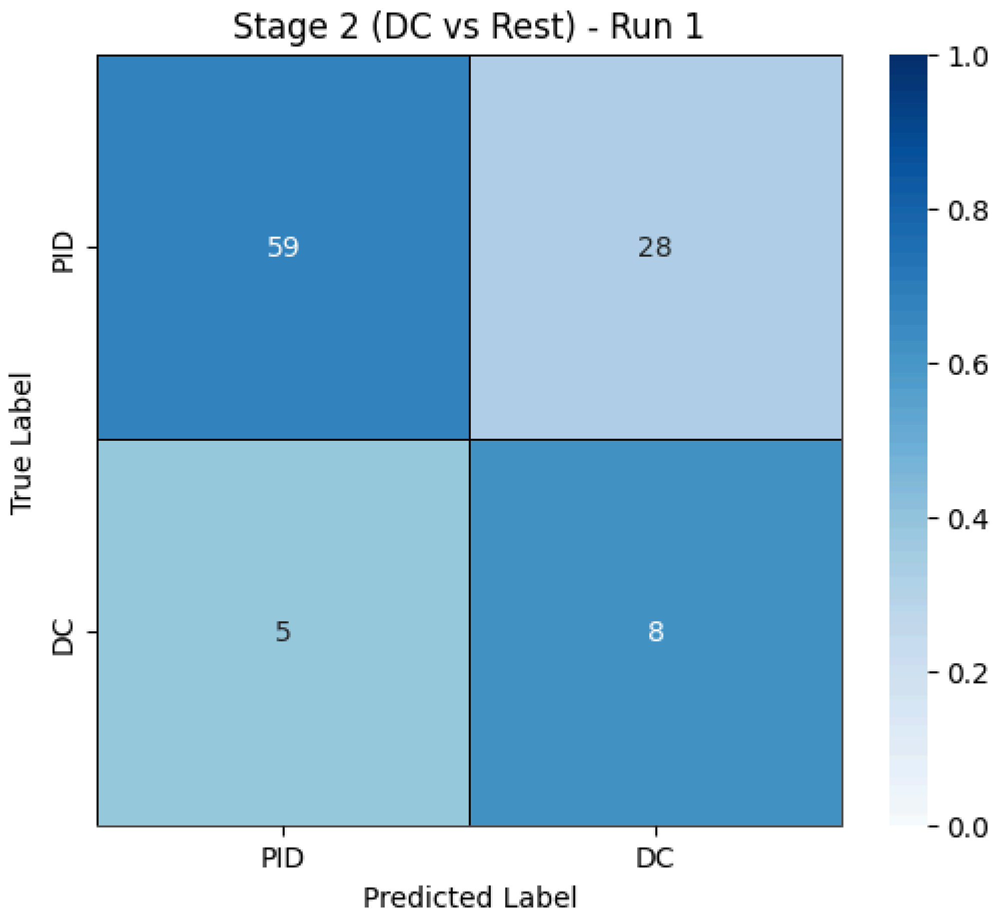


```
-- Stage 2 (DC vs Rest) --
```

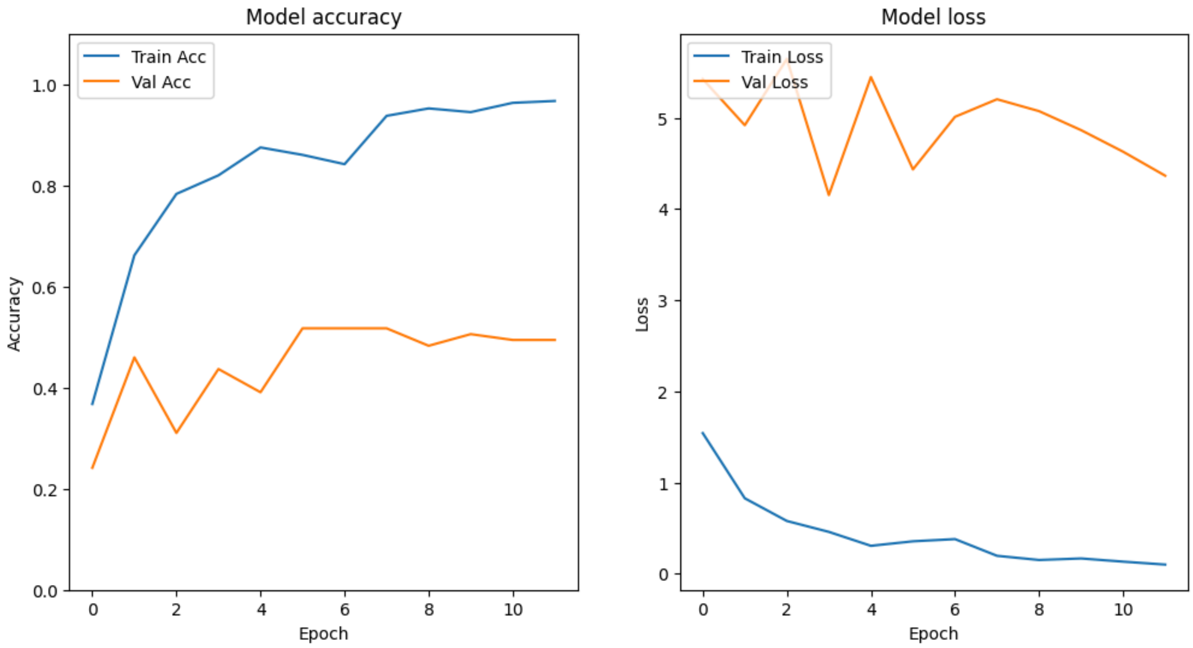


```
[Threshold Optimization] Best balanced_accuracy: 0.6351 at threshold=0.161
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

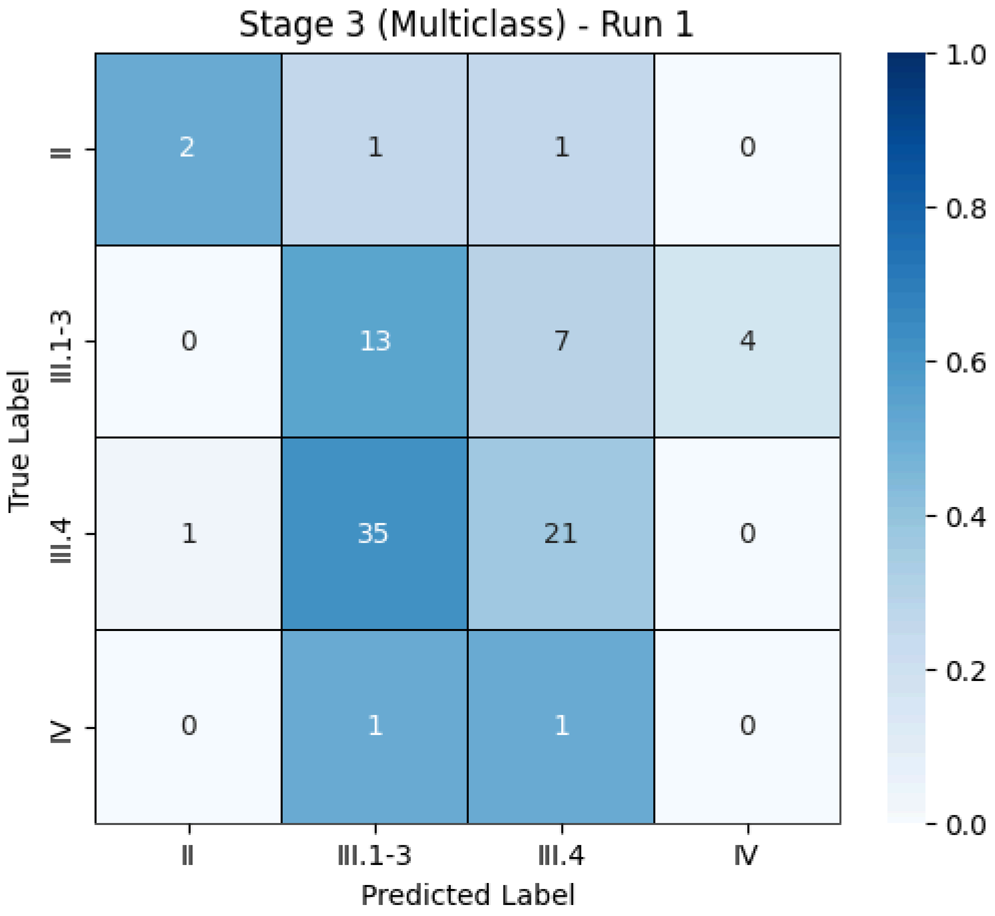
```
           0     0.9219    0.6782    0.7815        87
           1     0.2222    0.6154    0.3265        13

    accuracy                         0.6700       100
   macro avg     0.5720    0.6468    0.5540       100
weighted avg     0.8309    0.6700    0.7223       100

Balanced Accuracy: 0.6467727674624226
```



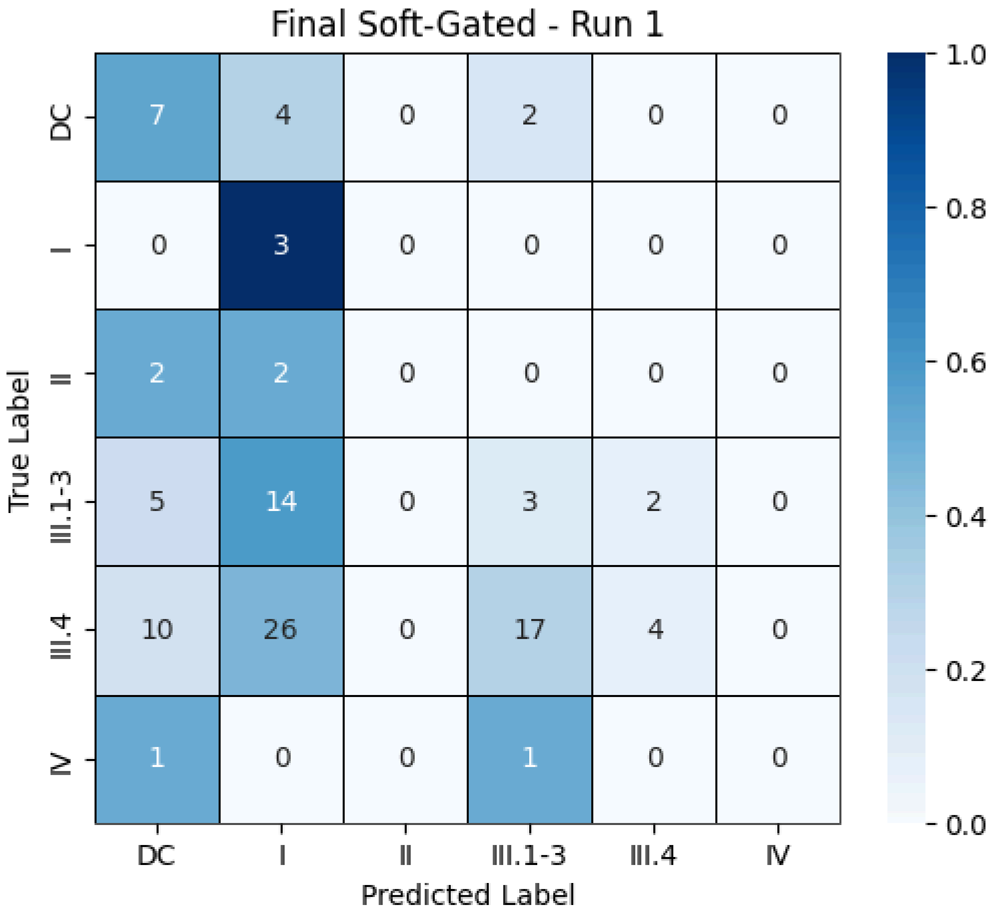
-- Stage 3 (Multiclass) --



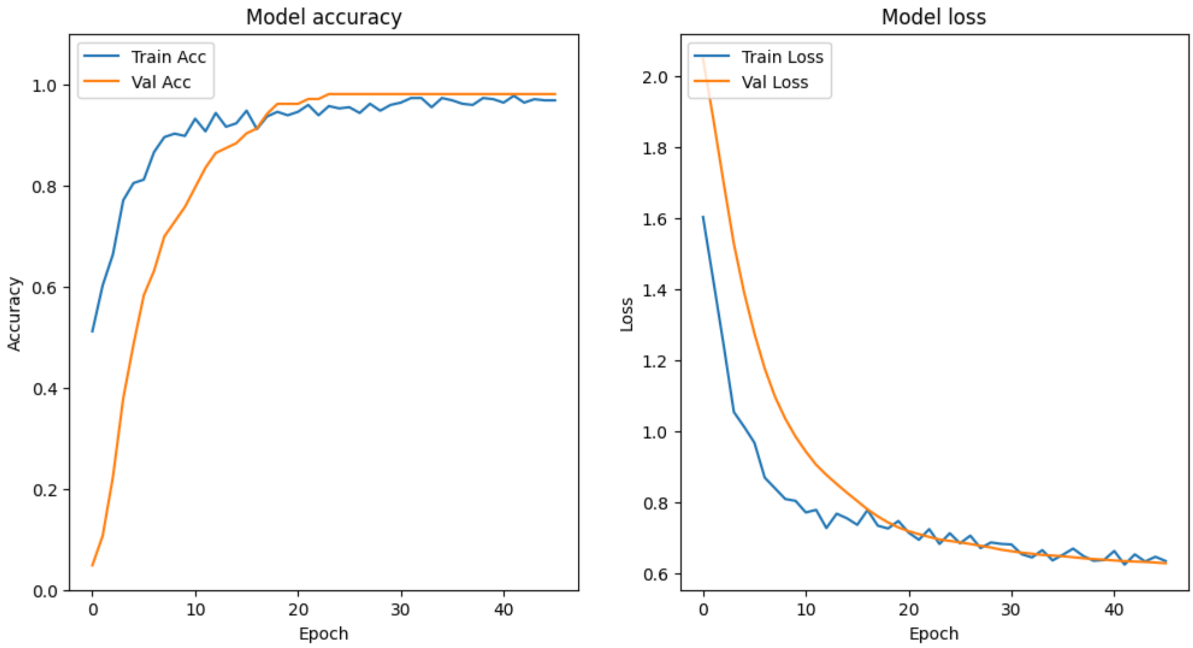
```
           precision    recall  f1-score   support

         0    0.6667    0.5000    0.5714         4
         1    0.2600    0.5417    0.3514        24
         2    0.7000    0.3684    0.4828        57
         3    0.0000    0.0000    0.0000         2

  accuracy                        0.4138        87
 macro avg    0.4067    0.3525    0.3514        87
weighted avg    0.5610    0.4138    0.4395        87

Balanced Accuracy: 0.3525219298245614
```



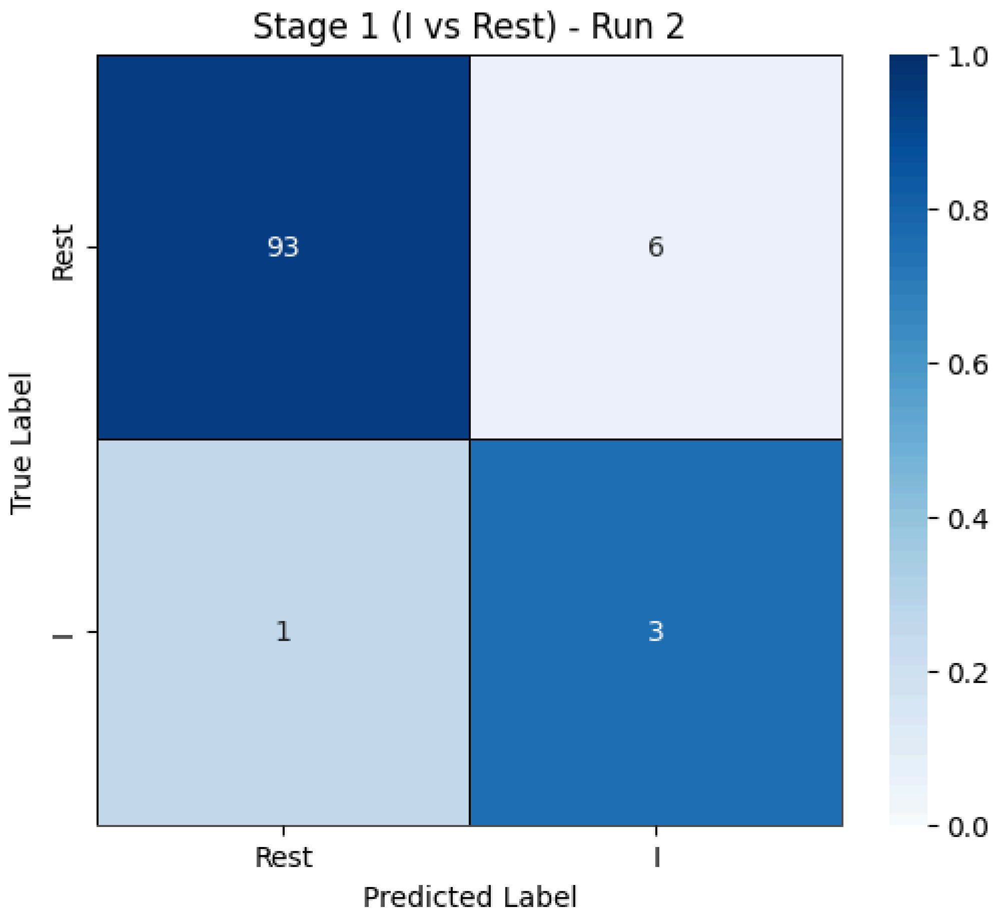
```
== Soft-Gated Overall (Test Set) - Run 1 ==
              precision    recall  f1-score   support

          DC     0.2800    0.5385    0.3684        13
           I     0.0612    1.0000    0.1154         3
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.1304    0.1250    0.1277        24
       III.4     0.6667    0.0702    0.1270        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.1650       103
   macro avg     0.1897    0.2889    0.1231       103
weighted avg     0.4364    0.1650    0.1499       103

Balanced Accuracy: 0.28893949617633824
```



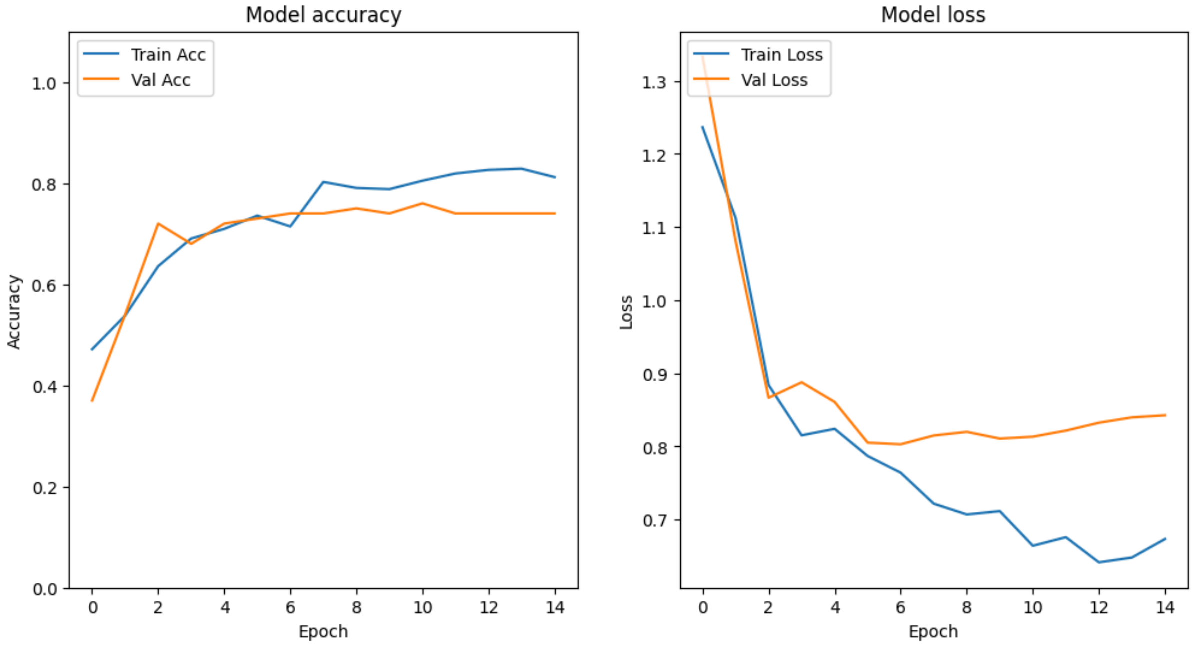
```
==============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
==============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.9900 at threshold=0.362
Optimal threshold (Stage 1 (I vs Rest)): 0.362
              precision    recall  f1-score   support

           0     0.9894    0.9394    0.9637        99
           1     0.3333    0.7500    0.4615         4

    accuracy                         0.9320       103
   macro avg     0.6613    0.8447    0.7126       103
weighted avg     0.9639    0.9320    0.9442       103

Balanced Accuracy: 0.8446969696969697
```

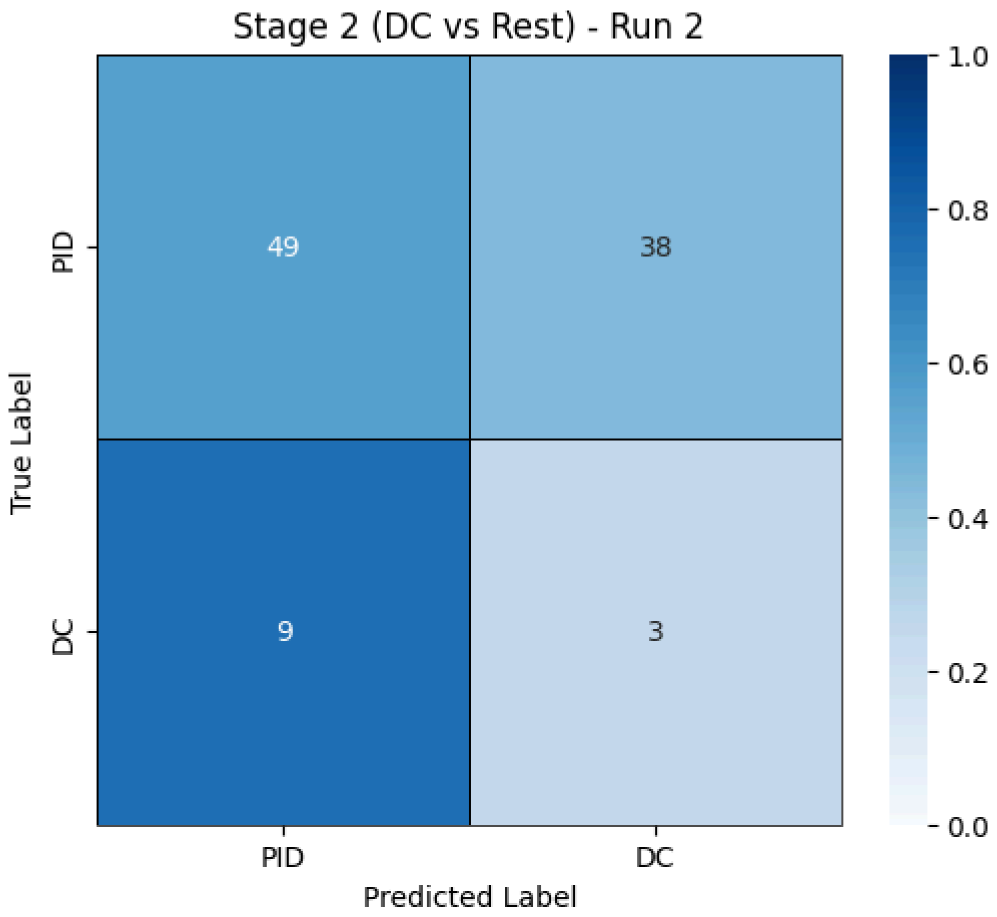


```
-- Stage 2 (DC vs Rest) --
```

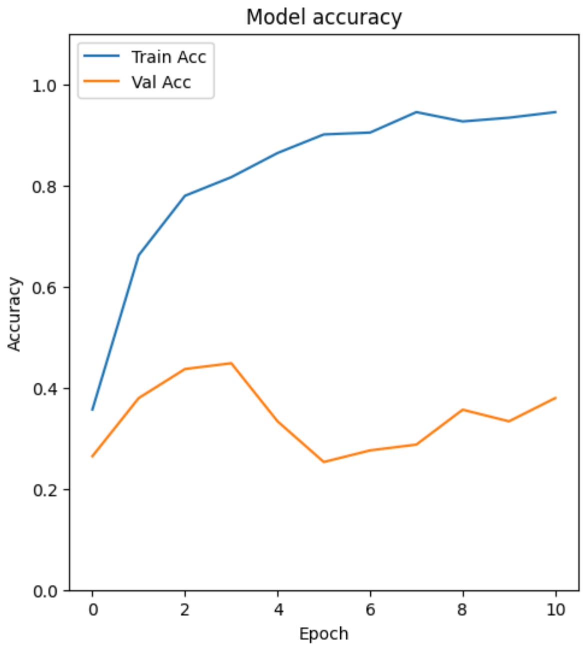


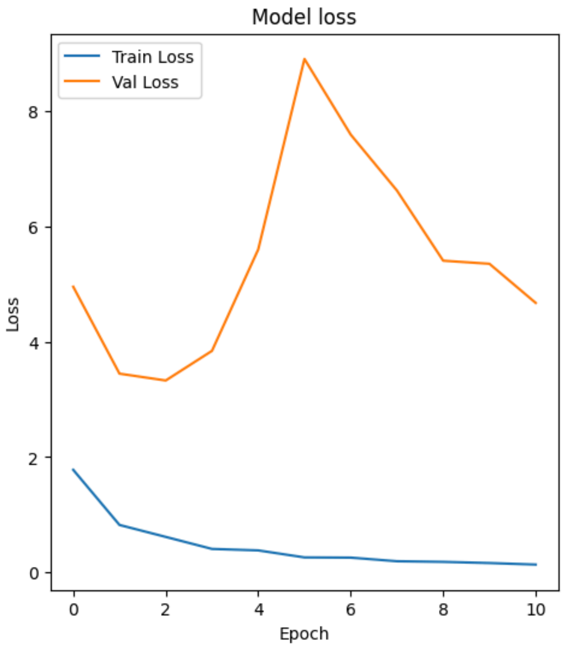
```
[Threshold Optimization] Best balanced_accuracy: 0.6220 at threshold=0.101
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

```
           0     0.8448    0.5632    0.6759        87
           1     0.0732    0.2500    0.1132        12

    accuracy                         0.5253        99
   macro avg     0.4590    0.4066    0.3945        99
weighted avg     0.7513    0.5253    0.6077        99

Balanced Accuracy: 0.40660919540229884
```

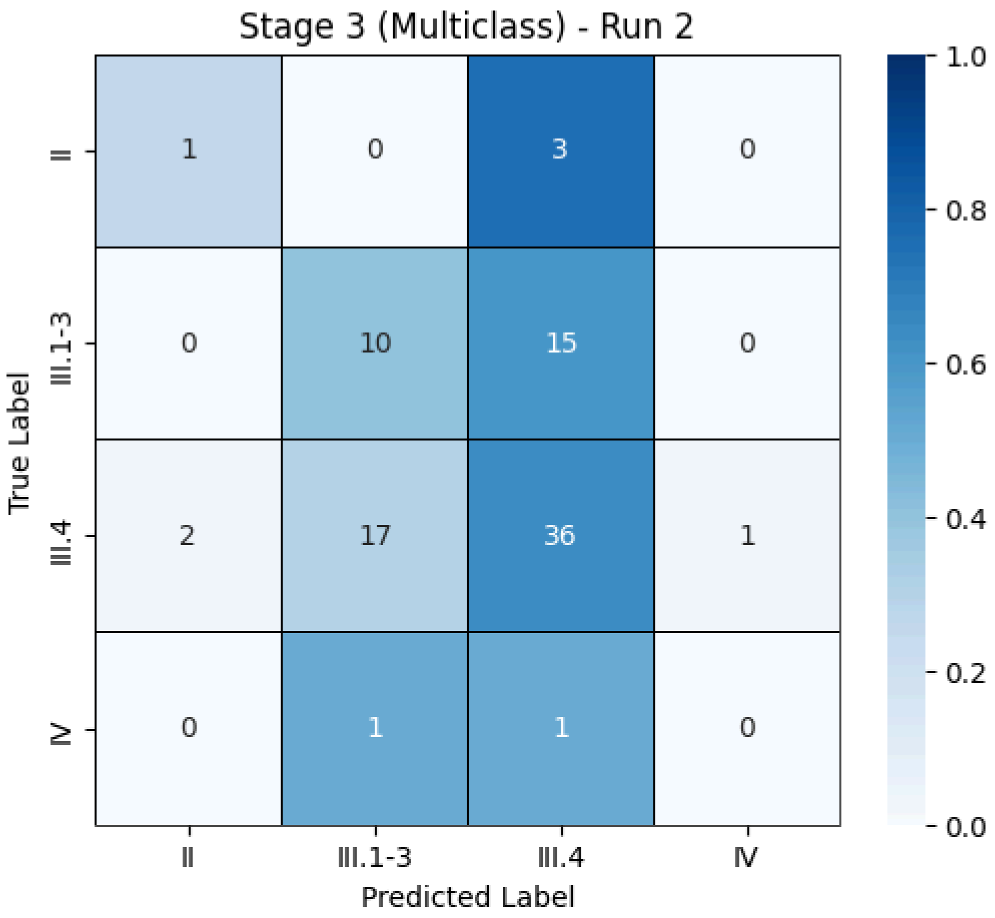


-- Stage 3 (Multiclass) --

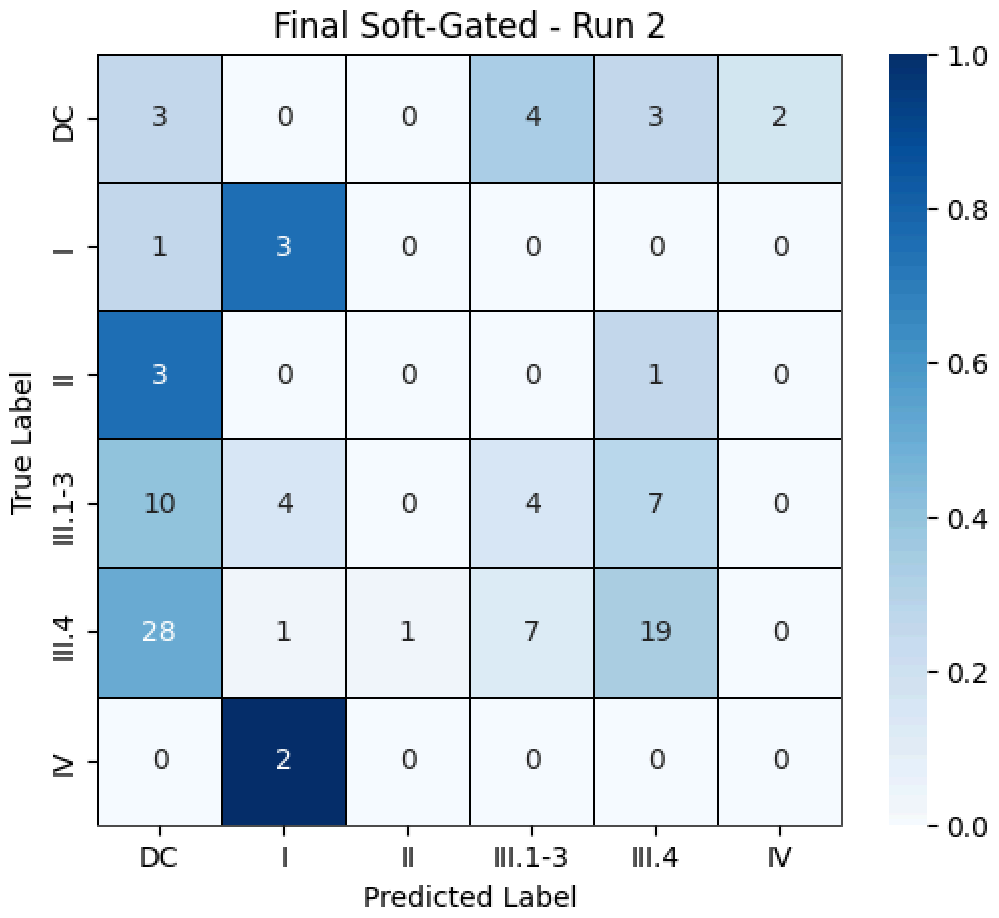




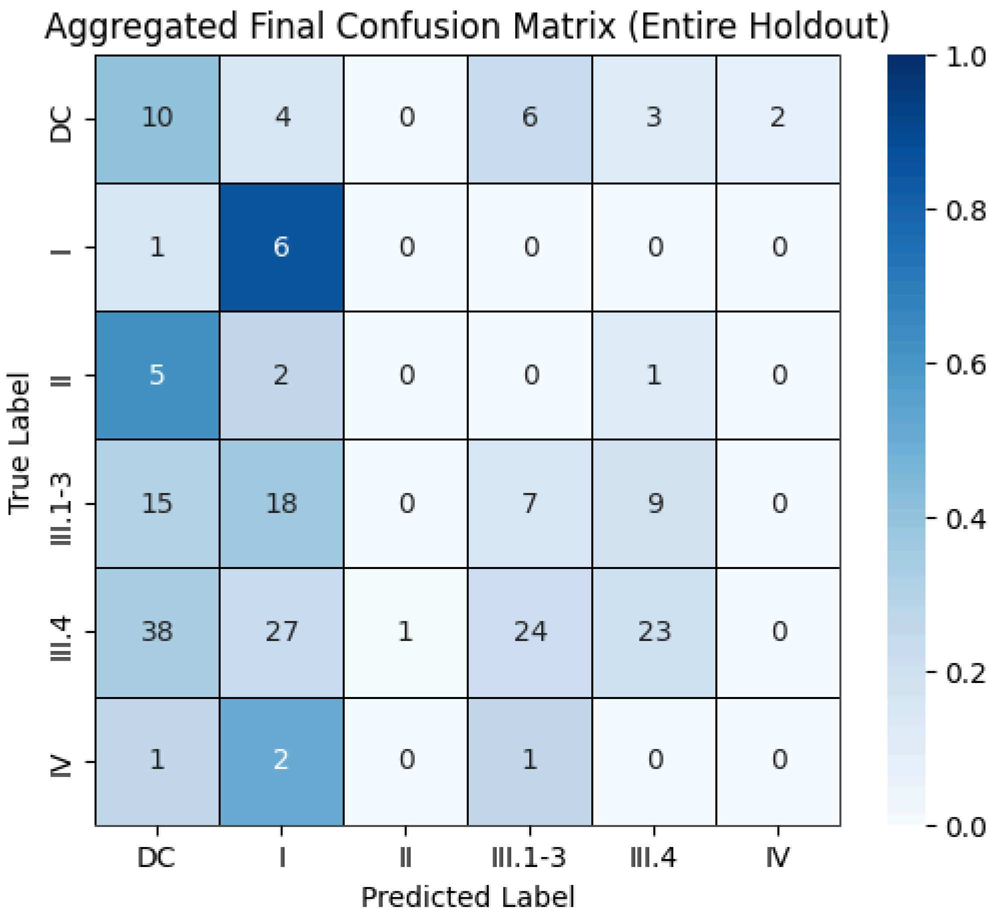
```
       precision    recall  f1-score   support

     0    0.3333    0.2500    0.2857         4
     1    0.3571    0.4000    0.3774        25
     2    0.6545    0.6429    0.6486        56
     3    0.0000    0.0000    0.0000         2

  accuracy                      0.5402        87
 macro avg    0.3363    0.3232    0.3279        87
weighted avg    0.5393    0.5402    0.5391        87

Balanced Accuracy: 0.32321428571428573
```



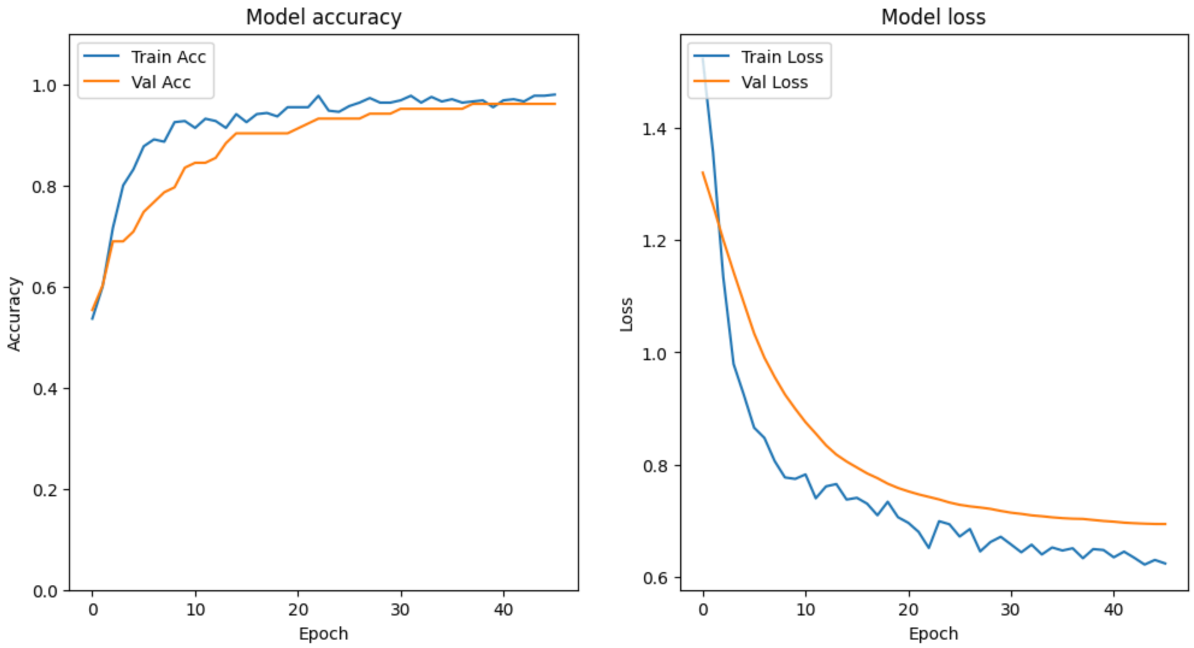
```
== Soft-Gated Overall (Test Set) - Run 2 ==
              precision    recall  f1-score   support

          DC     0.0667    0.2500    0.1053        12
           I     0.3000    0.7500    0.4286         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.2667    0.1600    0.2000        25
       III.4     0.6333    0.3393    0.4419        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.2816       103
   macro avg     0.2111    0.2499    0.1959       103
weighted avg     0.4285    0.2816    0.3177       103

Balanced Accuracy: 0.24988095238095234
```



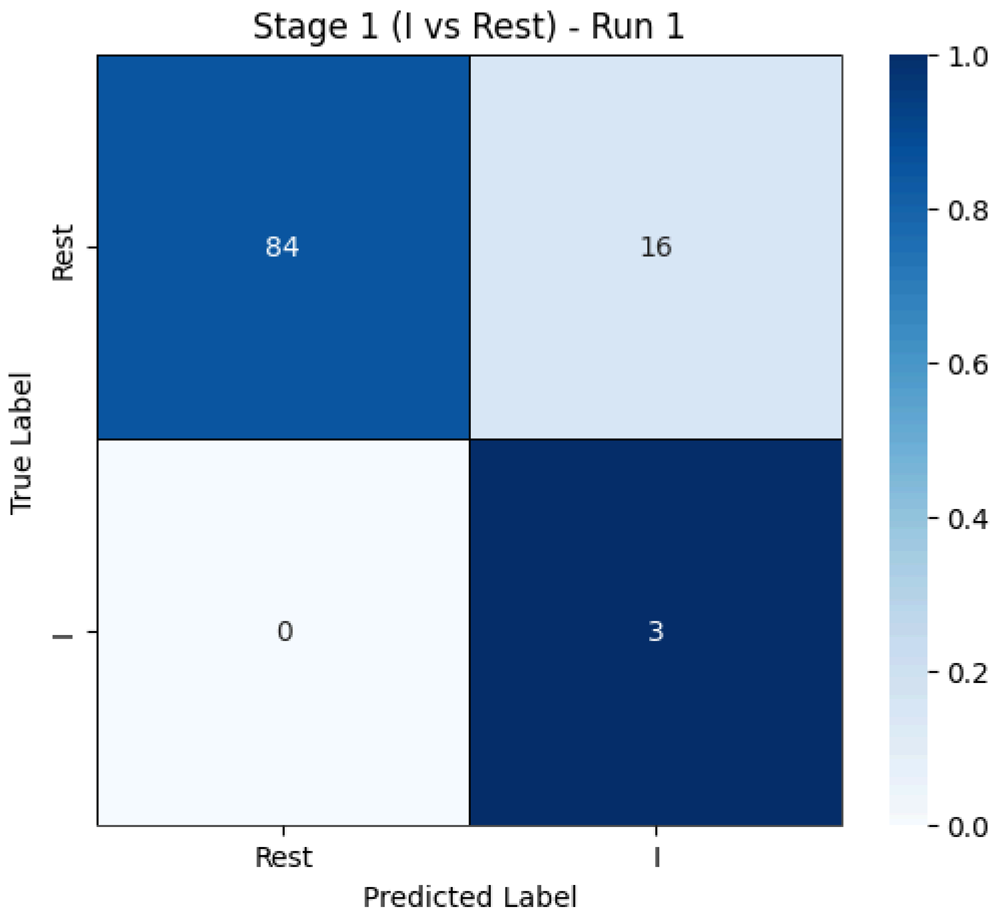
```
==============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
==============================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.1429    0.4000    0.2105        25
           I     0.1017    0.8571    0.1818         7
          II     0.0000    0.0000    0.0000         8
     III.1-3     0.1842    0.1429    0.1609        49
       III.4     0.6389    0.2035    0.3087       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.2233       206
   macro avg     0.1779    0.2673    0.1437       206
weighted avg     0.4151    0.2233    0.2394       206


Aggregated Balanced Accuracy: 0.2673
```



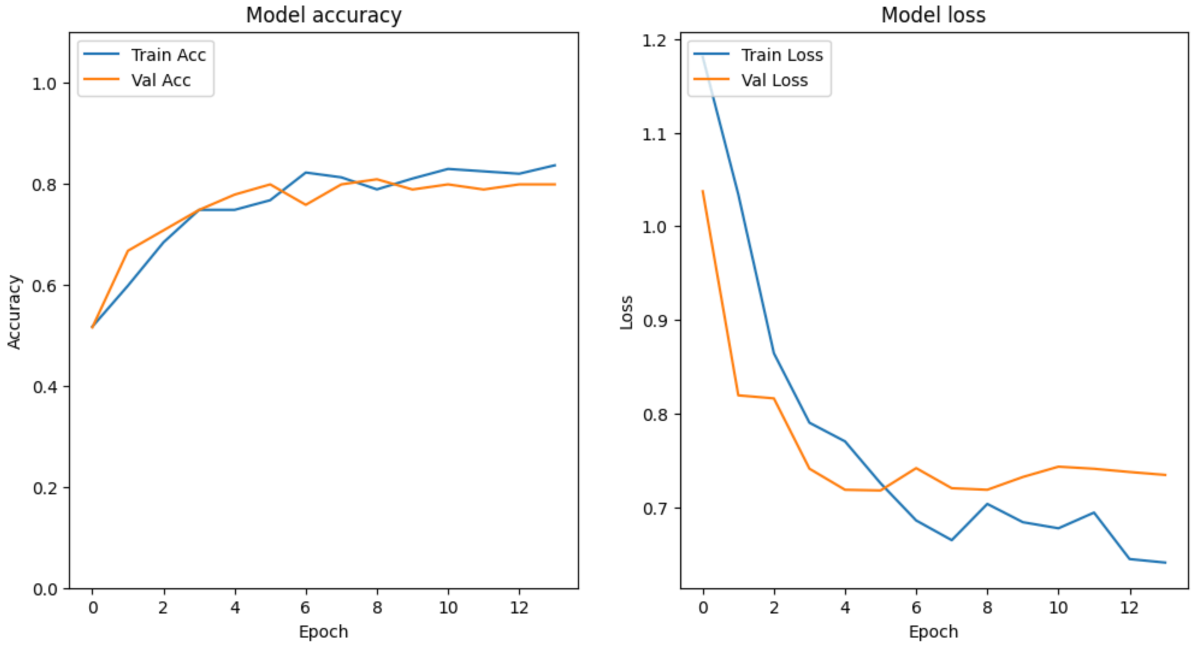
```
== Average Stage Balanced Accuracies ==
Stage 1 (I vs Rest): 0.8898
Stage 2 (DC vs Rest): 0.5267
Stage 3 (Multiclass): 0.3379
Final (Soft-Gated): 0.2694

============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.8737 at threshold=0.080
Optimal threshold (Stage 1 (I vs Rest)): 0.080
              precision    recall  f1-score   support

           0     1.0000    0.8400    0.9130       100
           1     0.1579    1.0000    0.2727         3

    accuracy                         0.8447       103
   macro avg     0.5789    0.9200    0.5929       103
weighted avg     0.9755    0.8447    0.8944       103

Balanced Accuracy: 0.9199999999999999
```

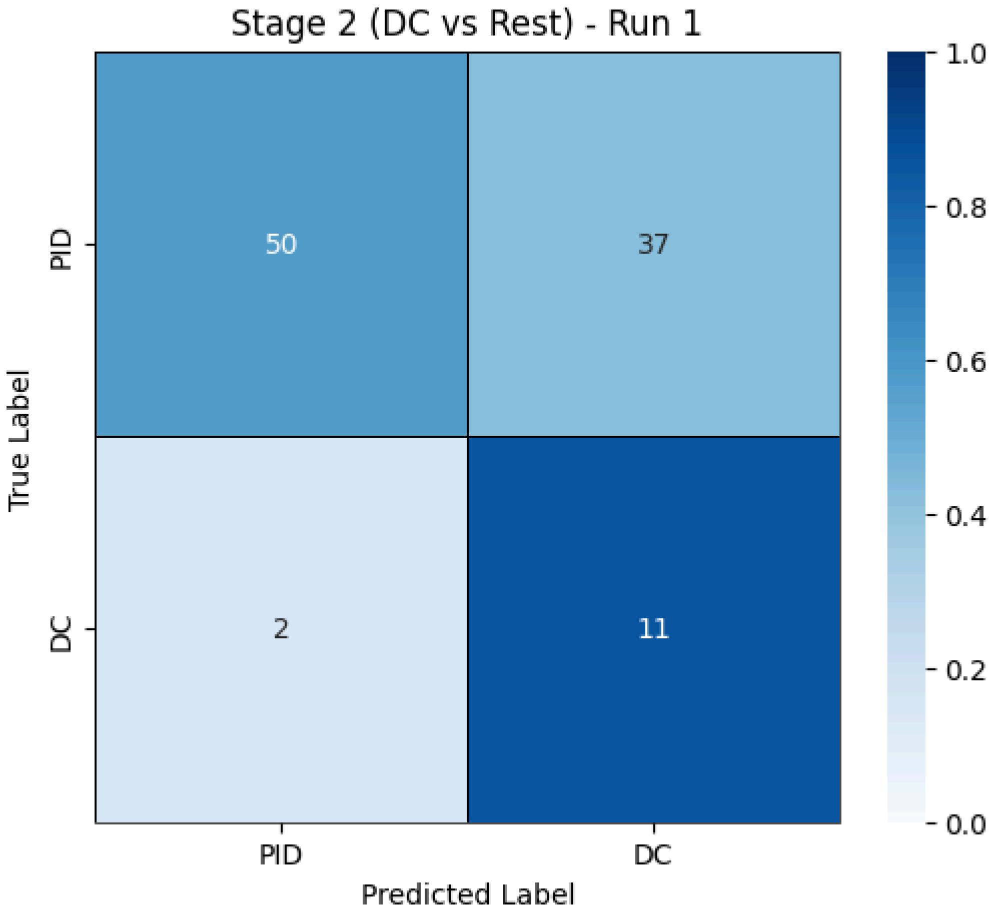


```
-- Stage 2 (DC vs Rest) --
```

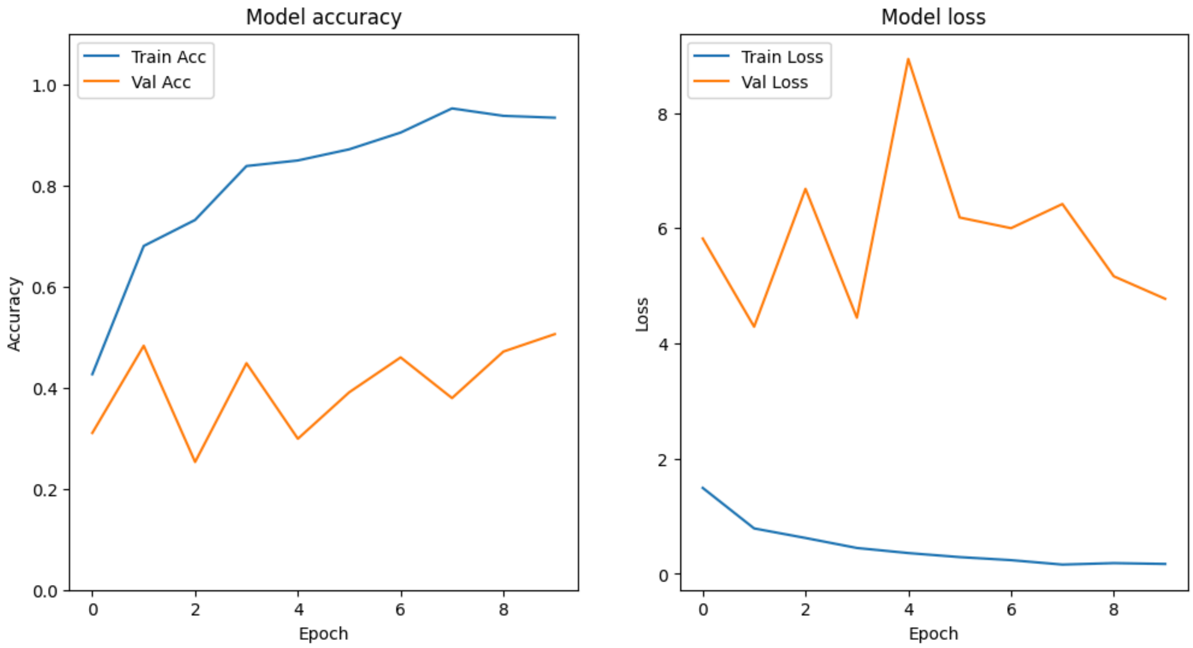


```
[Threshold Optimization] Best balanced_accuracy: 0.6351 at threshold=0.101
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

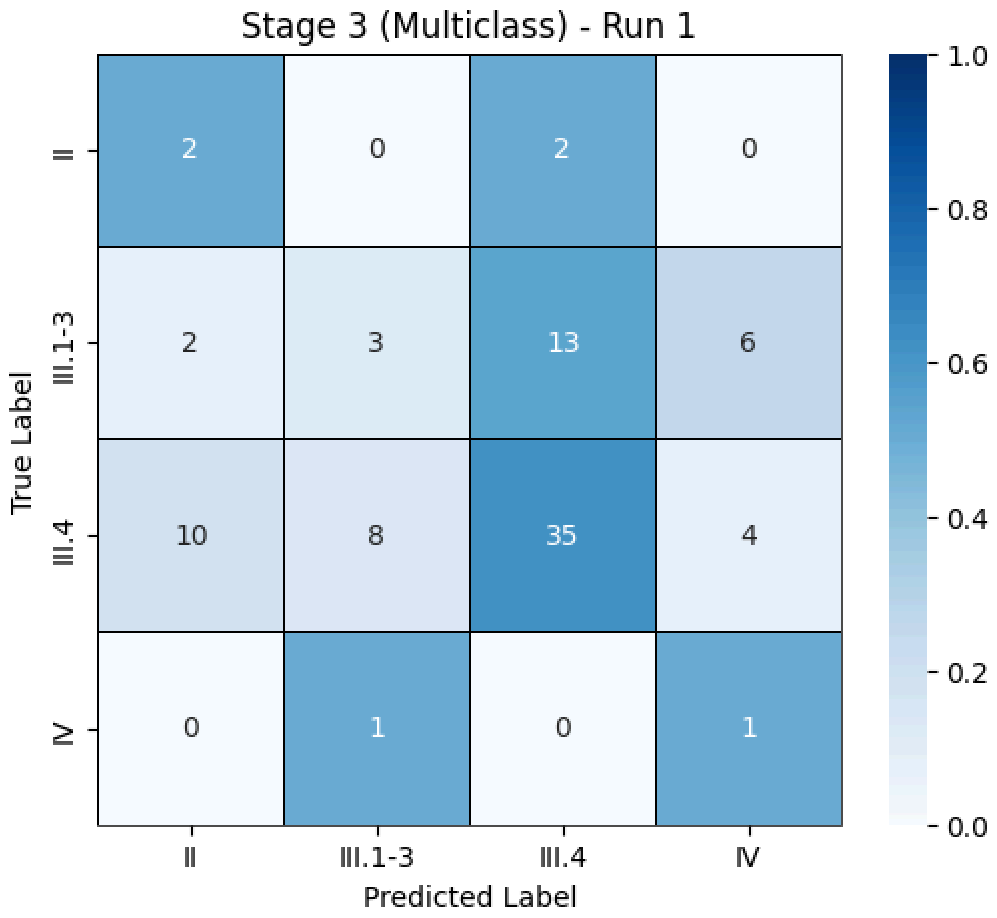
```
           0      0.9615    0.5747    0.7194        87
           1      0.2292    0.8462    0.3607        13

    accuracy                          0.6100       100
   macro avg      0.5954    0.7104    0.5400       100
weighted avg      0.8663    0.6100    0.6728       100

Balanced Accuracy: 0.7104332449160036
```



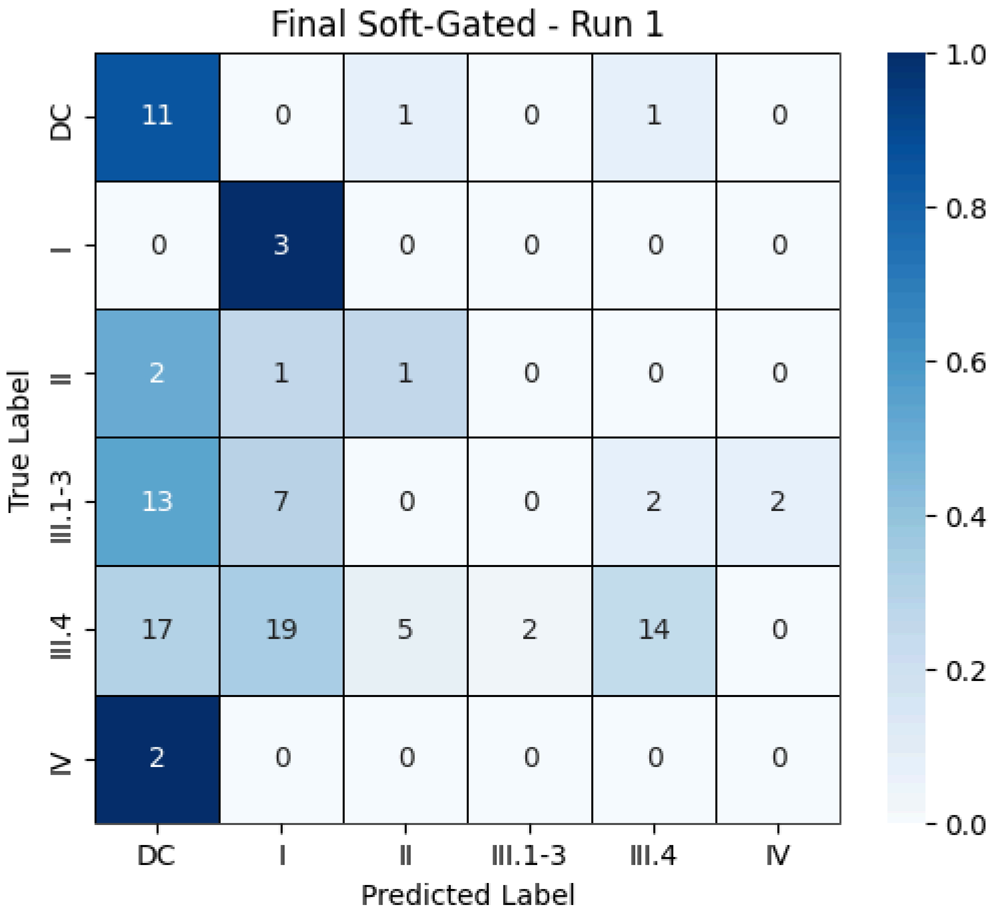
```
-- Stage 3 (Multiclass) --
```



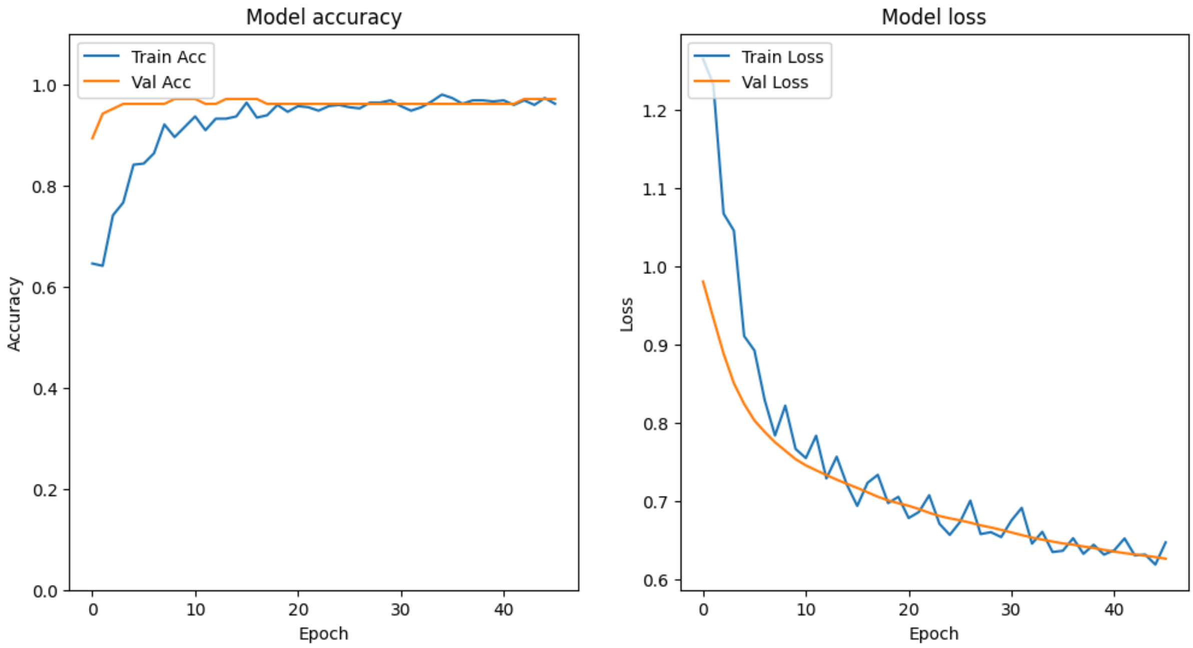
```
              precision    recall  f1-score   support

           0     0.1429    0.5000    0.2222         4
           1     0.2500    0.1250    0.1667        24
           2     0.7000    0.6140    0.6542        57
           3     0.0909    0.5000    0.1538         2

    accuracy                         0.4713        87
   macro avg     0.2959    0.4348    0.2992        87
weighted avg     0.5362    0.4713    0.4883        87

Balanced Accuracy: 0.4347587719298246
```



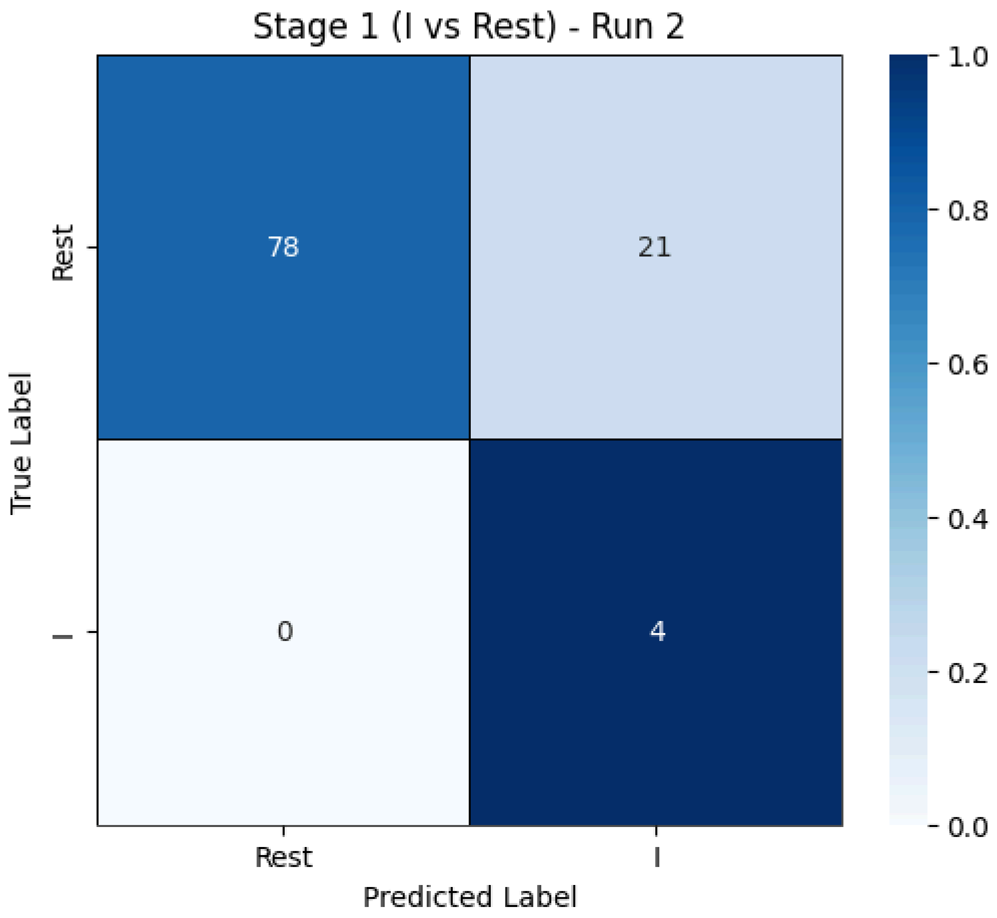
```
== Soft-Gated Overall (Test Set) - Run 1 ==
              precision    recall  f1-score   support

          DC     0.2444    0.8462    0.3793        13
           I     0.1000    1.0000    0.1818         3
          II     0.1429    0.2500    0.1818         4
     III.1-3     0.0000    0.0000    0.0000        24
       III.4     0.8235    0.2456    0.3784        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.2816       103
   macro avg     0.2185    0.3903    0.1869       103
weighted avg     0.4951    0.2816    0.2696       103

Balanced Accuracy: 0.3902946468735942
```



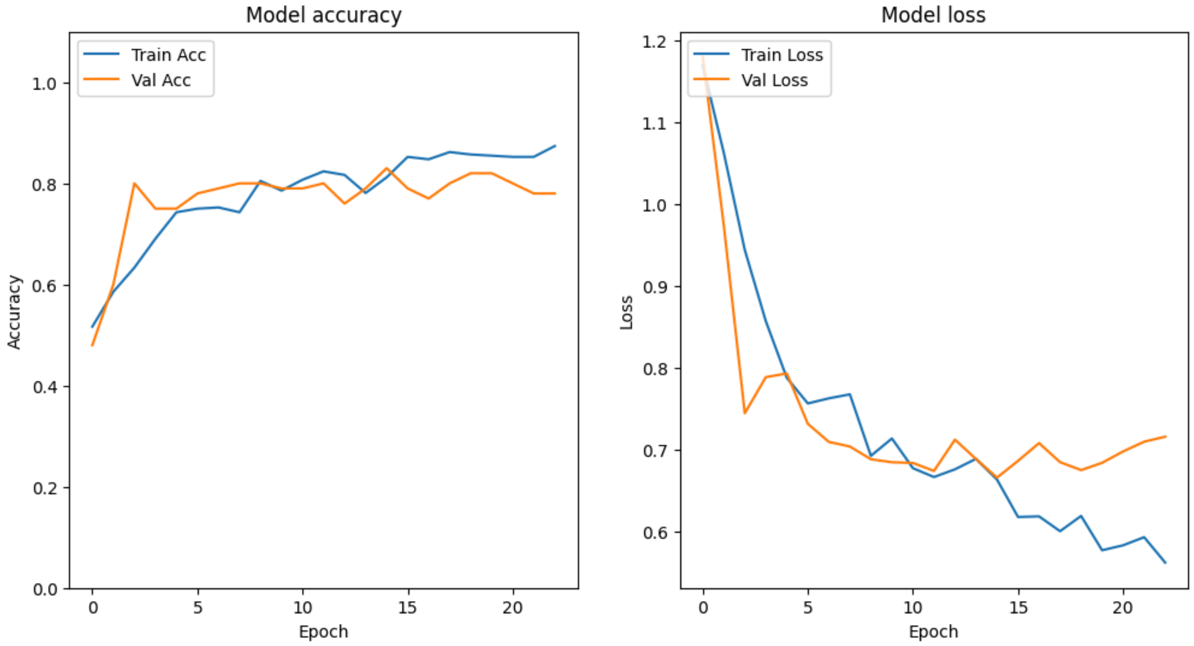
```
============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']

-- Stage 1 (I vs Rest) --
```



```
[Threshold Optimization] Best balanced_accuracy: 0.9550 at threshold=0.151
Optimal threshold (Stage 1 (I vs Rest)): 0.151
              precision    recall  f1-score   support

           0     1.0000    0.7879    0.8814        99
           1     0.1600    1.0000    0.2759         4

    accuracy                         0.7961       103
   macro avg     0.5800    0.8939    0.5786       103
weighted avg     0.9674    0.7961    0.8578       103

Balanced Accuracy: 0.8939393939393939
```

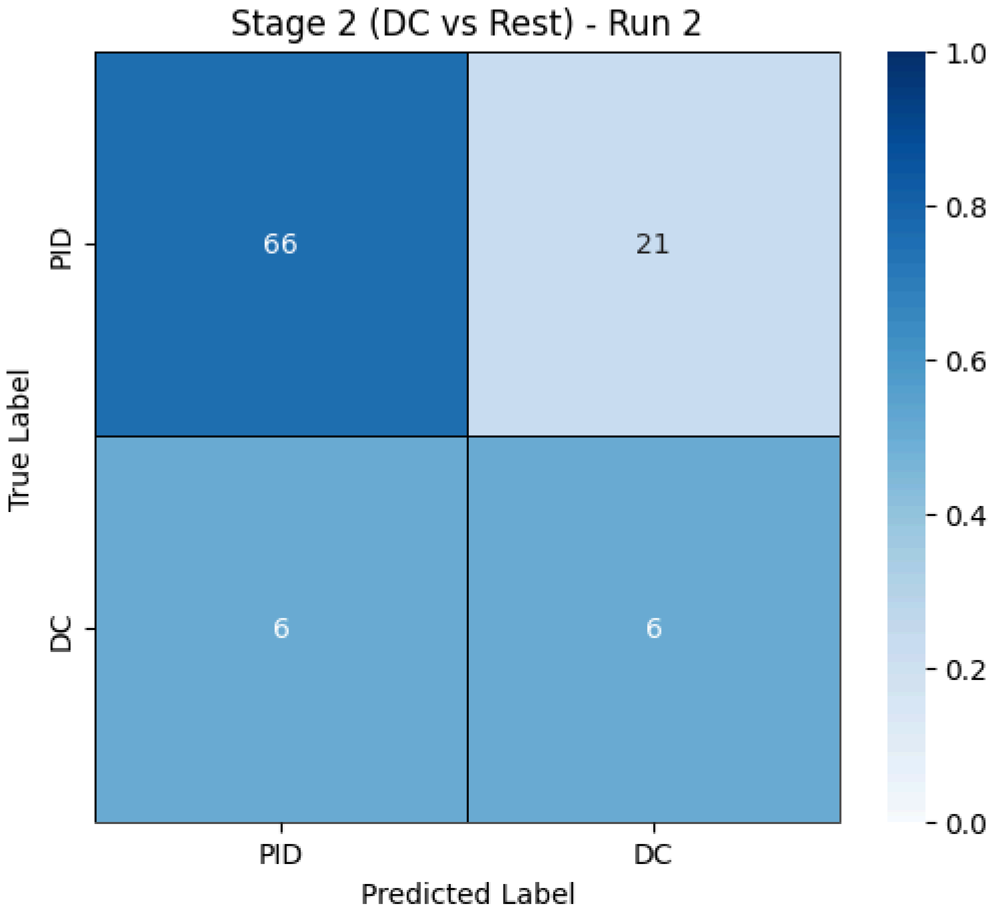


```
-- Stage 2 (DC vs Rest) --
```

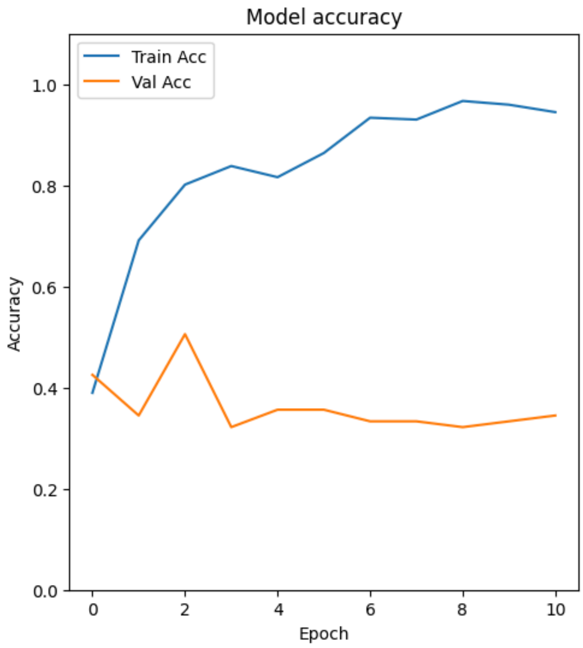


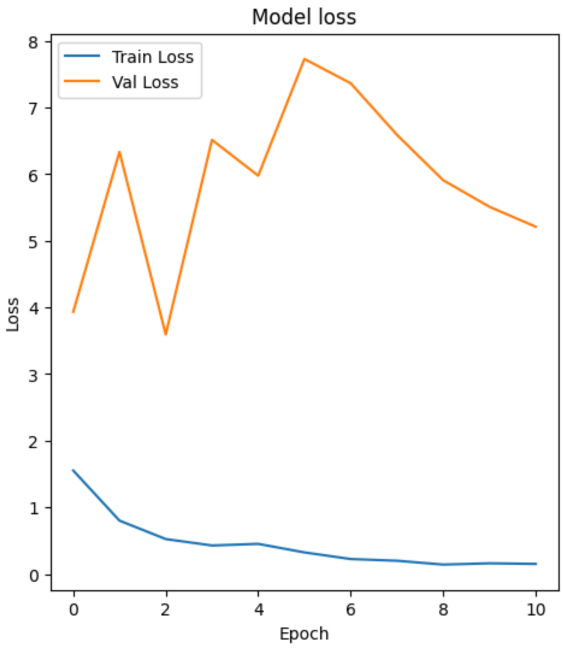
```
[Threshold Optimization] Best balanced_accuracy: 0.7776 at threshold=0.146
Optimal threshold (Stage 2 (DC vs Rest)): 0.200
              precision    recall  f1-score   support
```

```
           0     0.9167    0.7586    0.8302        87
           1     0.2222    0.5000    0.3077        12

    accuracy                         0.7273        99
   macro avg     0.5694    0.6293    0.5689        99
weighted avg     0.8325    0.7273    0.7669        99

Balanced Accuracy: 0.6293103448275862
```

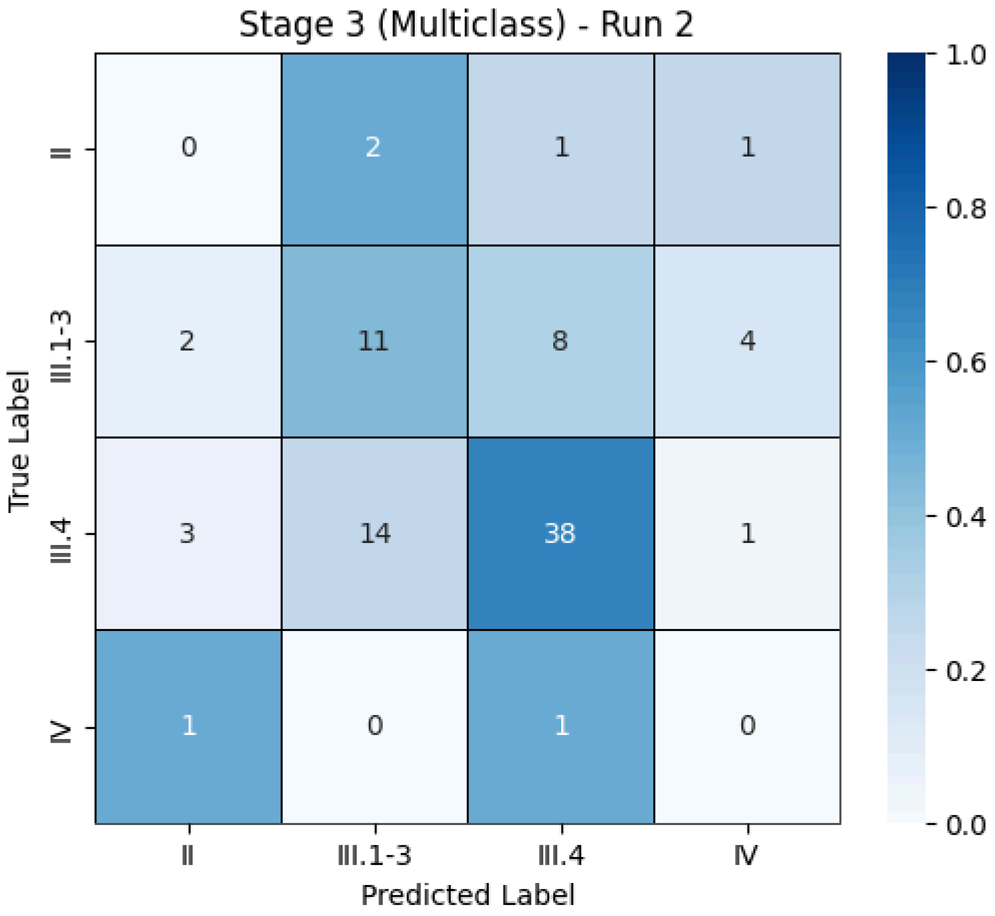


```
-- Stage 3 (Multiclass) --
```

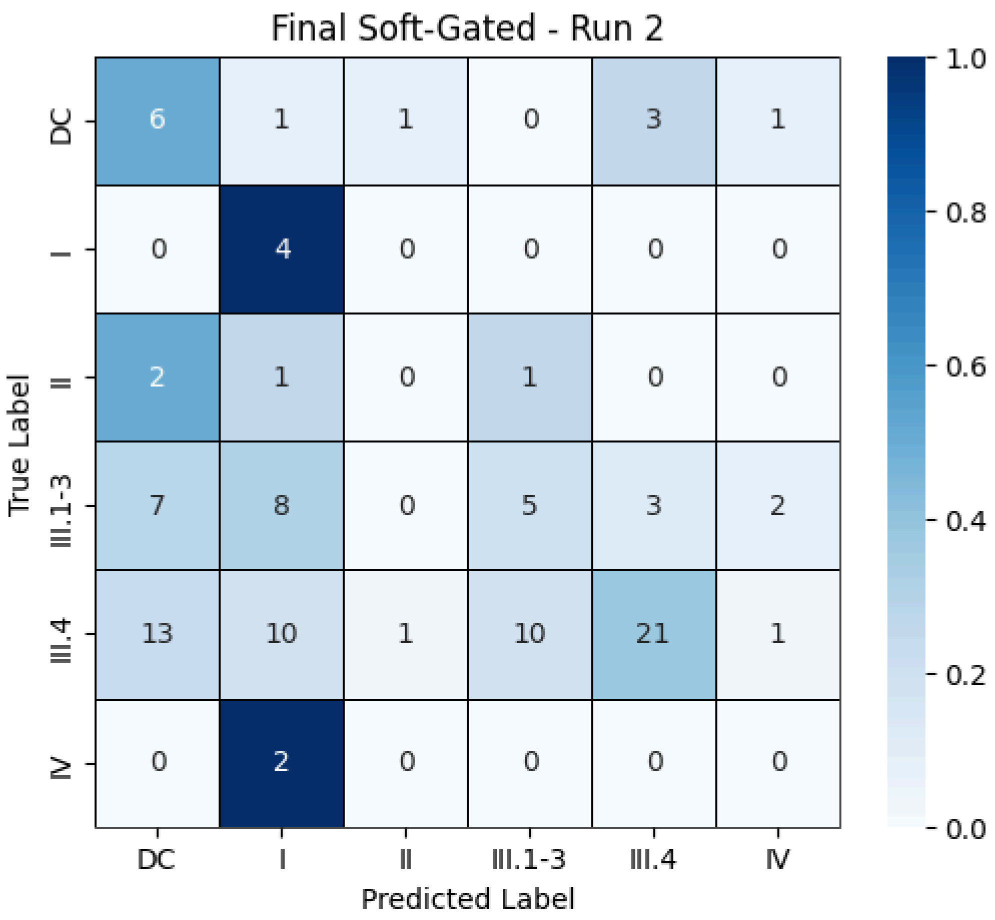




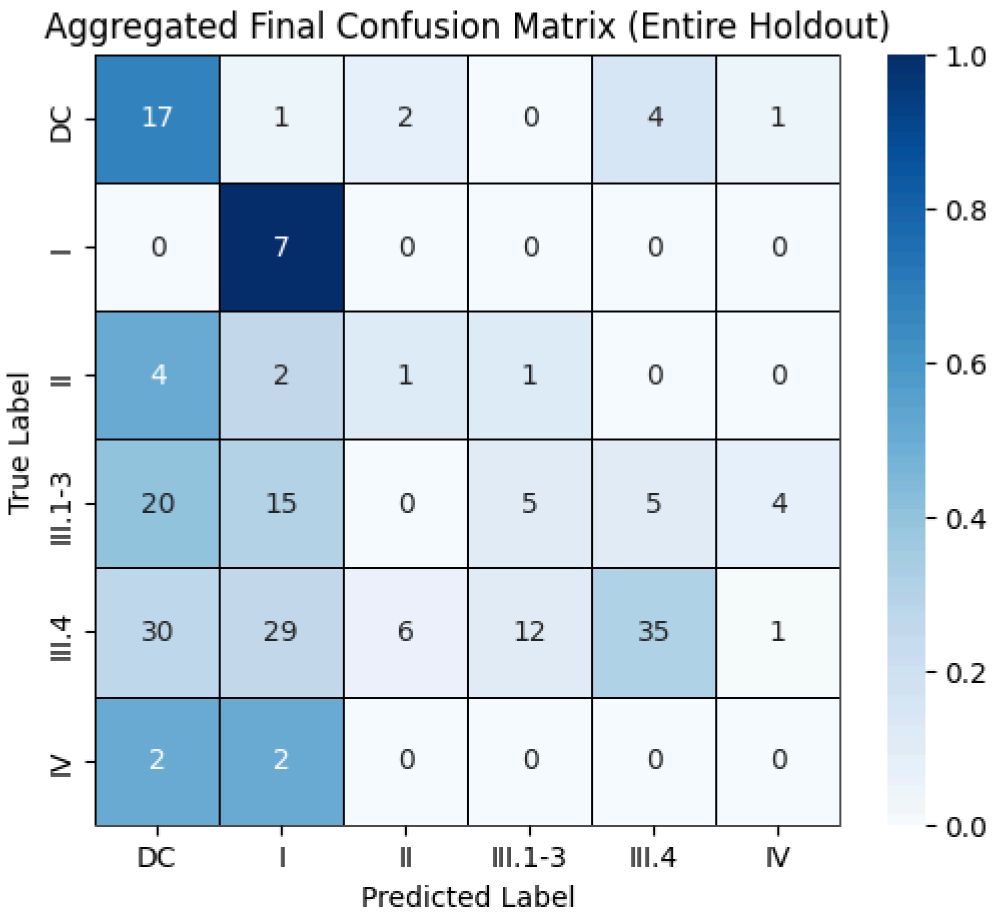
```
       precision    recall  f1-score   support

   0     0.0000    0.0000    0.0000         4
   1     0.4074    0.4400    0.4231        25
   2     0.7917    0.6786    0.7308        56
   3     0.0000    0.0000    0.0000         2

 accuracy                    0.5632        87
macro avg     0.2998    0.2796    0.2885        87
weighted avg     0.6266    0.5632    0.5920        87

Balanced Accuracy: 0.27964285714285714
```



```
== Soft-Gated Overall (Test Set) - Run 2 ==
           precision    recall  f1-score   support

        DC     0.2143    0.5000    0.3000        12
         I     0.1538    1.0000    0.2667         4
        II     0.0000    0.0000    0.0000         4
   III.1-3     0.3125    0.2000    0.2439        25
     III.4     0.7778    0.3750    0.5060        56
        IV     0.0000    0.0000    0.0000         2

  accuracy                         0.3495       103
 macro avg     0.2431    0.3458    0.2194       103
weighted avg     0.5297    0.3495    0.3796       103

Balanced Accuracy: 0.3458333333333334
```

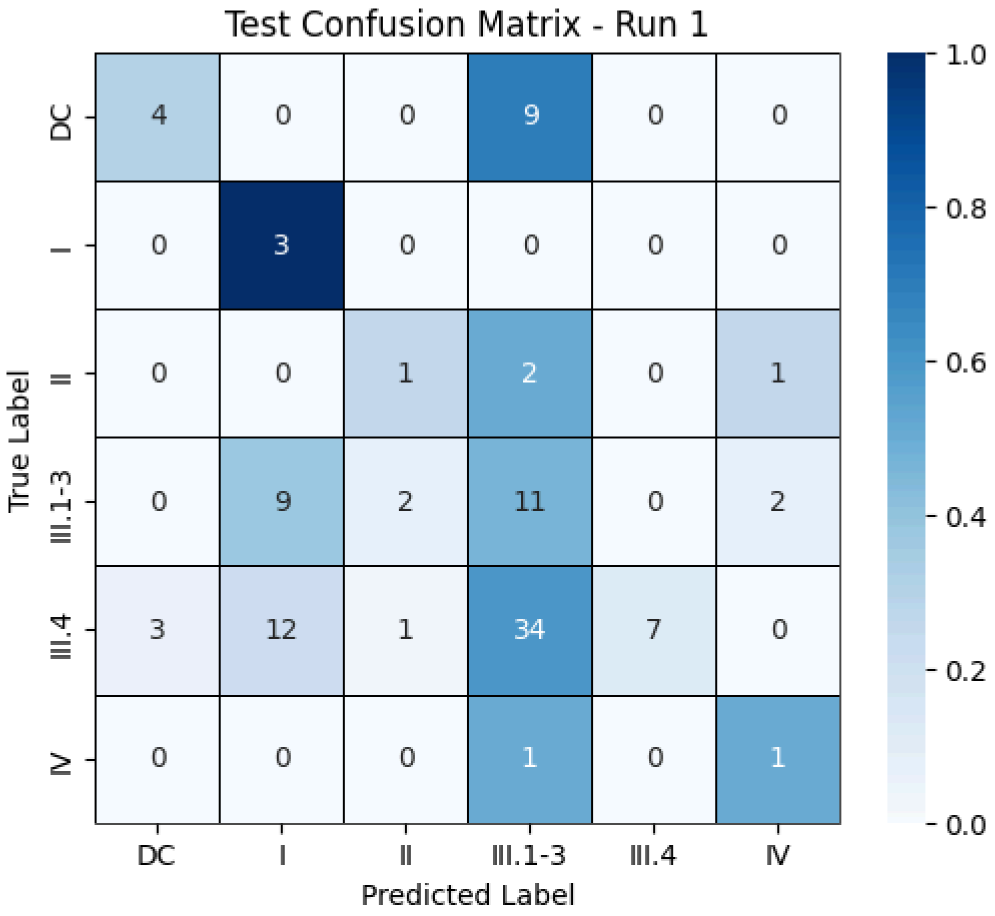


```
============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
============================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.2329    0.6800    0.3469        25
           I     0.1250    1.0000    0.2222         7
          II     0.1111    0.1250    0.1176         8
     III.1-3     0.2778    0.1020    0.1493        49
       III.4     0.7955    0.3097    0.4459       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.3155       206
   macro avg     0.2570    0.3695    0.2137       206
weighted avg     0.5392    0.3155    0.3343       206


Aggregated Balanced Accuracy: 0.3695
```



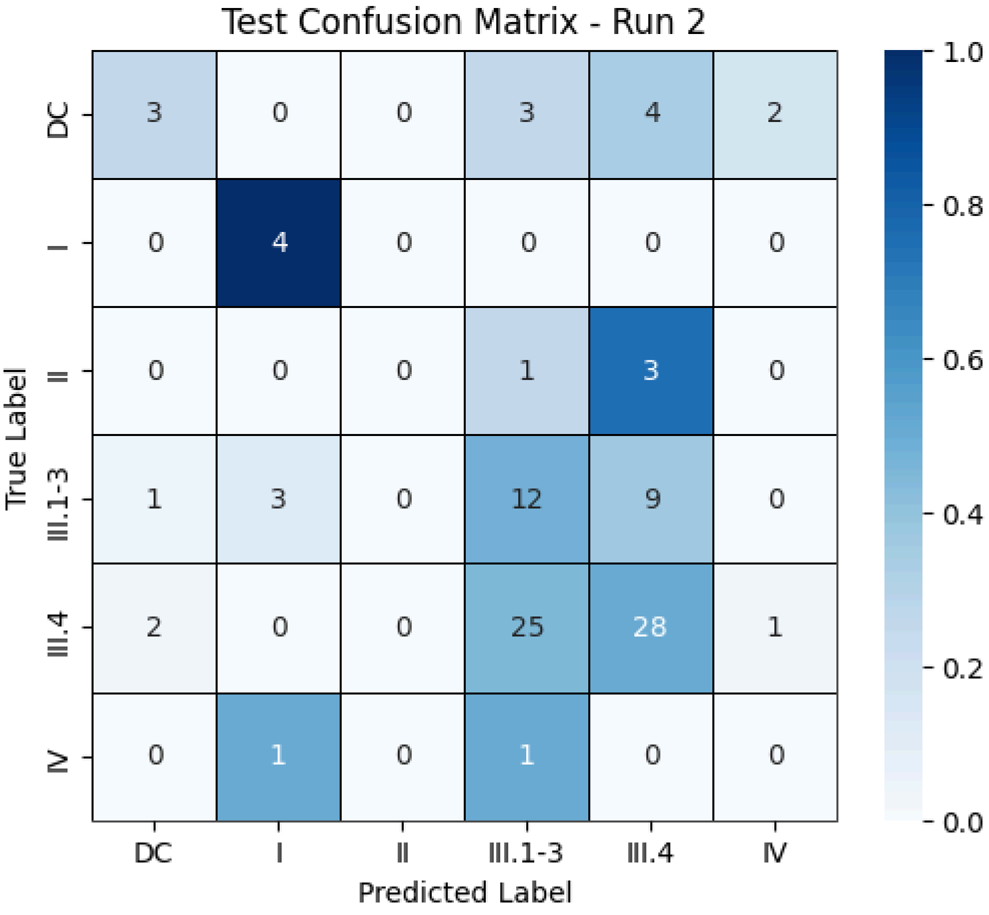
```
== Average Stage Balanced Accuracies ==
Stage 1 (I vs Rest): 0.9070
Stage 2 (DC vs Rest): 0.6699
Stage 3 (Multiclass): 0.3572
Final (Soft-Gated): 0.3681

[0.8673232323232323, 0.5627763041556144, 0.2800031328320802, 0.24578171091445425]
[0.861919191919192, 0.5250884173297966, 0.36178258145363407, 0.2895755824453675]
[0.8722474747474747, 0.5644341290893015, 0.30152882205513787, 0.3070182409246885]
[0.8898484848484849, 0.5266909814323607, 0.33786810776942355, 0.2672566371681416]
[0.906969696969697, 0.6698717948717949, 0.3572008145363409, 0.3694625549334778]

0.880 ± 0.007
0.570 ± 0.024
0.328 ± 0.014
0.296 ± 0.019
```

```
FLAT
============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step

== Test Set Evaluation - Run 1 ==
              precision    recall  f1-score   support

          DC     0.5714    0.3077    0.4000        13
           I     0.1250    1.0000    0.2222         3
          II     0.2500    0.2500    0.2500         4
     III.1-3     0.1930    0.4583    0.2716        24
       III.4     1.0000    0.1228    0.2188        57
          IV     0.2500    0.5000    0.3333         2

    accuracy                         0.2621       103
   macro avg     0.3982    0.4398    0.2827       103
weighted avg     0.6887    0.2621    0.2575       103

Balanced Accuracy: 0.43980544309491676
Confusion Matrix:
 [[ 4  0  0  9  0  0]
 [ 0  3  0  0  0  0]
 [ 0  0  1  2  0  1]
 [ 0  9  2 11  0  2]
 [ 3 12  1 34  7  0]
 [ 0  0  0  1  0  1]]
```



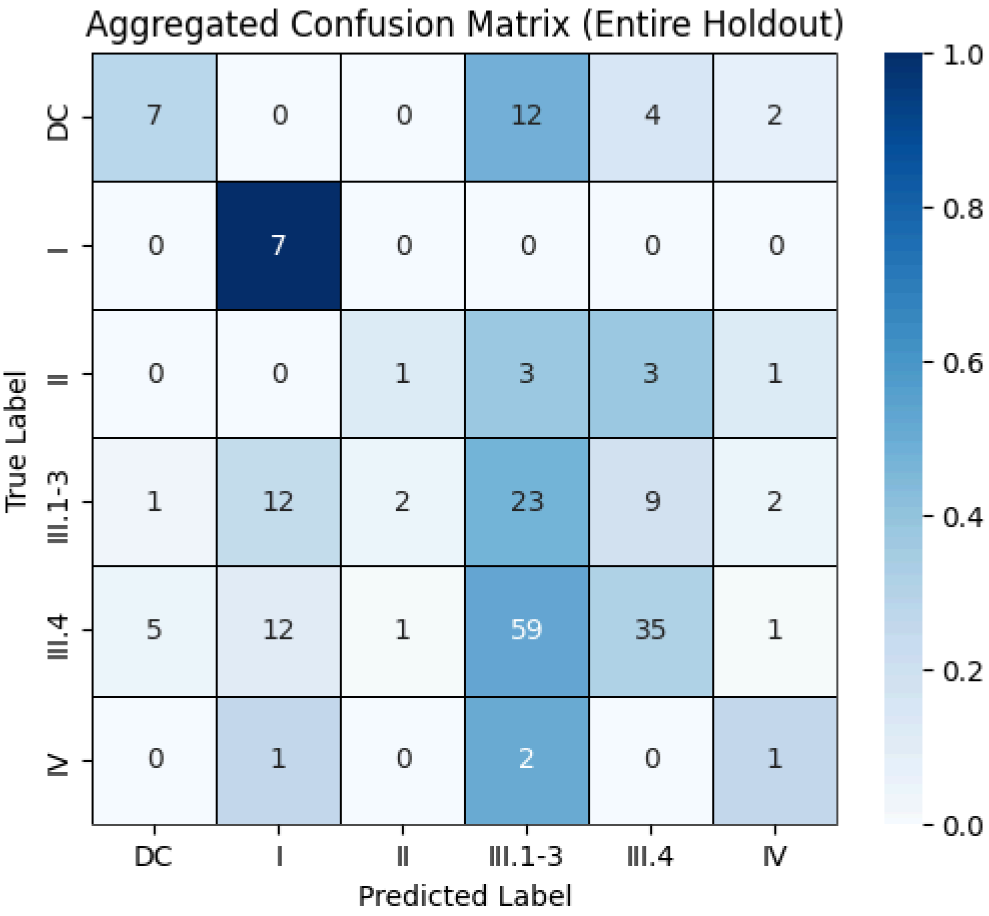
```
============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 20ms/step

== Test Set Evaluation - Run 2 ==
              precision    recall  f1-score   support

          DC     0.5000    0.2500    0.3333        12
           I     0.5000    1.0000    0.6667         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.2857    0.4800    0.3582        25
       III.4     0.6364    0.5000    0.5600        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.4563       103
   macro avg     0.3203    0.3717    0.3197       103
weighted avg     0.4930    0.4563    0.4561       103

Balanced Accuracy: 0.37166666666666665
```

```
Confusion Matrix:
[[ 3  0  0  3  4  2]
 [ 0  4  0  0  0  0]
 [ 0  0  0  1  3  0]
 [ 1  3  0 12  9  0]
 [ 2  0  0 25 28  1]
 [ 0  1  0  1  0  0]]
```



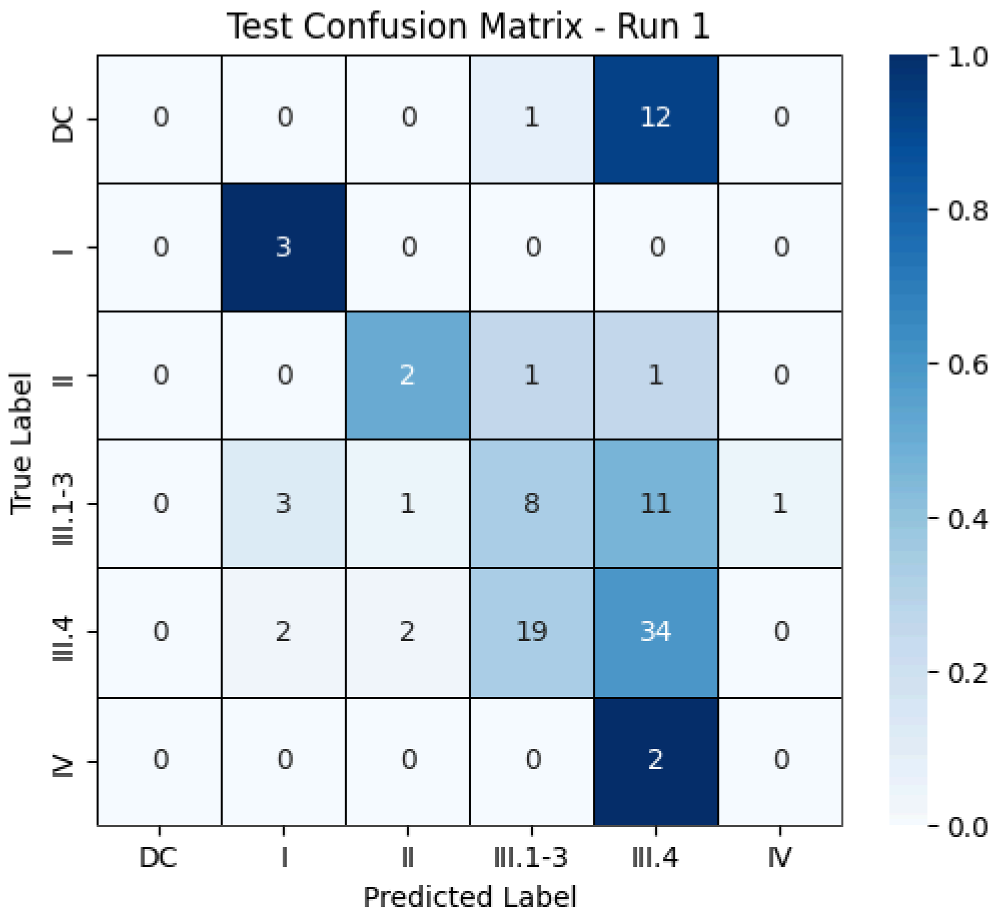
```
============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
============================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.5385    0.2800    0.3684        25
           I     0.2188    1.0000    0.3590         7
          II     0.2500    0.1250    0.1667         8
     III.1-3     0.2323    0.4694    0.3108        49
       III.4     0.6863    0.3097    0.4268       113
          IV     0.1429    0.2500    0.1818         4

    accuracy                         0.3592       206
```

```
   macro avg     0.3448    0.4057    0.3023       206
weighted avg     0.5170    0.3592    0.3750       206


Aggregated Balanced Accuracy: 0.4057
Aggregated Confusion Matrix:
 [[ 7  0  0 12  4  2]
 [ 0  7  0  0  0  0]
 [ 0  0  1  3  3  1]
 [ 1 12  2 23  9  2]
 [ 5 12  1 59 35  1]
 [ 0  1  0  2  0  1]]
```



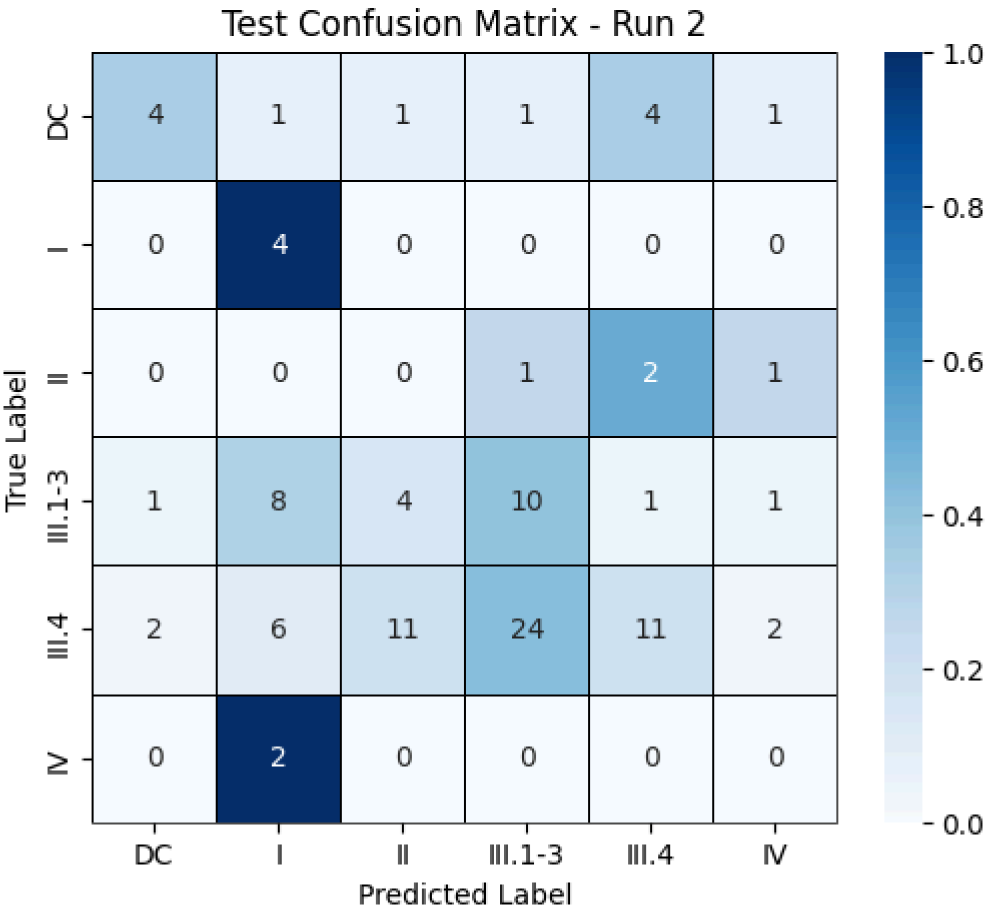
```
== Average Balanced Accuracy Across Runs ==
Run 1: 0.4398
Run 2: 0.3717
Average: 0.4057

============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
```

```
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step

== Test Set Evaluation - Run 1 ==
              precision    recall  f1-score   support

          DC     0.0000    0.0000    0.0000        13
           I     0.3750    1.0000    0.5455         3
          II     0.4000    0.5000    0.4444         4
     III.1-3     0.2759    0.3333    0.3019        24
       III.4     0.5667    0.5965    0.5812        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.4563       103
   macro avg     0.2696    0.4050    0.3122       103
weighted avg     0.4043    0.4563    0.4251       103

Balanced Accuracy: 0.4049707602339181
Confusion Matrix:
 [[ 0  0  0  1 12  0]
 [ 0  3  0  0  0  0]
 [ 0  0  2  1  1  0]
 [ 0  3  1  8 11  1]
 [ 0  2  2 19 34  0]
 [ 0  0  0  0  2  0]]
```



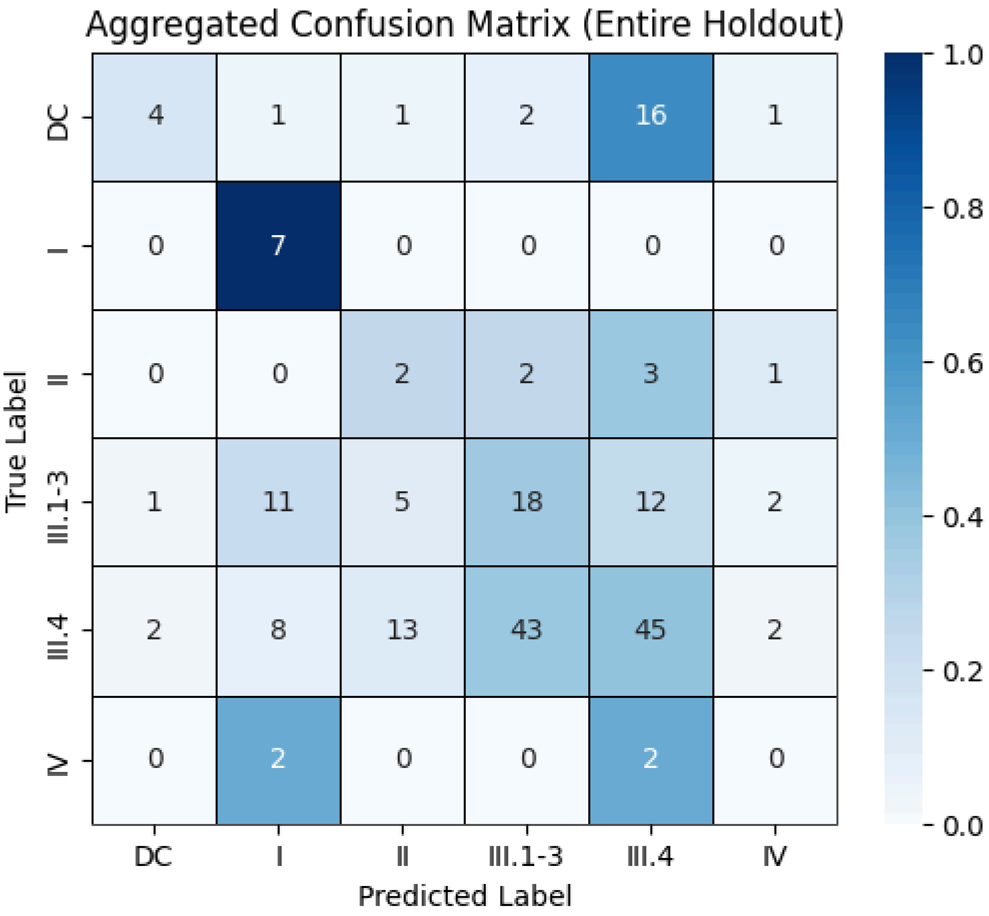
```
================================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
================================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 24ms/step

== Test Set Evaluation - Run 2 ==
              precision    recall  f1-score   support

          DC     0.5714    0.3333    0.4211        12
           I     0.1905    1.0000    0.3200         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.2778    0.4000    0.3279        25
       III.4     0.6111    0.1964    0.2973        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.2816       103
   macro avg     0.2751    0.3216    0.2277       103
weighted avg     0.4736    0.2816    0.3027       103

Balanced Accuracy: 0.32162698412698415
```

```
Confusion Matrix:
 [[ 4  1  1  1  4  1]
 [ 0  4  0  0  0  0]
 [ 0  0  0  1  2  1]
 [ 1  8  4 10  1  1]
 [ 2  6 11 24 11  2]
 [ 0  2  0  0  0  0]]
```



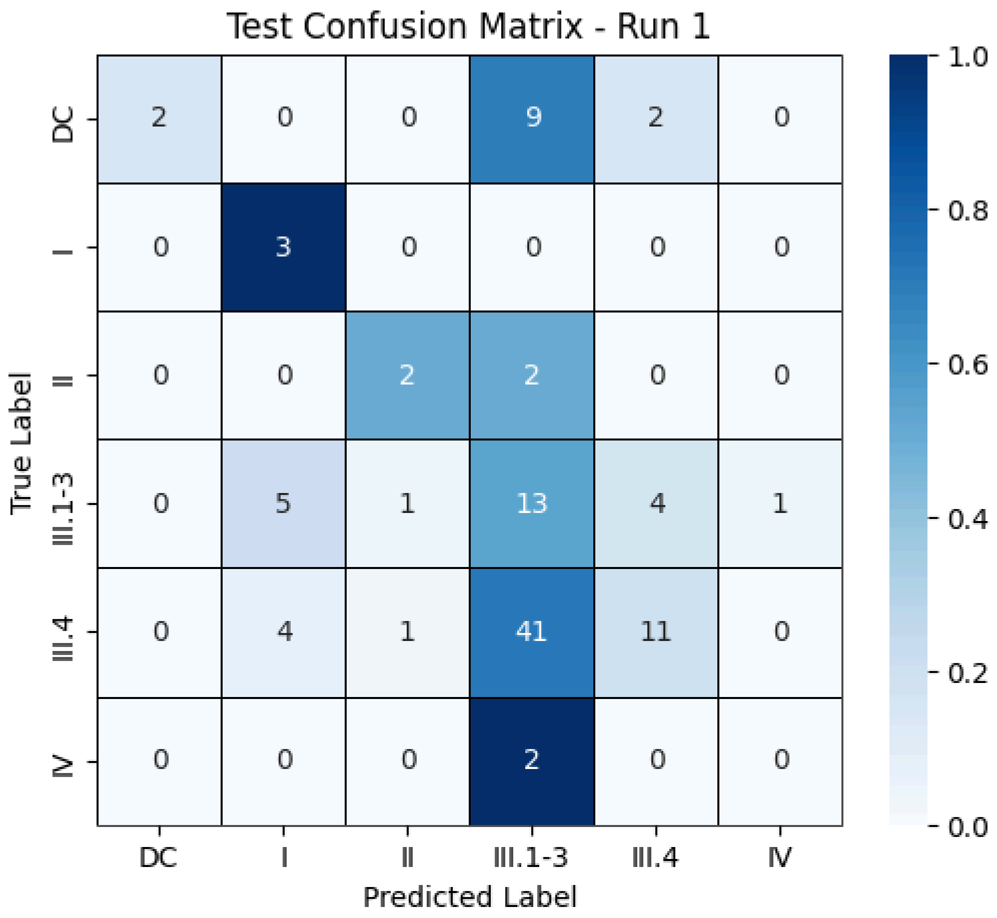
```
============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
============================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.5714    0.1600    0.2500        25
           I     0.2414    1.0000    0.3889         7
          II     0.0952    0.2500    0.1379         8
     III.1-3     0.2769    0.3673    0.3158        49
       III.4     0.5769    0.3982    0.4712       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.3689       206
```

```
   macro avg     0.2936    0.3626    0.2606       206
weighted avg     0.4636    0.3689    0.3825       206


Aggregated Balanced Accuracy: 0.3626
Aggregated Confusion Matrix:
 [[ 4  1  1  2 16  1]
 [ 0  7  0  0  0  0]
 [ 0  0  2  2  3  1]
 [ 1 11  5 18 12  2]
 [ 2  8 13 43 45  2]
 [ 0  2  0  0  2  0]]
```



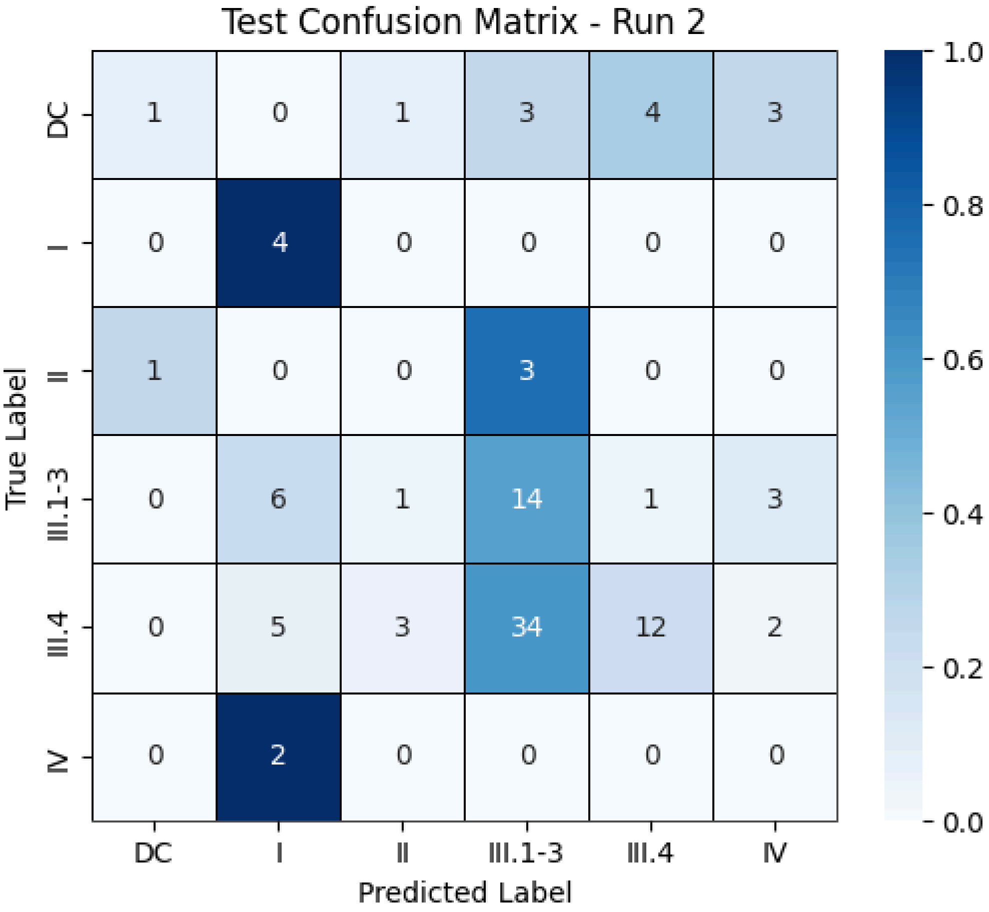
```
== Average Balanced Accuracy Across Runs ==
Run 1: 0.4050
Run 2: 0.3216
Average: 0.3633

============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
```

```
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step

== Test Set Evaluation - Run 1 ==
              precision    recall  f1-score   support

          DC     1.0000    0.1538    0.2667        13
           I     0.2500    1.0000    0.4000         3
          II     0.5000    0.5000    0.5000         4
     III.1-3     0.1940    0.5417    0.2857        24
       III.4     0.6471    0.1930    0.2973        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.3010       103
   macro avg     0.4318    0.3981    0.2916       103
weighted avg     0.5562    0.3010    0.2958       103

Balanced Accuracy: 0.3980825461088619
Confusion Matrix:
 [[ 2  0  0  9  2  0]
 [ 0  3  0  0  0  0]
 [ 0  0  2  2  0  0]
 [ 0  5  1 13  4  1]
 [ 0  4  1 41 11  0]
 [ 0  0  0  2  0  0]]
```



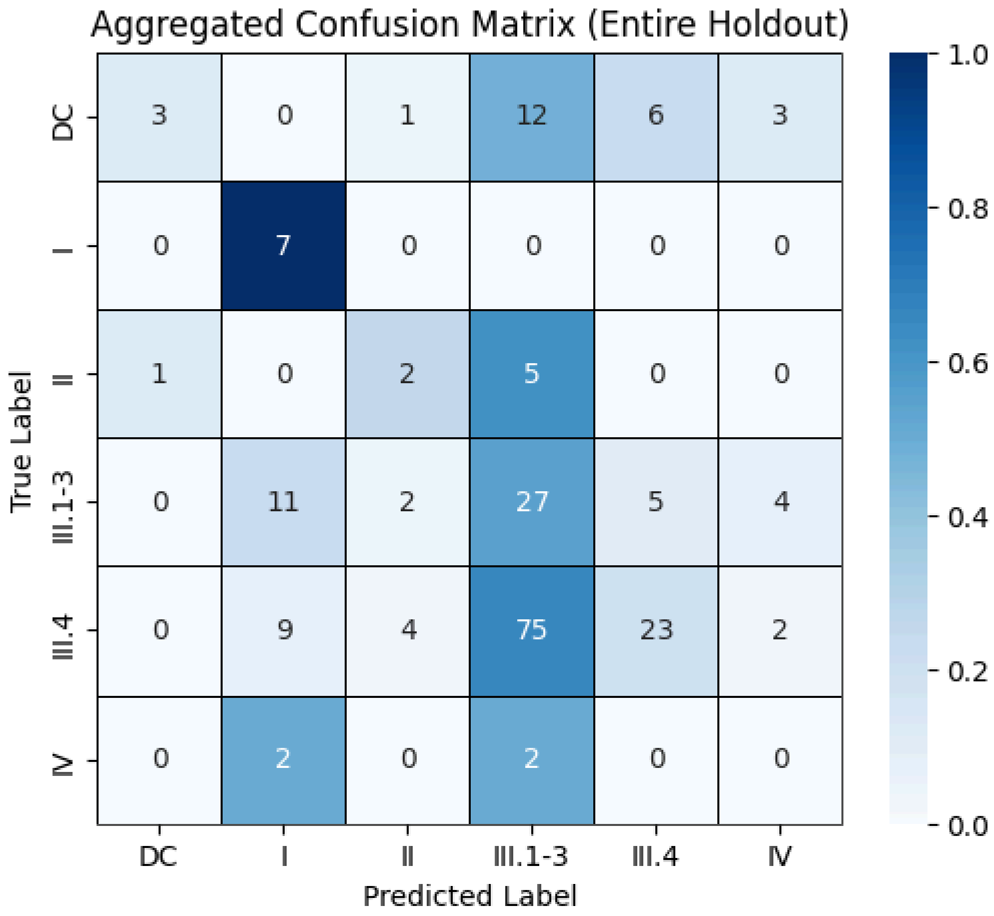
```
=============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
=============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step

== Test Set Evaluation - Run 2 ==
              precision    recall  f1-score   support

          DC     0.5000    0.0833    0.1429        12
           I     0.2353    1.0000    0.3810         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.2593    0.5600    0.3544        25
       III.4     0.7059    0.2143    0.3288        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.3010       103
   macro avg     0.2834    0.3096    0.2012       103
weighted avg     0.5141    0.3010    0.2962       103

Balanced Accuracy: 0.3096031746031746
```

```
Confusion Matrix:
 [[ 1  0  1  3  4  3]
 [ 0  4  0  0  0  0]
 [ 1  0  0  3  0  0]
 [ 0  6  1 14  1  3]
 [ 0  5  3 34 12  2]
 [ 0  2  0  0  0  0]]
```



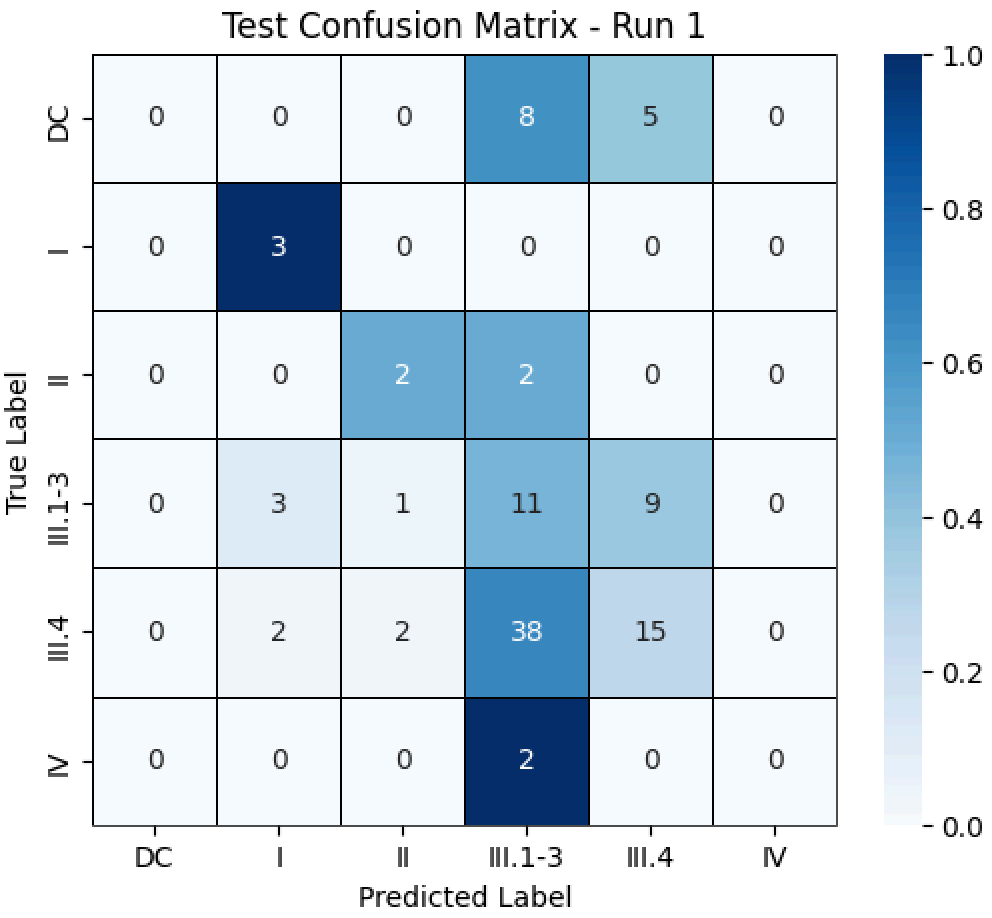
```
============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
============================================================


== Aggregated Classification Report ==
          precision    recall  f1-score   support


      DC     0.7500    0.1200    0.2069        25
       I     0.2414    1.0000    0.3889         7
      II     0.2222    0.2500    0.2353         8
  III.1-3     0.2231    0.5510    0.3176        49
    III.4     0.6765    0.2035    0.3129       113
      IV     0.0000    0.0000    0.0000         4

  accuracy                        0.3010       206
```

```
   macro avg     0.3522    0.3541    0.2436       206
weighted avg     0.5320    0.3010    0.2947       206


Aggregated Balanced Accuracy: 0.3541
Aggregated Confusion Matrix:
 [[ 3  0  1 12  6  3]
 [ 0  7  0  0  0  0]
 [ 1  0  2  5  0  0]
 [ 0 11  2 27  5  4]
 [ 0  9  4 75 23  2]
 [ 0  2  0  2  0  0]]
```



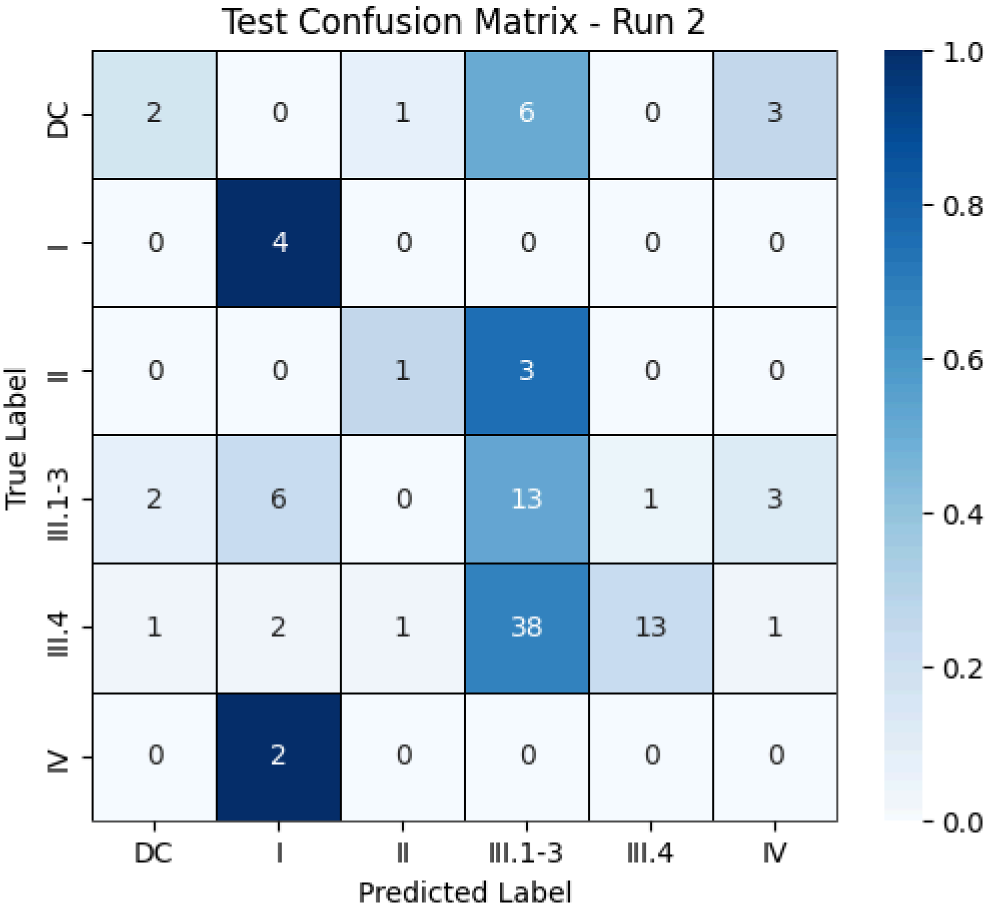
```
== Average Balanced Accuracy Across Runs ==
Run 1: 0.3981
Run 2: 0.3096
Average: 0.3538

==============================================================
RUN 1: Using Split A for validation, Split B for testing
==============================================================
These features will be dropped:
```

```
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 26ms/step

== Test Set Evaluation - Run 1 ==
              precision    recall  f1-score   support

          DC     0.0000    0.0000    0.0000        13
           I     0.3750    1.0000    0.5455         3
          II     0.4000    0.5000    0.4444         4
     III.1-3     0.1803    0.4583    0.2588        24
       III.4     0.5172    0.2632    0.3488        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.3010       103
   macro avg     0.2454    0.3702    0.2663       103
weighted avg     0.3547    0.3010    0.2865       103

Balanced Accuracy: 0.37024853801169594
Confusion Matrix:
 [[ 0  0  0  8  5  0]
 [ 0  3  0  0  0  0]
 [ 0  0  2  2  0  0]
 [ 0  3  1 11  9  0]
 [ 0  2  2 38 15  0]
 [ 0  0  0  2  0  0]]
```



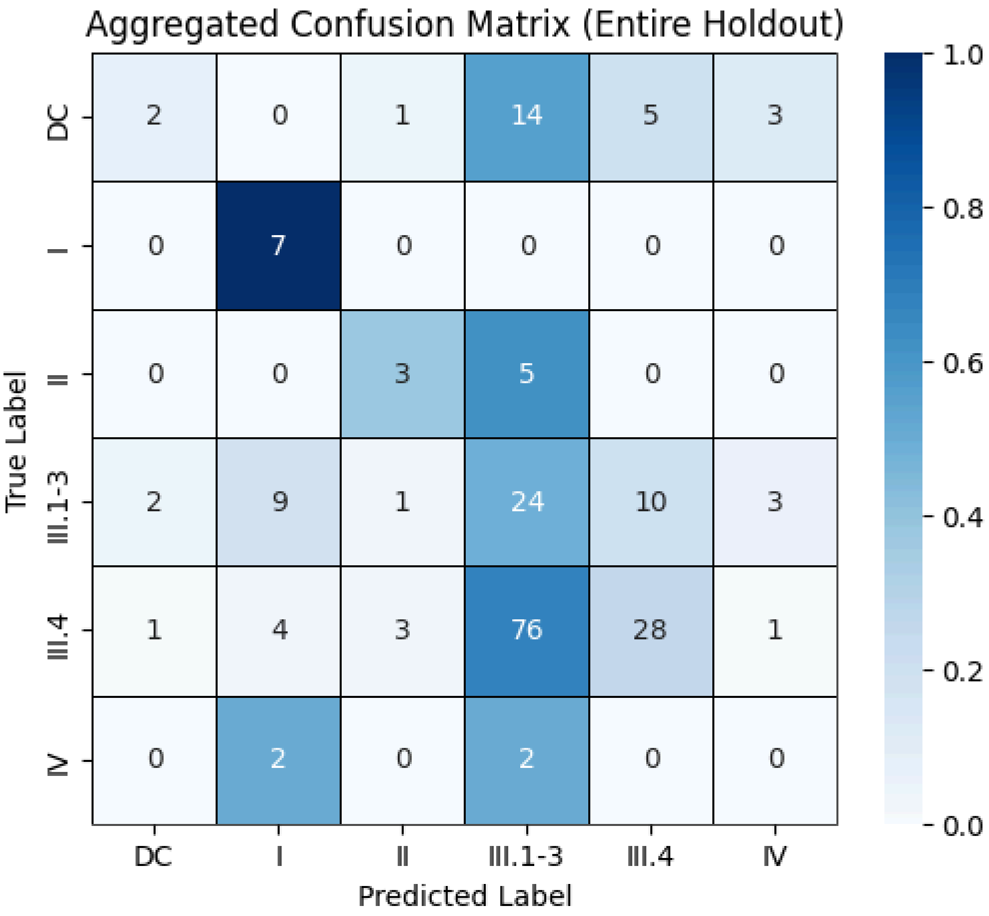
```
============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 24ms/step

== Test Set Evaluation - Run 2 ==
              precision    recall  f1-score   support

          DC     0.4000    0.1667    0.2353        12
           I     0.2857    1.0000    0.4444         4
          II     0.3333    0.2500    0.2857         4
     III.1-3     0.2167    0.5200    0.3059        25
       III.4     0.9286    0.2321    0.3714        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.3204       103
   macro avg     0.3607    0.3615    0.2738       103
weighted avg     0.6281    0.3204    0.3320       103

Balanced Accuracy: 0.36146825396825394
```

```
Confusion Matrix:
 [[ 2  0  1  6  0  3]
 [ 0  4  0  0  0  0]
 [ 0  0  1  3  0  0]
 [ 2  6  0 13  1  3]
 [ 1  2  1 38 13  1]
 [ 0  2  0  0  0  0]]
```



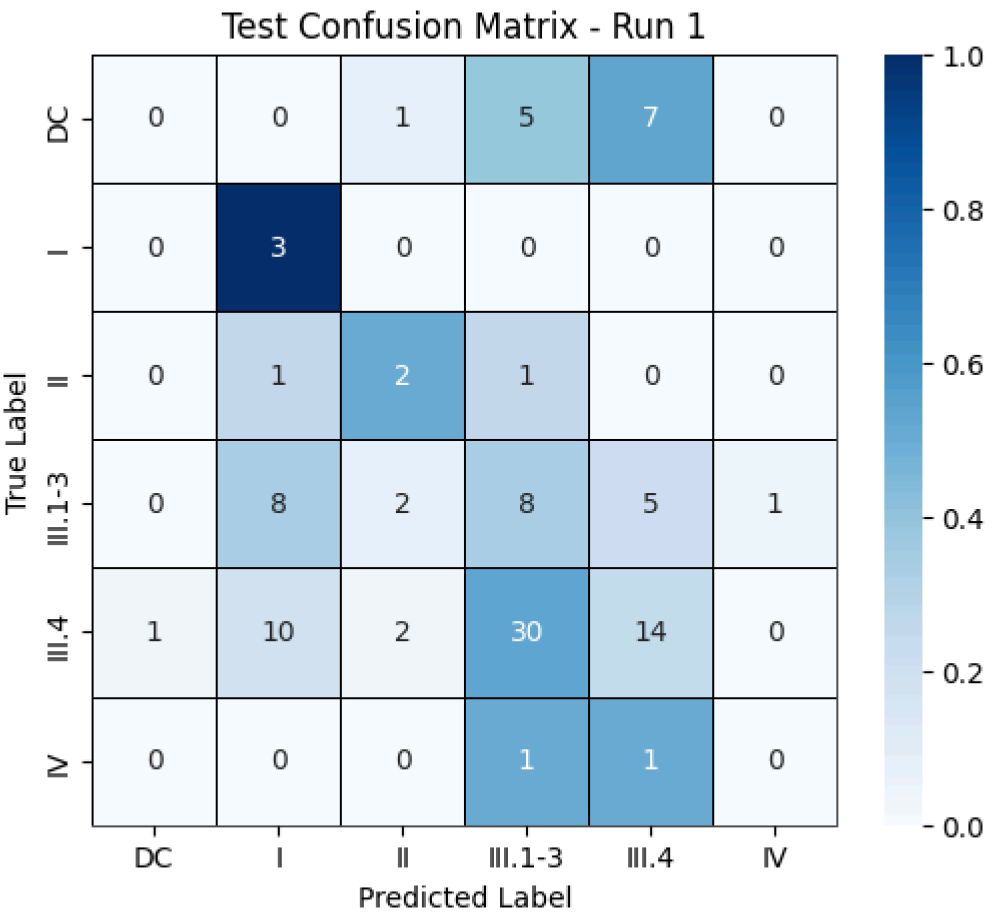
```
================================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
================================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.4000    0.0800    0.1333        25
           I     0.3182    1.0000    0.4828         7
          II     0.3750    0.3750    0.3750         8
     III.1-3     0.1983    0.4898    0.2824        49
       III.4     0.6512    0.2478    0.3590       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.3107       206
```

```
   macro avg     0.3238    0.3654    0.2721       206
weighted avg     0.4783    0.3107    0.3112       206


Aggregated Balanced Accuracy: 0.3654
Aggregated Confusion Matrix:
 [[ 2  0  1 14  5  3]
 [ 0  7  0  0  0  0]
 [ 0  0  3  5  0  0]
 [ 2  9  1 24 10  3]
 [ 1  4  3 76 28  1]
 [ 0  2  0  2  0  0]]
```



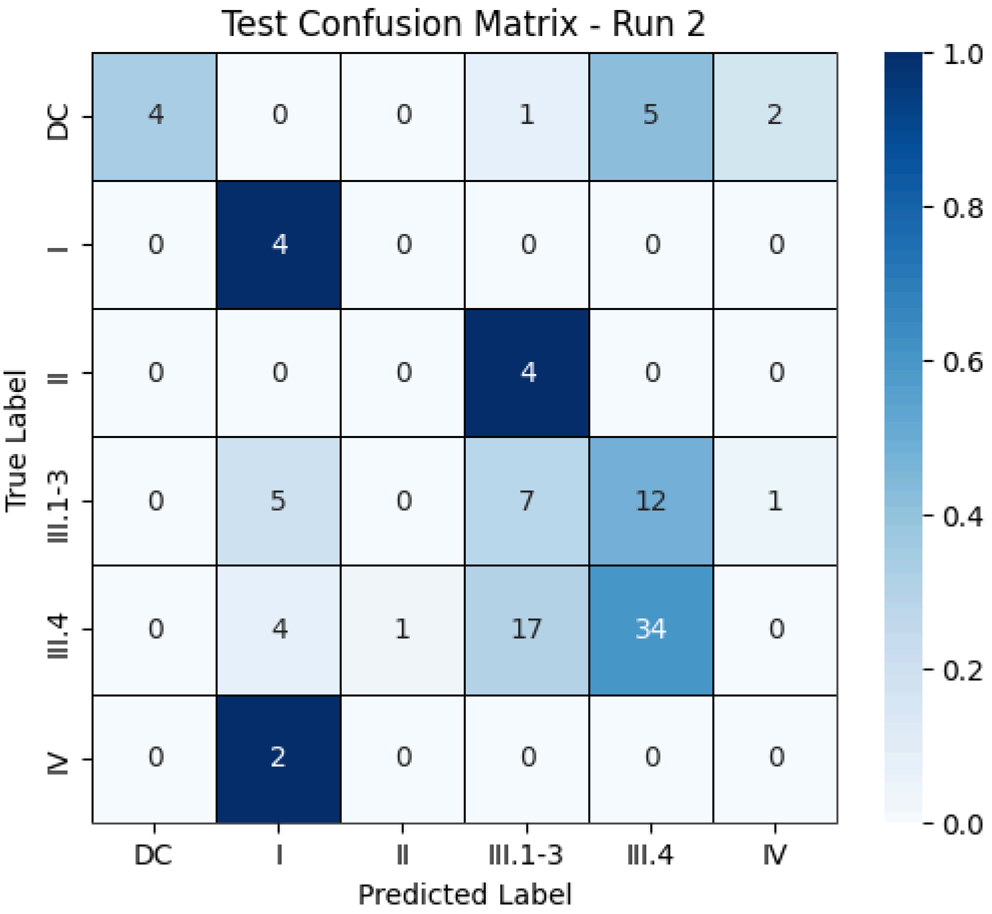
```
== Average Balanced Accuracy Across Runs ==
Run 1: 0.3702
Run 2: 0.3615
Average: 0.3659

============================================================
RUN 1: Using Split A for validation, Split B for testing
============================================================
These features will be dropped:
```

```
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 25ms/step

== Test Set Evaluation - Run 1 ==
              precision    recall  f1-score   support

          DC     0.0000    0.0000    0.0000        13
           I     0.1364    1.0000    0.2400         3
          II     0.2857    0.5000    0.3636         4
     III.1-3     0.1778    0.3333    0.2319        24
       III.4     0.5185    0.2456    0.3333        57
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.2621       103
   macro avg     0.1864    0.3465    0.1948       103
weighted avg     0.3434    0.2621    0.2596       103

Balanced Accuracy: 0.34649122807017546
Confusion Matrix:
 [[ 0  0  1  5  7  0]
 [ 0  3  0  0  0  0]
 [ 0  1  2  1  0  0]
 [ 0  8  2  8  5  1]
 [ 1 10  2 30 14  0]
 [ 0  0  0  1  1  0]]
```



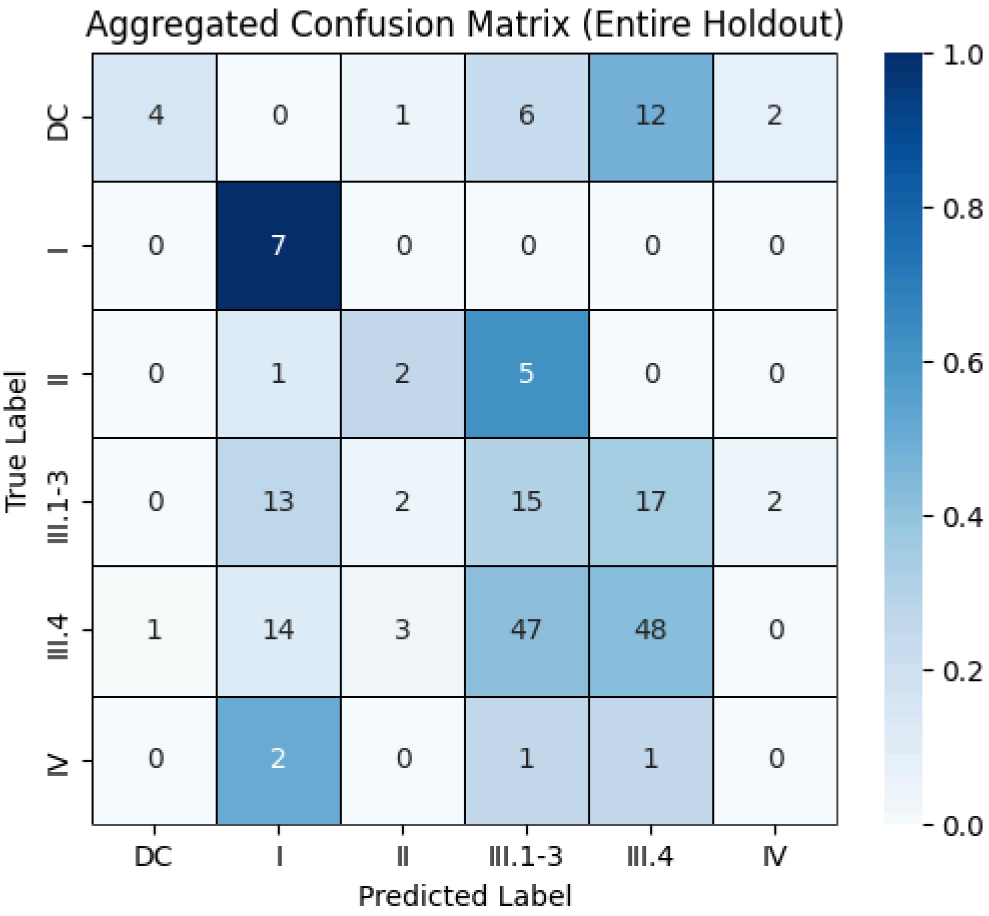
```
============================================================
RUN 2: Using Split B for validation, Split A for testing (SWAP)
============================================================
These features will be dropped:
['C113', 'C126', 'C100', 'C66', 'C11', 'C82', 'C135', 'C51', 'C86', 'C85', 'C96', 'C64', 'C87', 'C139', 'C127',
'C101', 'C92', 'C40']
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 24ms/step

== Test Set Evaluation - Run 2 ==
              precision    recall  f1-score   support

          DC     1.0000    0.3333    0.5000        12
           I     0.2667    1.0000    0.4211         4
          II     0.0000    0.0000    0.0000         4
     III.1-3     0.2414    0.2800    0.2593        25
       III.4     0.6667    0.6071    0.6355        56
          IV     0.0000    0.0000    0.0000         2

    accuracy                         0.4757       103
   macro avg     0.3625    0.3701    0.3026       103
weighted avg     0.5479    0.4757    0.4831       103

Balanced Accuracy: 0.37007936507936506
```

```
Confusion Matrix:
[[ 4  0  0  1  5  2]
 [ 0  4  0  0  0  0]
 [ 0  0  0  4  0  0]
 [ 0  5  0  7 12  1]
 [ 0  4  1 17 34  0]
 [ 0  2  0  0  0  0]]
```



```
============================================================
AGGREGATED RESULTS ACROSS ENTIRE HOLDOUT SET
============================================================

== Aggregated Classification Report ==
              precision    recall  f1-score   support

          DC     0.8000    0.1600    0.2667        25
           I     0.1892    1.0000    0.3182         7
          II     0.2500    0.2500    0.2500         8
     III.1-3     0.2027    0.3061    0.2439        49
       III.4     0.6154    0.4248    0.5026       113
          IV     0.0000    0.0000    0.0000         4

    accuracy                         0.3689       206
```

```
   macro avg     0.3429    0.3568    0.2636       206
weighted avg     0.4990    0.3689    0.3866       206


Aggregated Balanced Accuracy: 0.3568
Aggregated Confusion Matrix:
 [[ 4  0  1  6 12  2]
 [ 0  7  0  0  0  0]
 [ 0  1  2  5  0  0]
 [ 0 13  2 15 17  2]
 [ 1 14  3 47 48  0]
 [ 0  2  0  1  1  0]]
```



	DC	I	II	III.1-3	III.4	IV
DC	4	0	1	6	12	2
I	0	7	0	0	0	0
II	0	1	2	5	0	0
III.1-3	0	13	2	15	17	2
III.4	1	14	3	47	48	0
IV	0	2	0	1	1	0

```
== Average Balanced Accuracy Across Runs ==
Run 1: 0.3465
Run 2: 0.3701
Average: 0.3583

== Average Balanced Accuracy Across Runs ==
Run 1: 0.3494
Run 2: 0.2712
Average: 0.3103
```

[0.40568704472939615, 0.36259617121184756, 0.35409337186201917, 0.3654305881644693, 0.3568168683402564]

0.369 ± 0.008