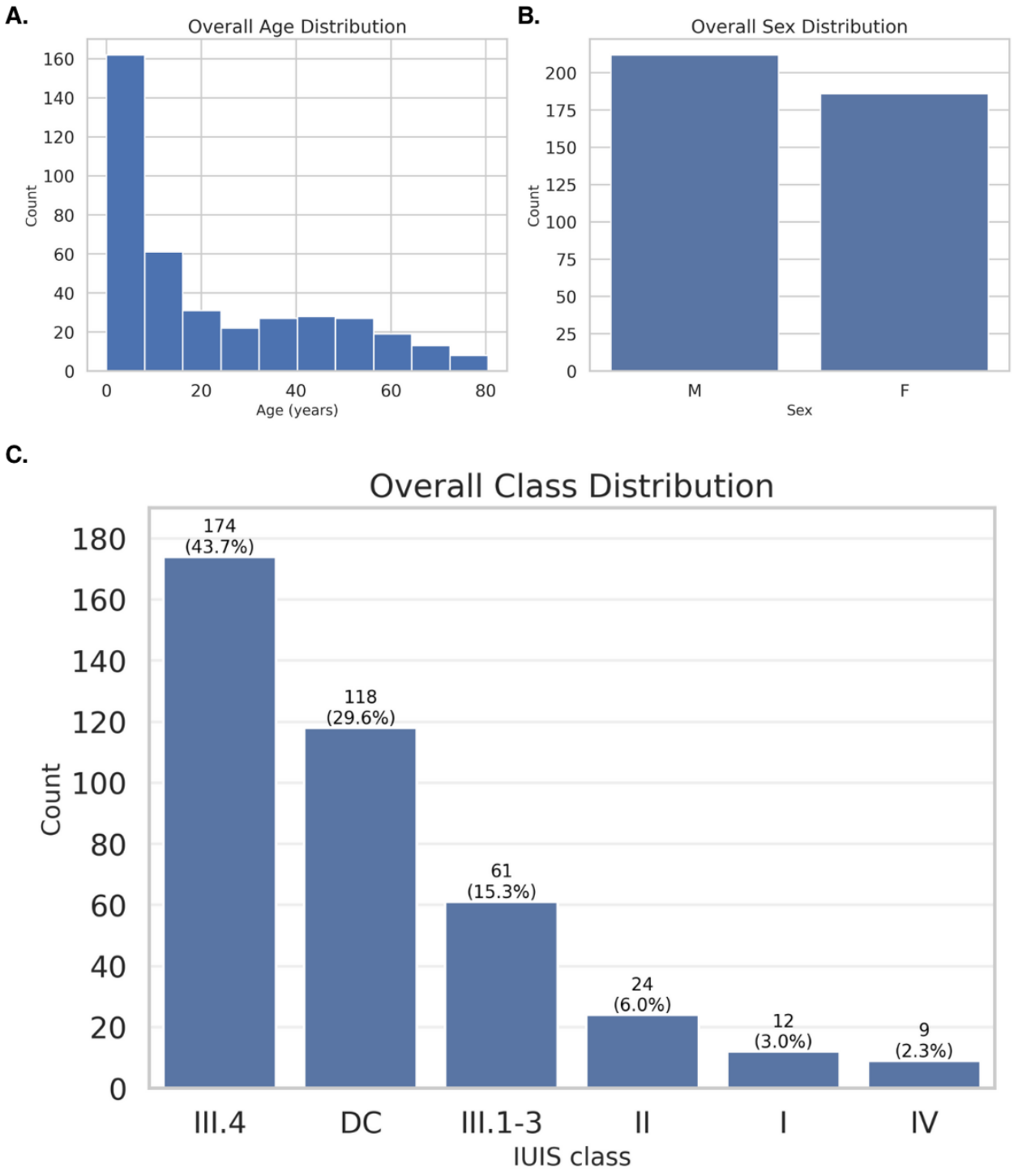


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**1: Age, sex and class distribution in the GHE_Train. A.** The age distribution of patients in the GHE_Train cohort. We can observe how an important fraction of the dataset (19.3%) consists of patients under 2 years old. **B.** The overall sex distribution of patients in the GHE_Train cohort. **C.** The overall class distribution in the GHE_Train cohort. We can observe the magnitude of the class imbalance that is affecting this dataset, with the less represented classes (IUIS class I, II and IV) accounting for less than 12% of the total dataset. Notably, a predominant proportion of the dataset comprises samples categorized under the IUIS class III.4 (43.7%) and diseased controls (29.6%).*

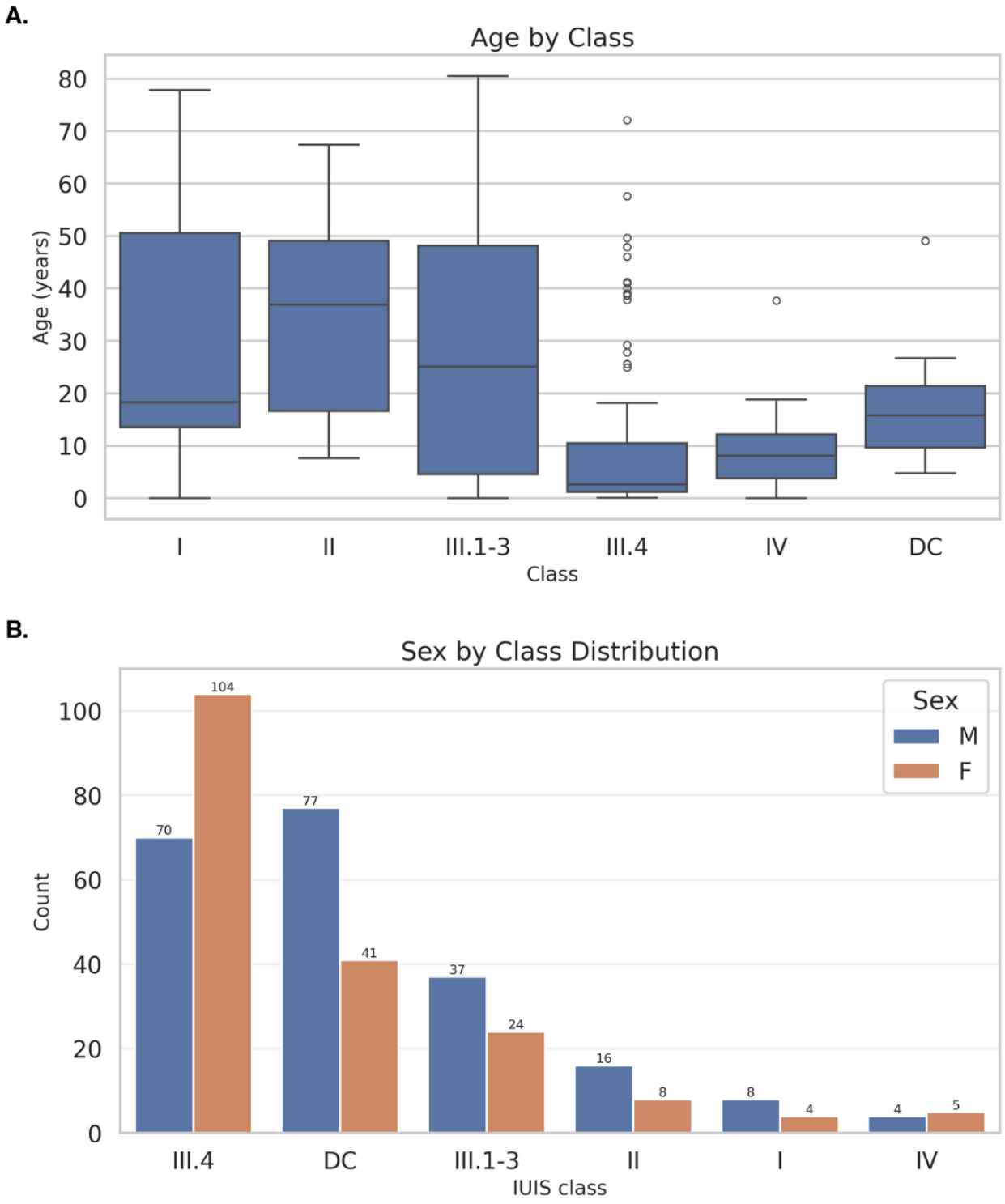


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**2: Age and Sex distribution by class in the GHE_Train cohort. A.** Age distribution by class in the GHE_Train cohort. **B.** Sex distribution by class in the GHE_Train cohort. Despite the overall sex distribution being quite equilibrated, most of the classes suffer from a noticeable unbalance between male and female patients.*

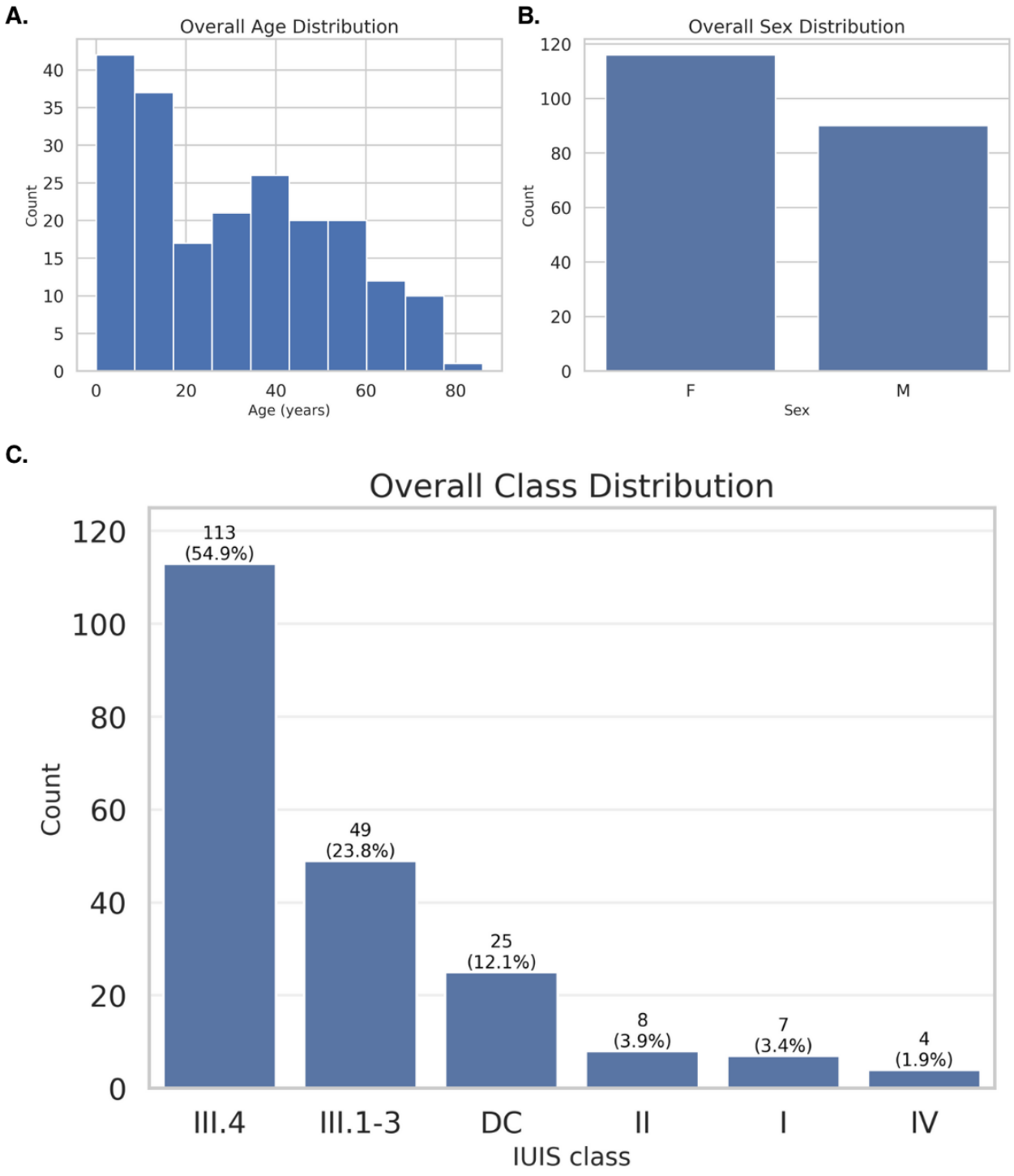
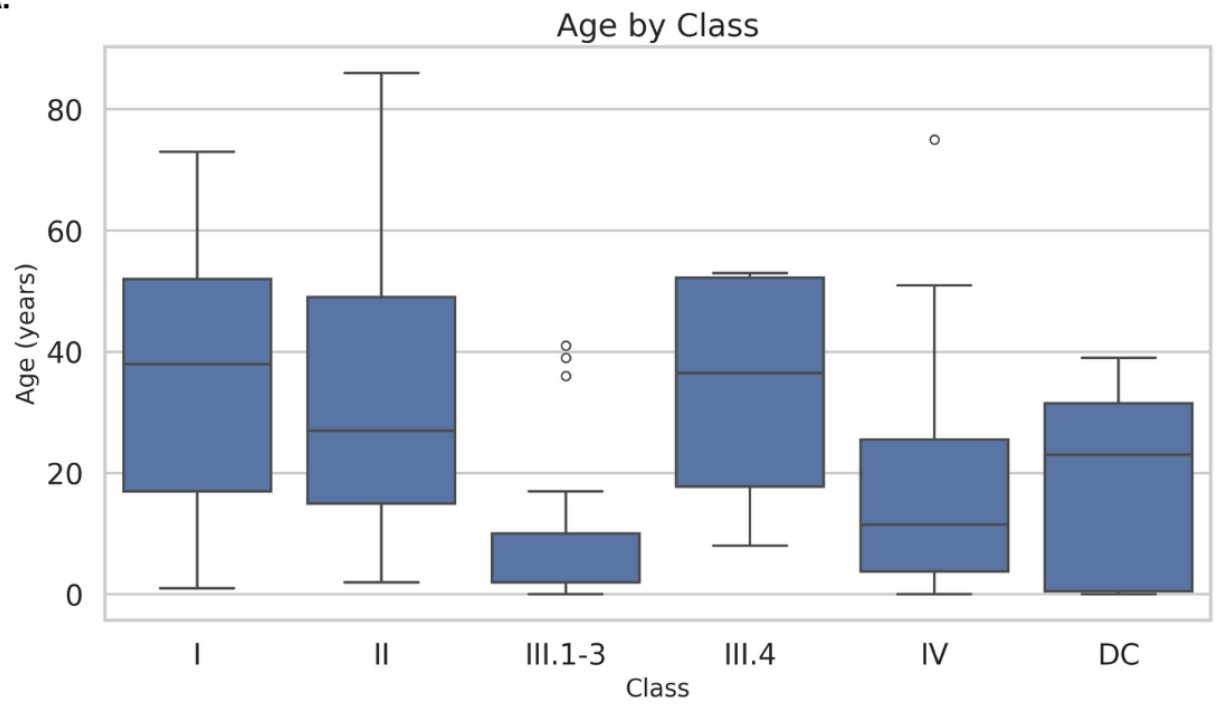


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**3: Age, sex and class distribution in the GHE_Val cohort. A.** The age distribution of patients in the GHE_Val cohort. We can observe how the fraction of patients under 2 years old (6.8%) is lower compared to the one of the GHE_Train cohort (19.3%) **B.** The overall sex distribution of patients in the GHE_Val cohort. **C.** The overall class distribution in the GHE_Val cohort. Heavy class imbalance is also affecting this dataset, but class abundances differ significantly compared to the GHE_Train cohort.*

A.


Age by Class
Age (years)
0
20
40
60
80
I
II
III.1-3
III.4
IV
DC
Class

B.

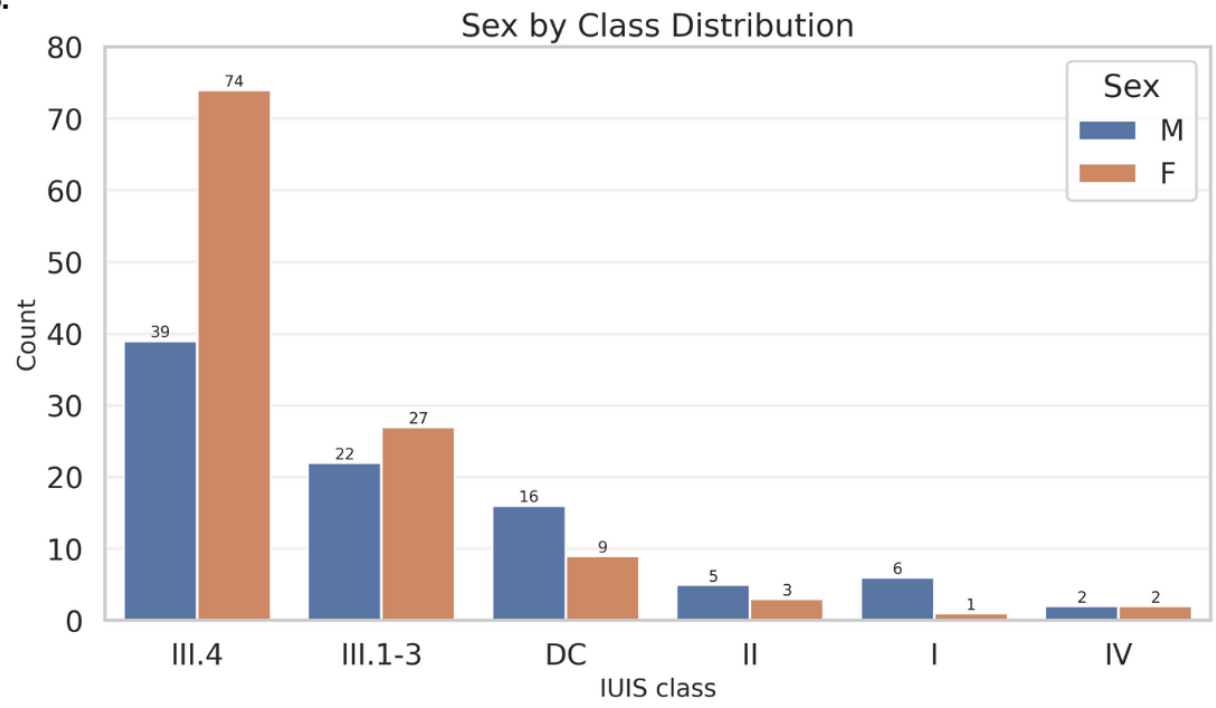

Sex by Class Distribution
Count
0
10
20
30
40
50
60
70
80
Sex
M
F
39
74
22
27
16
9
5
3
6
1
2
2
III.4
III.1-3
DC
II
I
IV
IUIS class

Figura** Errore. Nel documento non esiste testo dello stile specificato. **4: Age and Sex distribution by class

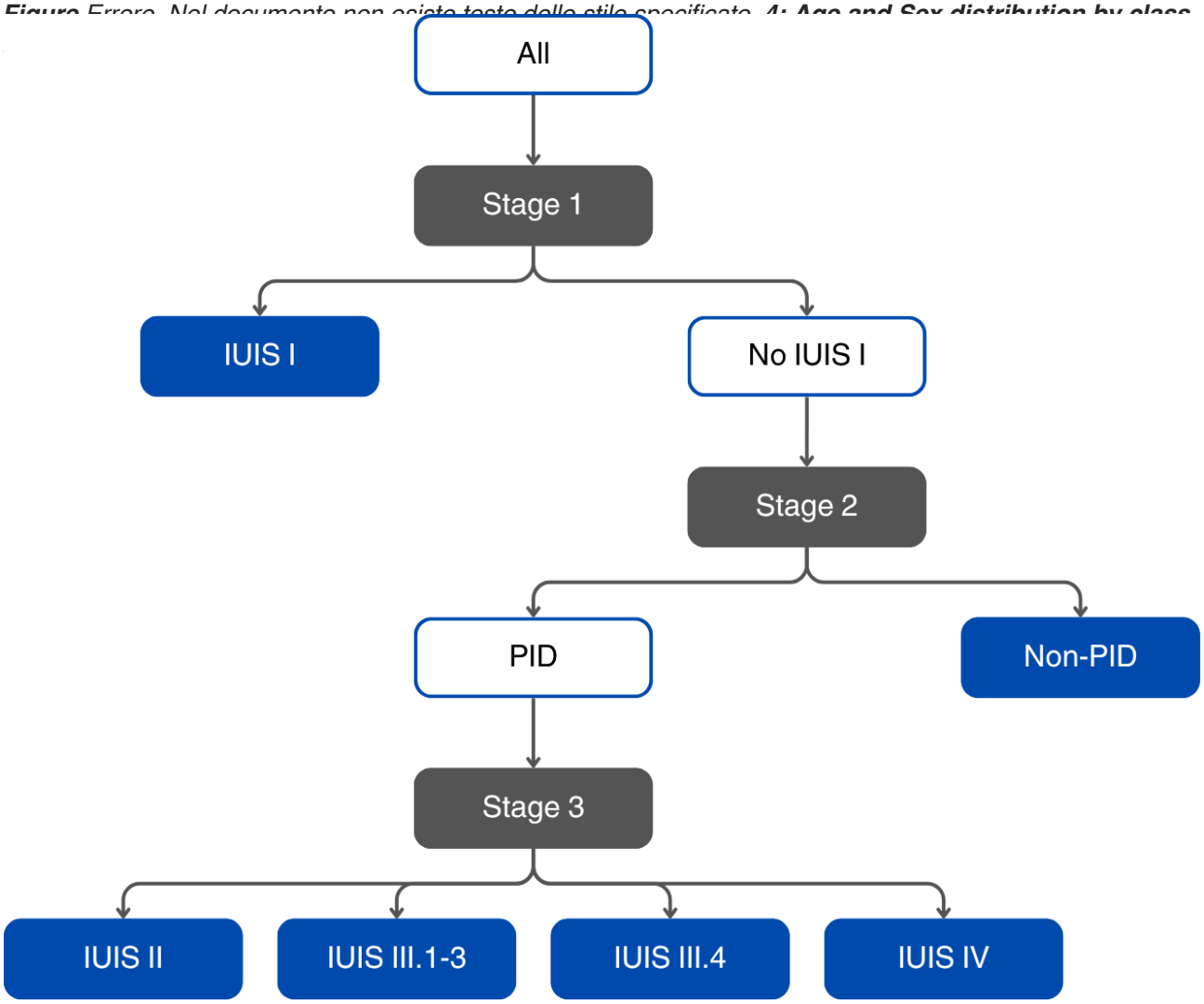


***Figure** E**classifier**classificat**immunode**I. T**before tac**interpreta*

***rarchical** *PIDgeon* *combined* *UIS class* *ategories* *ness and*

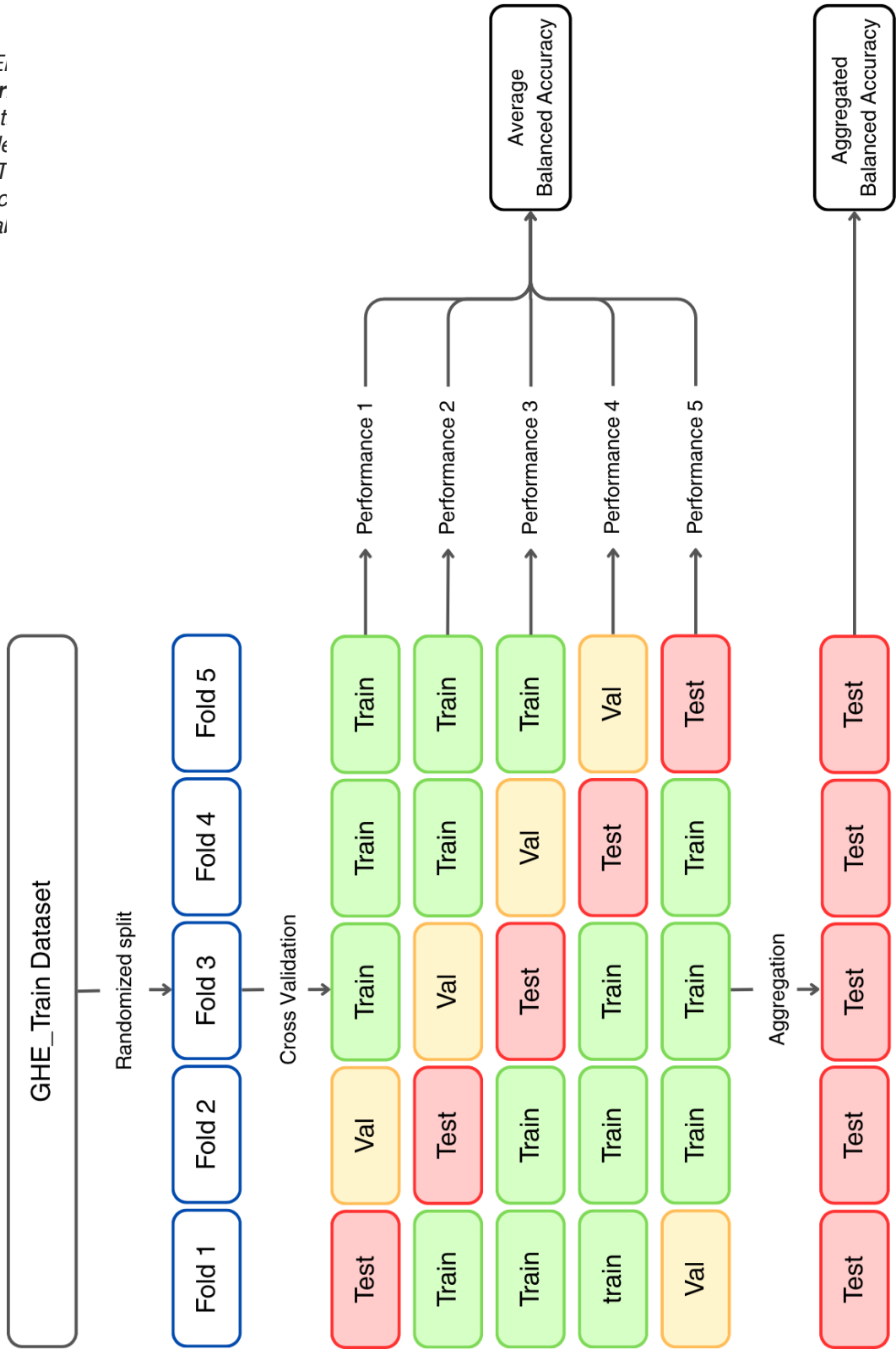


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**6: Diagram of the 5-fold cross validation routine.** The training dataset is split into 5 subset with equal class proportions. During each iteration of the CV routine, 3 subsets form the training set, one is used for monitoring, callbacks and threshold optimization (subset in yellow "Val"), while the last one is used to assess the performance of the model. Two type of performance metrics can be computed. The metrics obtained at each iteration can be averaged. Alternatively, the predictions from each iteration can be aggregated in a single confusion matrix, and performance metrics can be computed from this aggregated result.*

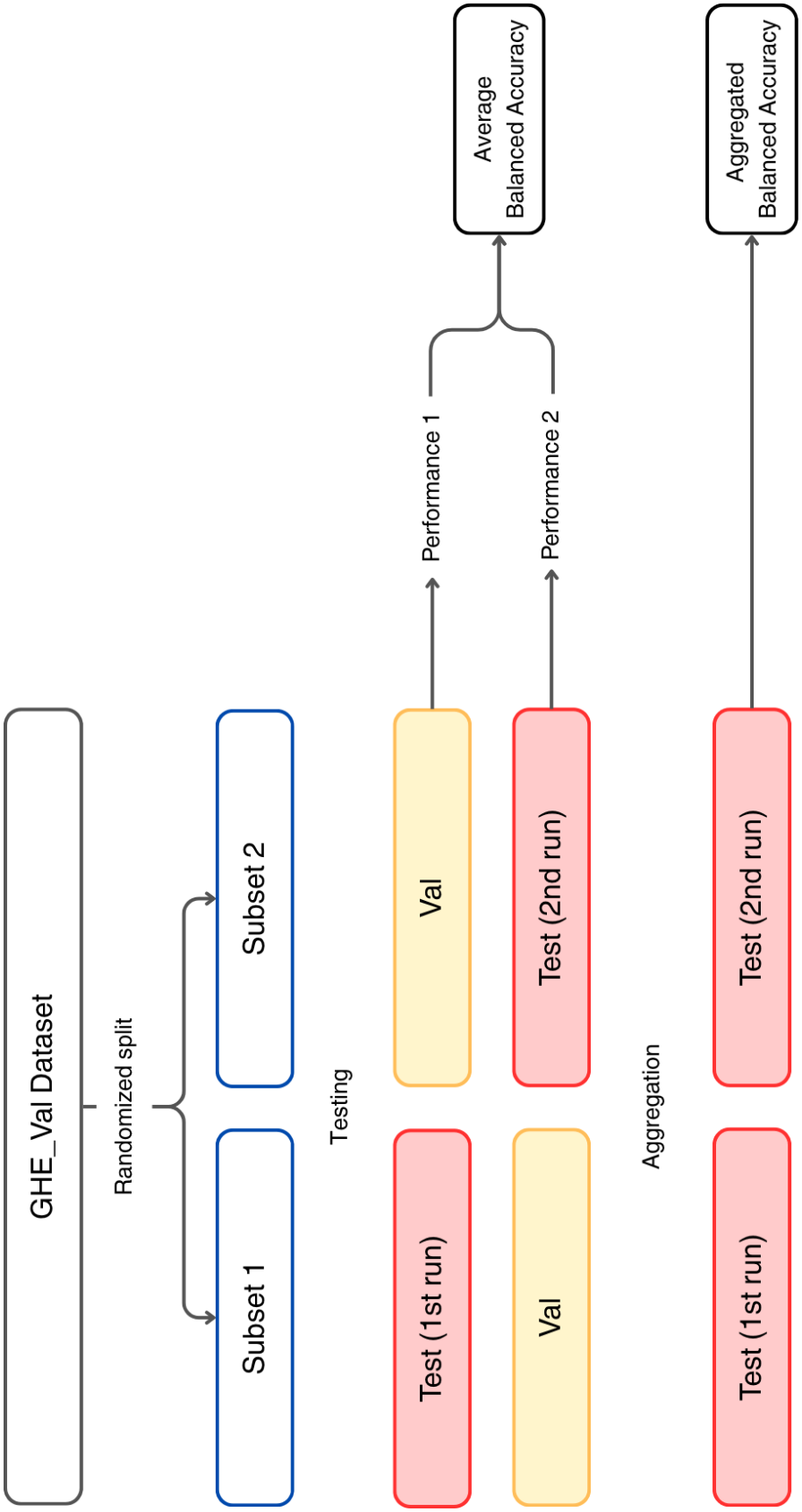


Figure** Errore. Nel documento non esiste testo dello stile specificato..**7: Diagram of the final testing routine.

In the final part of this thesis, the models were tested on a holdout dataset obtained from the GHE_Val cohort. The holdout dataset was partitioned into two equal subsets, designated as the validation and test sets, while preserving the relative distribution of classes. The two subsets were then swapped in a second, independent run, and the corresponding metrics and confusion matrices were aggregated to evaluate the models' performance across the entire holdout dataset.

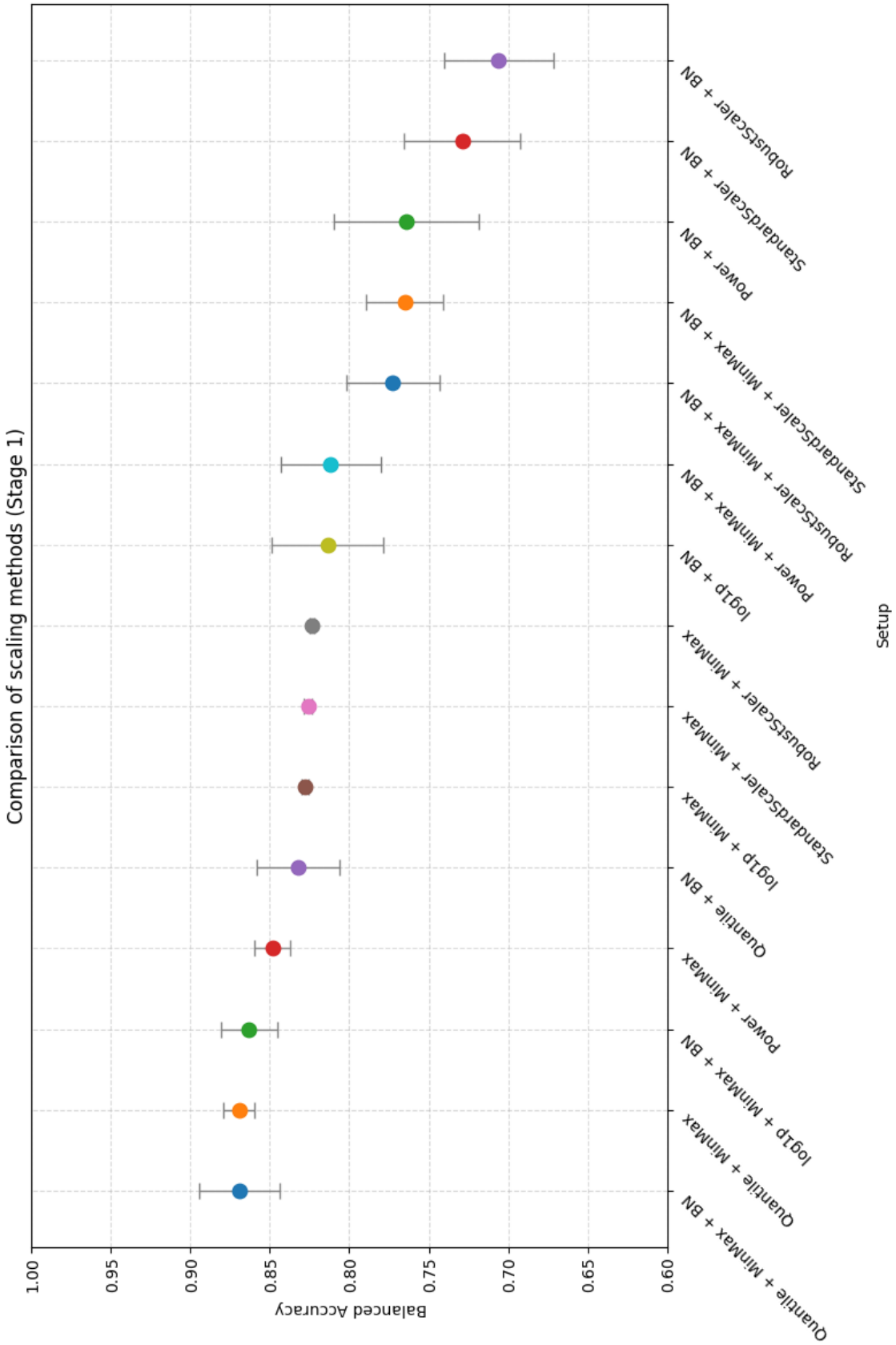


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**8: Comparison of scaling methods for Stage 1.** The best-performing preprocessing strategy was the QuantileTransformer combined with MinMax scaling and batch normalization, which achieved a mean balanced accuracy of 0.869 ± 0.025.*

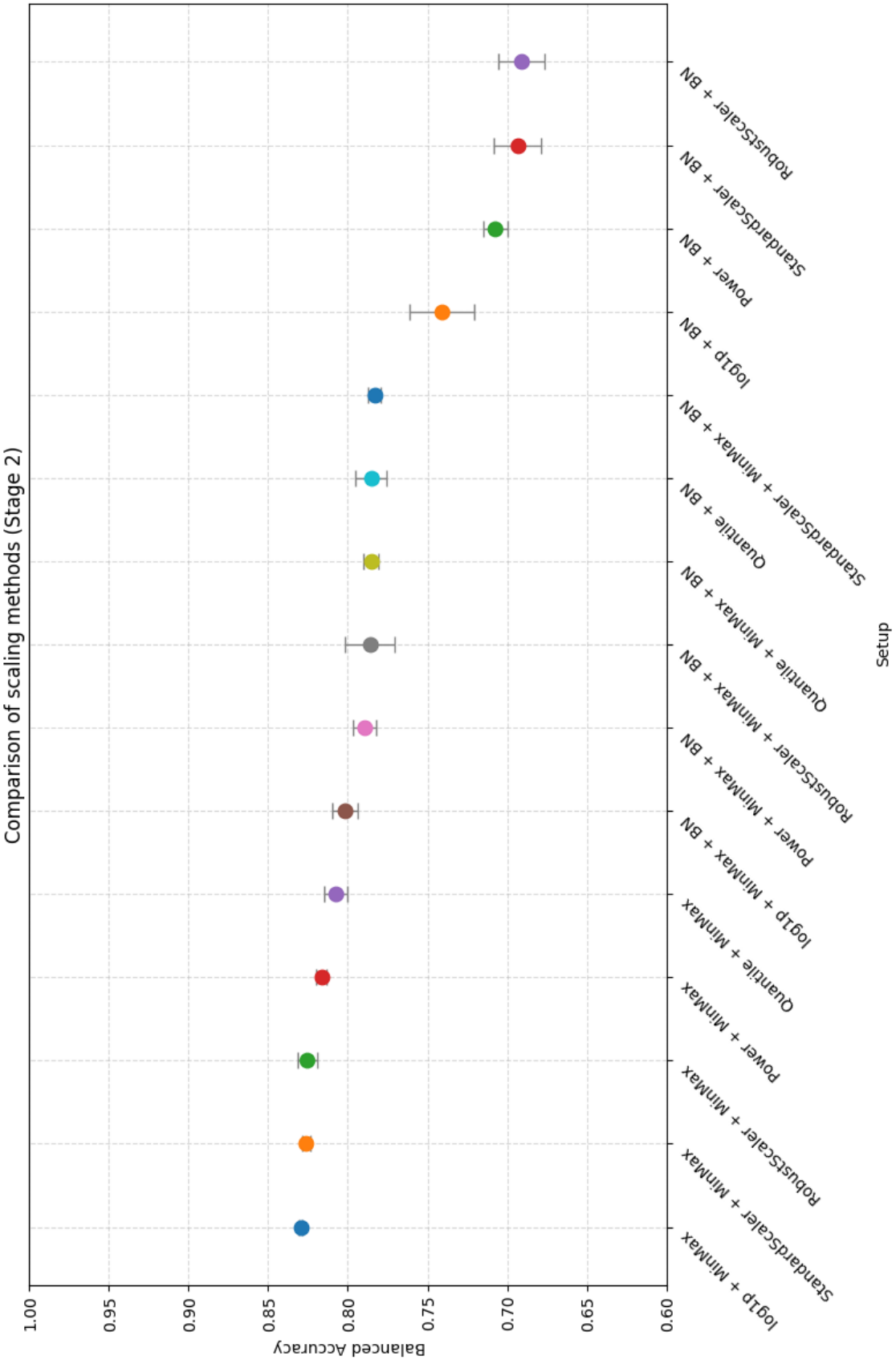


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**9: Comparison of scaling methods for Stage 2.** The best performing preprocessing strategy was the logarithmic transformation followed by MinMax scaling, which achieved a mean balanced accuracy of 0.830 ± 0.001.*

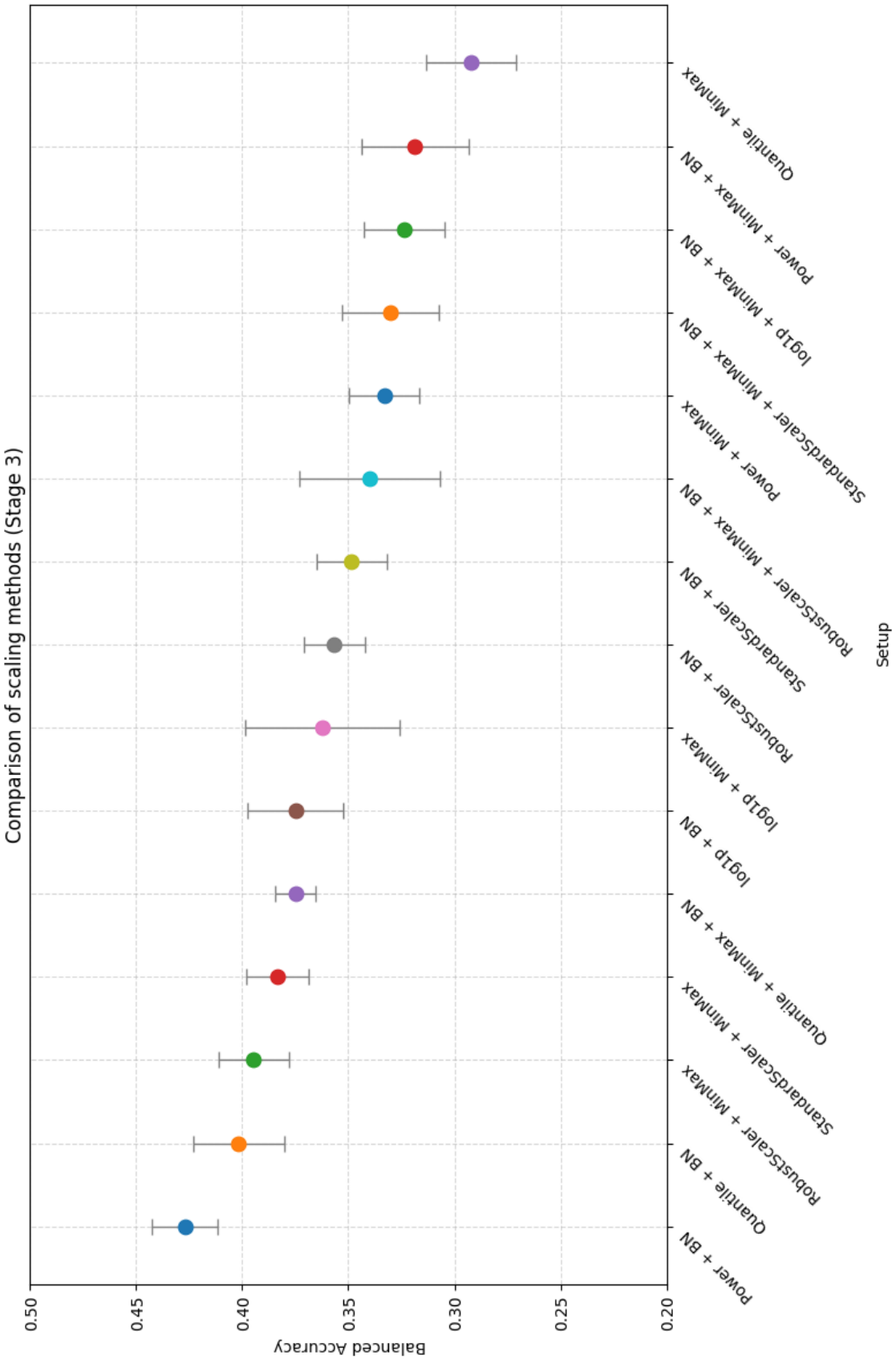


***Figure** Errore. Nel documento non esiste testo dello stile specificato.**.10: Comparison of scaling methods for Stage 3.** The best performing preprocessing strategy was the PowerTransformer combined with batch normalization, which achieved a balanced accuracy of 0.427 ± 0.016.*

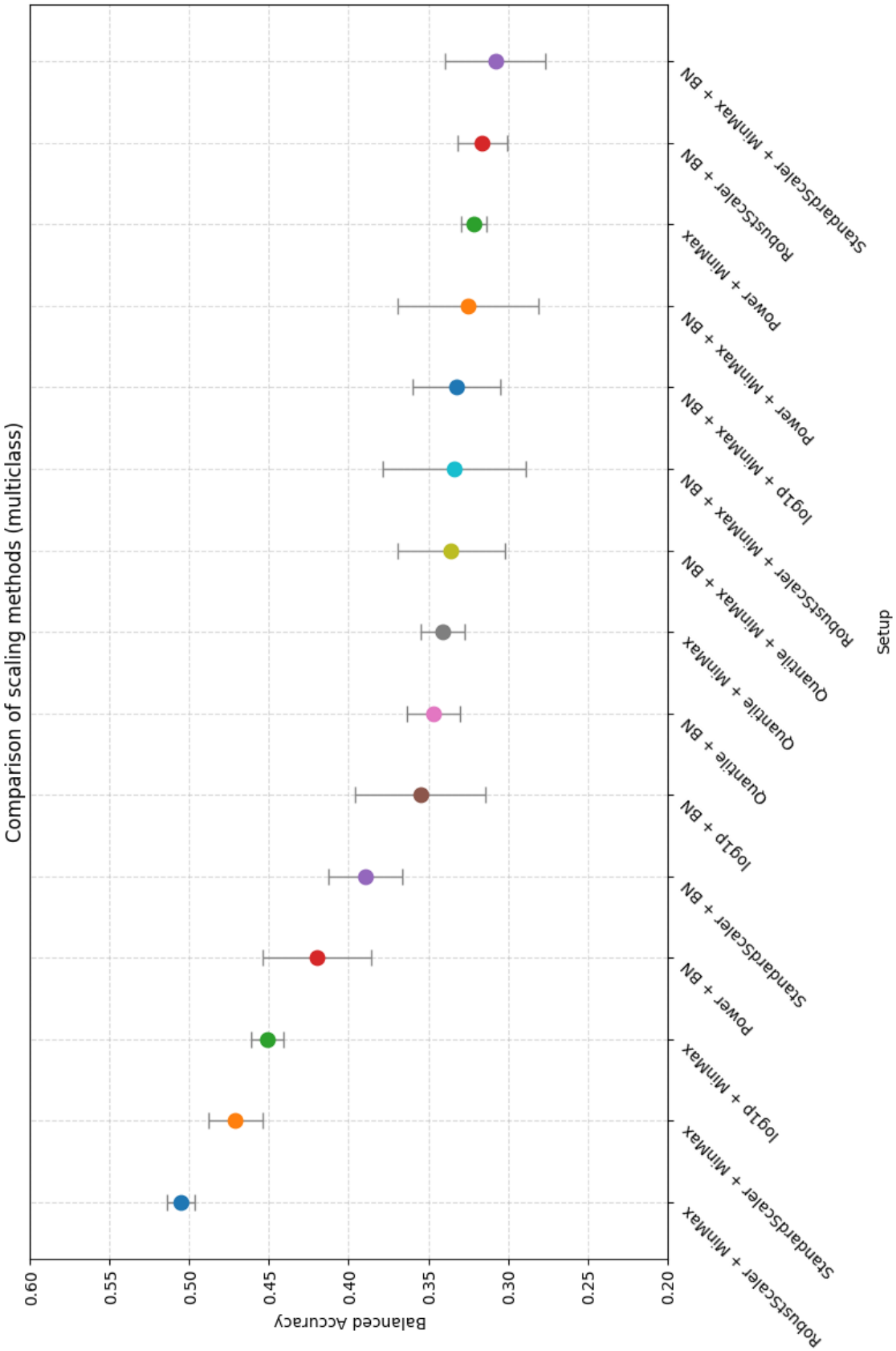


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**11: Comparison of scaling methods for the flat multiclass classifier.** The best performing preprocessing strategy was the RobustScaler combined with MinMax scaling, which achieved a balanced accuracy of 0.505 ± 0.009.*

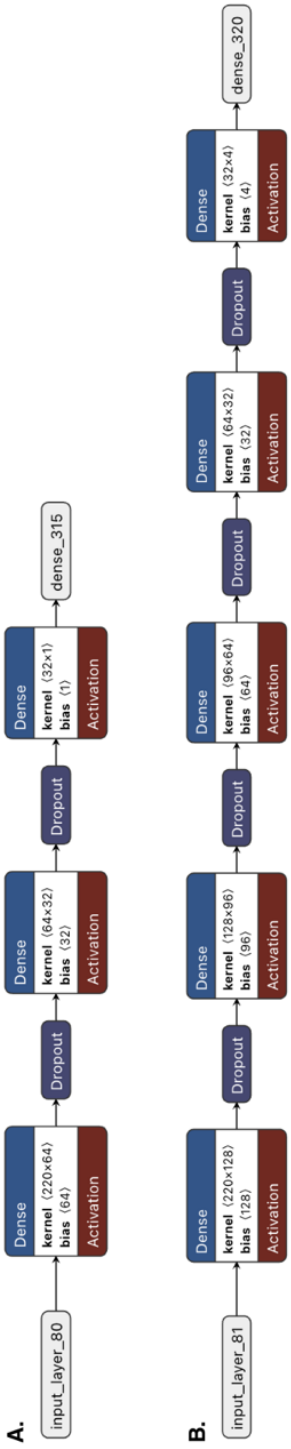


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**12: Baseline architectures for the binary and multiclass classifiers. A.** Baseline architecture for the binary classifier. This architecture was used for both Stage 1 and Stage 2 during the optimization of the preprocessing pipeline and as a starting point for the Bayesian Optimization. **B.** Baseline architecture for the multiclass classifier. This architecture was used for Stage 3 and the flat classifier during the optimization of the preprocessing pipeline and as a starting point for the Bayesian Optimization.*

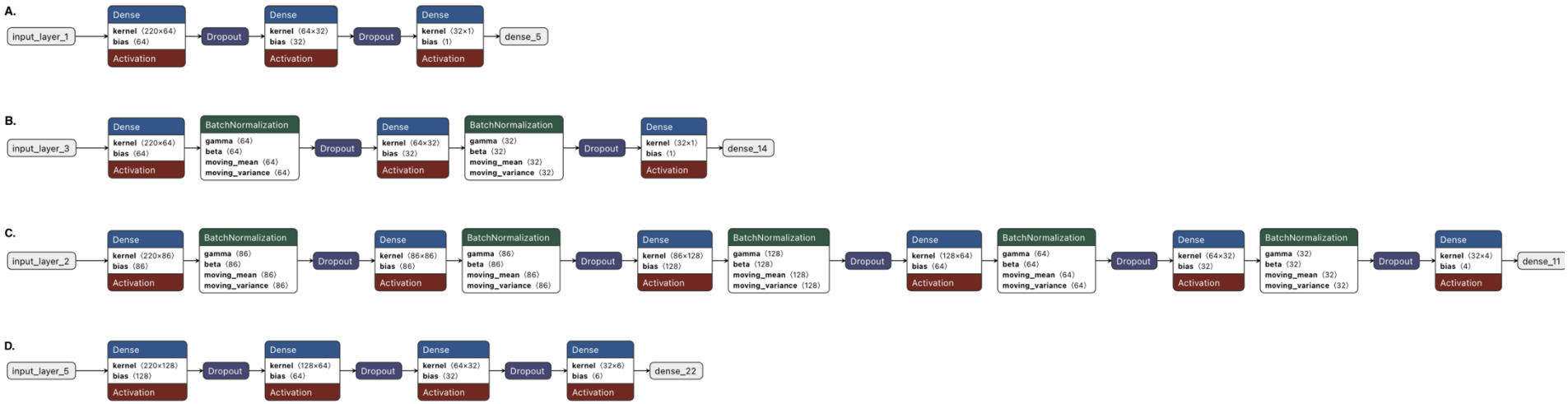


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**13: Optimized architecture for all models. A.** Optimized Architecture for Stage 1. No additional hidden layers were added compared to baseline. **B.** Optimized Architecture for Stage 2. Likewise, no supplementary hidden layer was added, but a batch normalization layer was implemented after each hidden layer following the results of the previous optimization **C.** Optimized architecture for Stage 3. Two additional hidden layers with 86 neurons each were added to the baseline model. **D.** Optimized architecture for the flat multiclass classifier. Optimization did not require an increase in depth, with no hidden layers added beyond the baseline.*

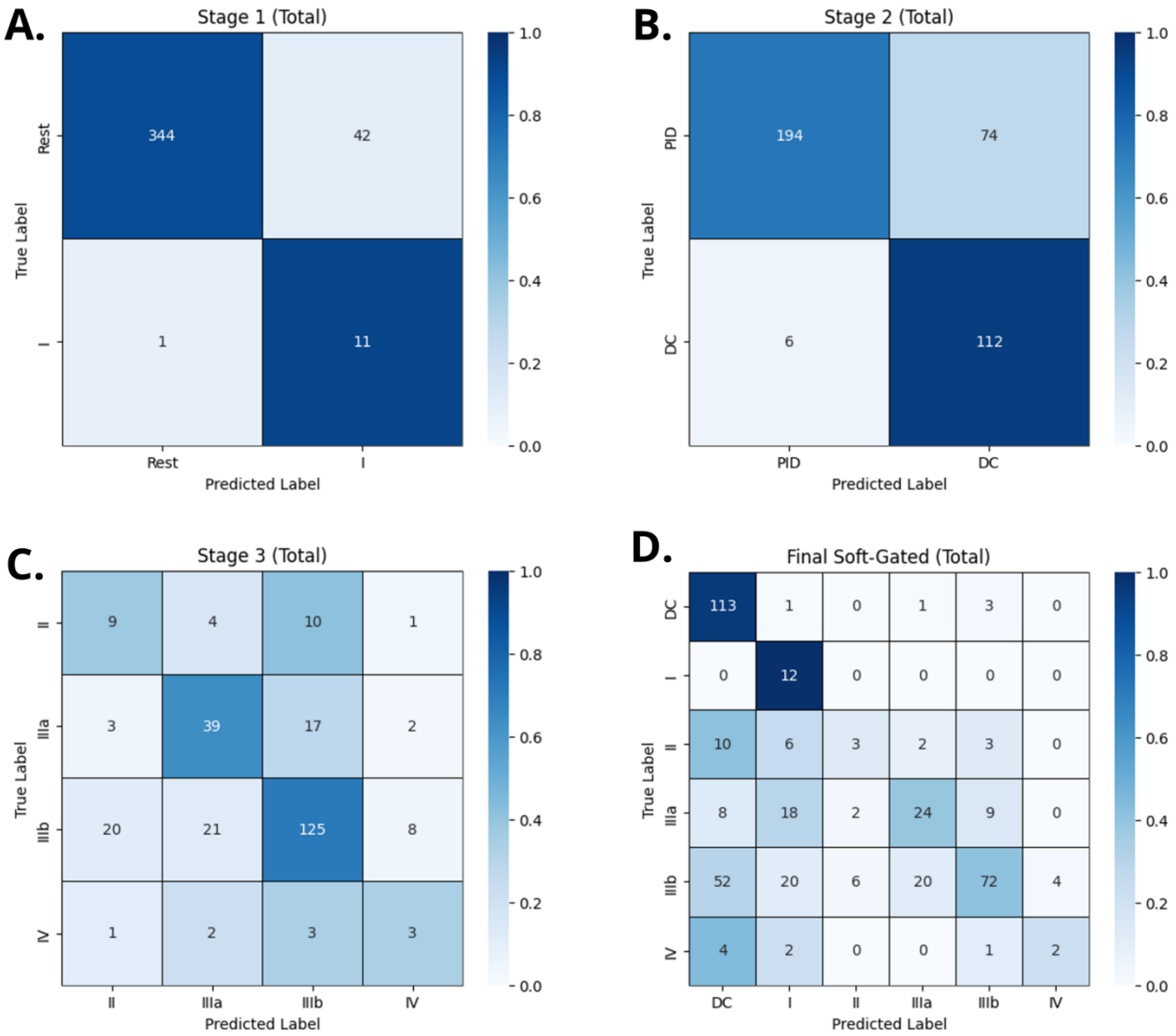


Figure *Errore. Nel documento non esiste testo dello stile specificato.**.14: Confusion matrices which achieved the mean balanced accuracy for each stage during cross-validation.** In the matrices, IIIa corresponds to IUIS class III.1-3 and IIIb to IUIS class III.4.* ***A.*** *Confusion matrix for Stage 1 (bACC = 0.904). The classifier correctly identified 11 of 12 class I cases, with a low precision (0.21) but very high recall (0.92).* ***B.*** *Confusion matrix for Stage 2 (bACC = 0.837). The model displays complementary error profiles across the two classes: high precision and low sensitivity for PID, while very high sensitivity but only moderate precision for DC, indicating a tendency to over-predict DC and generate more false positives.* ***C.*** *Confusion matrix for stage 3. The results show uneven performance across classes, with strong metrics for well-represented and distinct groups, and noticeably poorer performance for minority or heterogeneous classes.* ***D.*** *Confusion matrix for the aggregated predictions (bACC = 0.519). The soft-gated approach manages to be extremely sensitive for the IUIS class I samples, even though the first stage of this same run resulted in a false negative. However, the model appears to be biased towards predicting DC.*

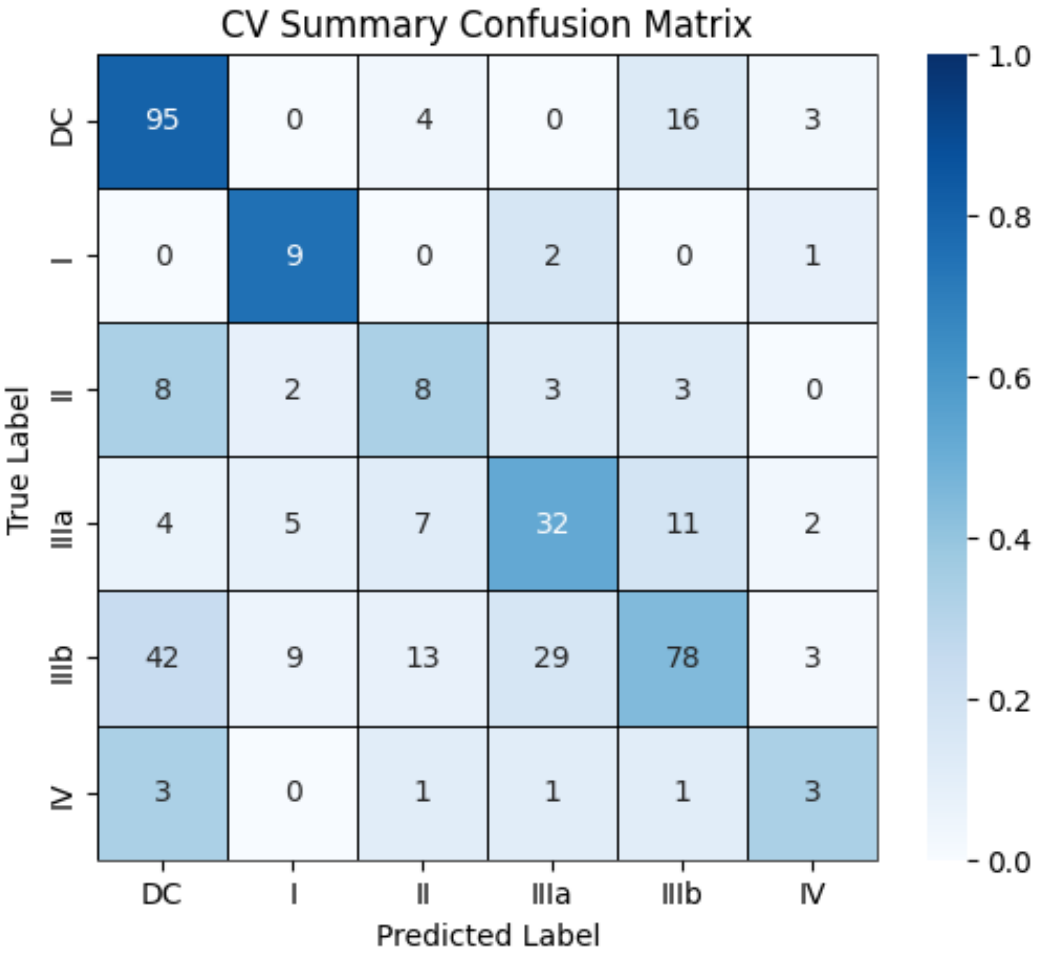


***Figure** Errore. Nel documento non esiste testo dello stile specificato.**.15: Confusion matrix which achieved the mean balanced accuracy for the flat multiclass classifier during cross-validation.** In the matrix, IIIa corresponds to IUIS class III.1-3 and IIIb to IUIS class III.4. The model's performance varies markedly across classes (bACC = 0.532): well-represented groups achieve high recall and reasonably strong precision, while minority or heterogeneous categories show weaker, less reliable detection. Underrepresented IUIS classes (II and IV) drive most misclassifications, whereas classes I, III.1-3, and III.4 display only moderate or imbalanced precision-recall profiles.*

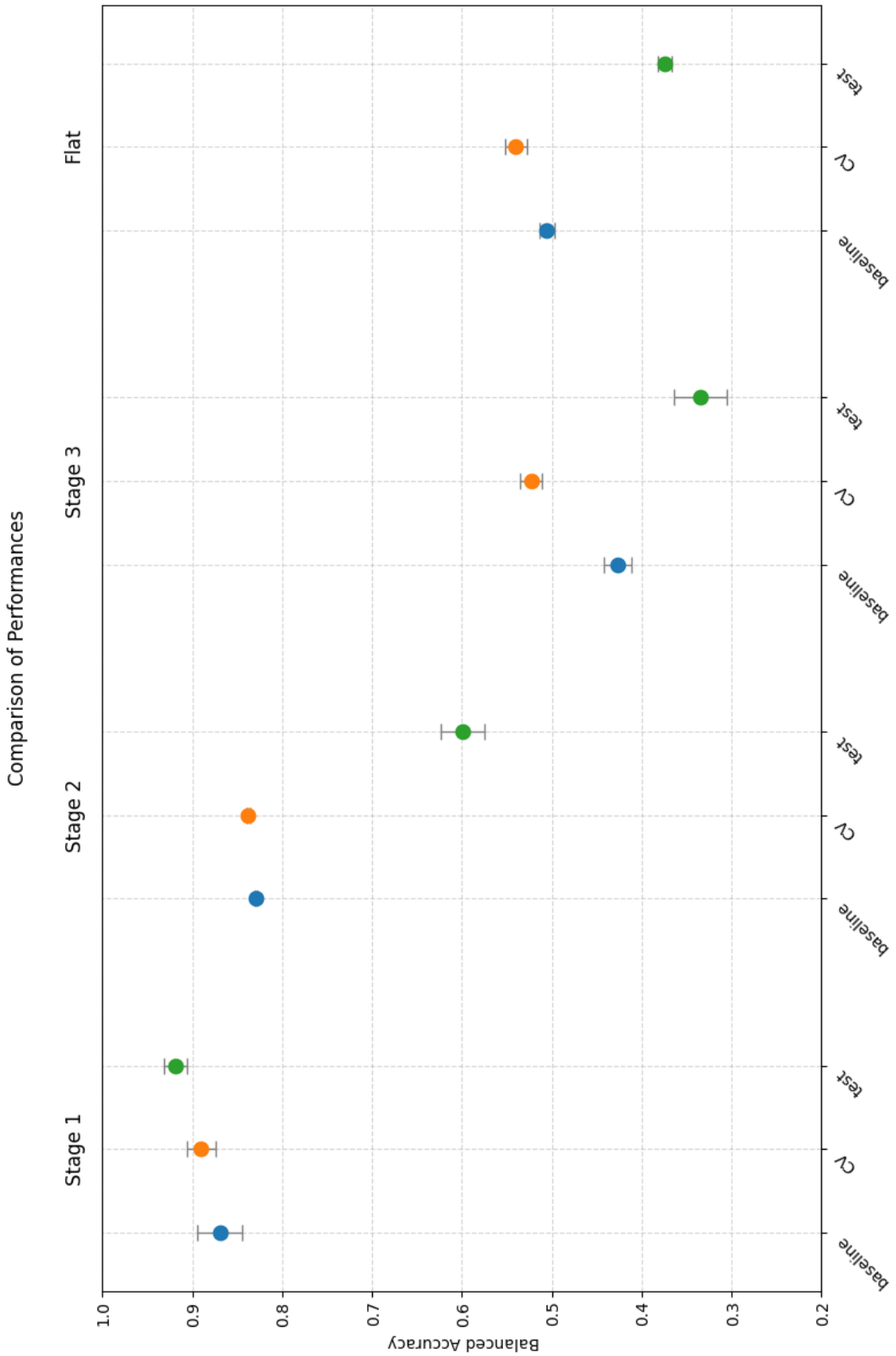


***Figure** Errore. Nel documento non esiste testo dello stile specificato..**16: Comparison of performances.** The figure reports the balanced accuracy recorded during preprocessing optimization ("baseline"), cross-validation ("CV") and testing ("test") across the four evaluation stages: Stage 1, Stage 2, Stage 3, and the Flat configuration. Each point represents the mean balanced accuracy over five runs for a given setup (baseline, cross-validation, and test), with error bars indicating the corresponding standard error. Results are grouped by stage. Aside from Stage 1, performances during testing are systematically lower compared to cross-*

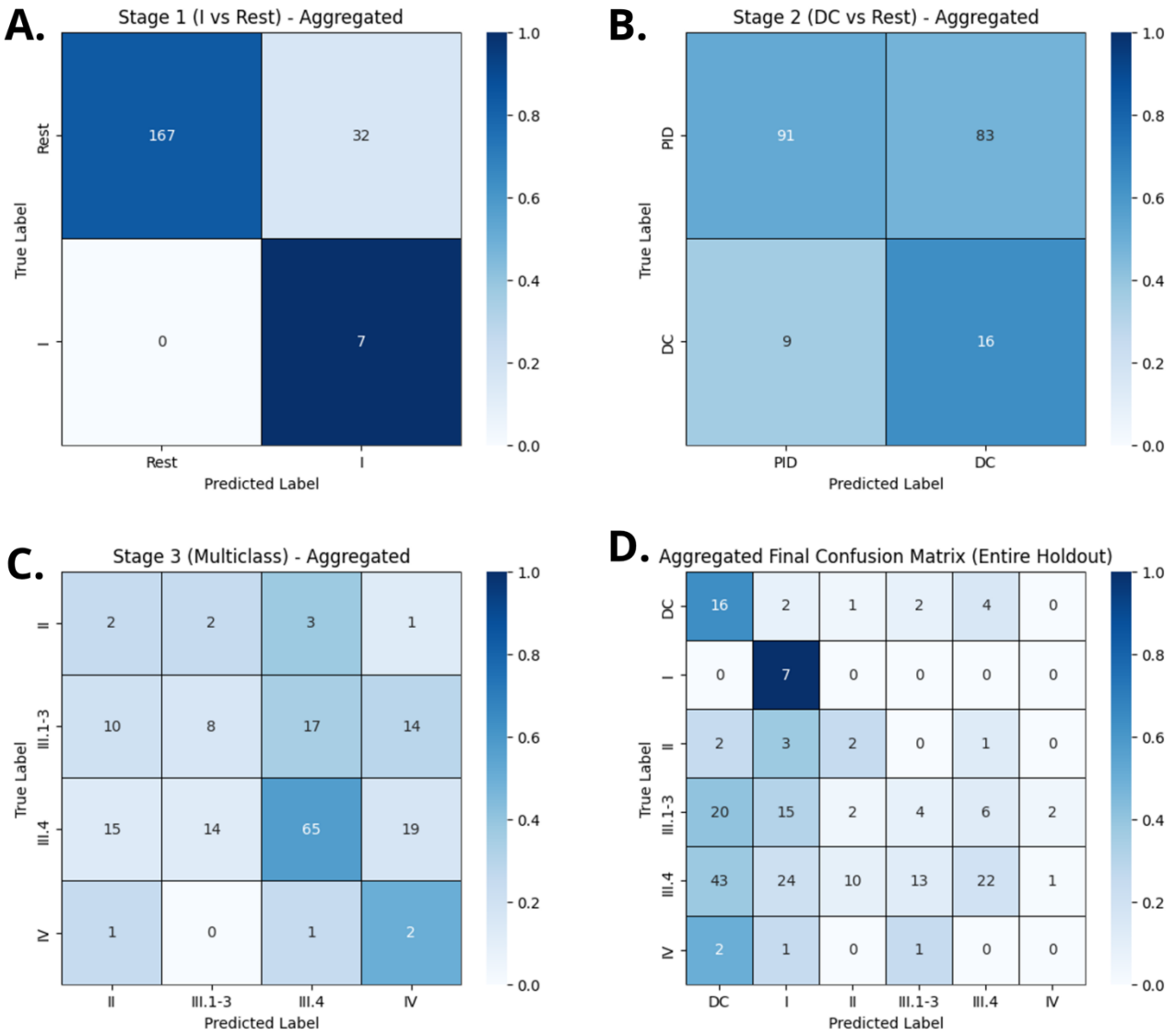


Figure *Errore. Nel documento non esiste testo dello stile specificato.**.17:* ***Confusion matrices which achieved the mean balanced accuracy for each stage during testing. A.*** *Confusion matrix for Stage 1 (bACC = 0.920). Here, the model scored perfect recall (1) but low precision (0.180) for IUIS class I.* ***B.*** *Confusion matrix for Stage 2 (bACC = 0.580). In this stage, the model struggles to establish robust decision boundaries between PID and DC cases. PID cases are identified with very high precision (0.91), but much lower recall (0.52). In contrast, DC samples show the opposite behaviour: precision is very low (0.16), but recall is comparatively high (0.64).* ***C.*** *Confusion matrix for Stage 3 (bACC = 0.372). Performance is strongly concentrated in class III.4, which achieves high precision (0.756) and moderate recall (0.575). Classes II and III.1-3 show weak reliability, with low precision (0.071 and 0.333) and limited recall (0.250 and 0.163). Class IV is particularly unstable, combining very low precision (0.056) with deceptively high recall (0.500), driven by its extremely small sample size.* ***D.*** *DC shows precision 0.193 and recall 0.640. Class I reaches precision 0.135 and recall 1.000. Class II has precision 0.133 and recall 0.250. Class III.1–3 records precision 0.200 and recall 0.082. Class III.4 attains precision 0.667 and recall 0.195. The model was unable to correctly identify any class IV sample.*

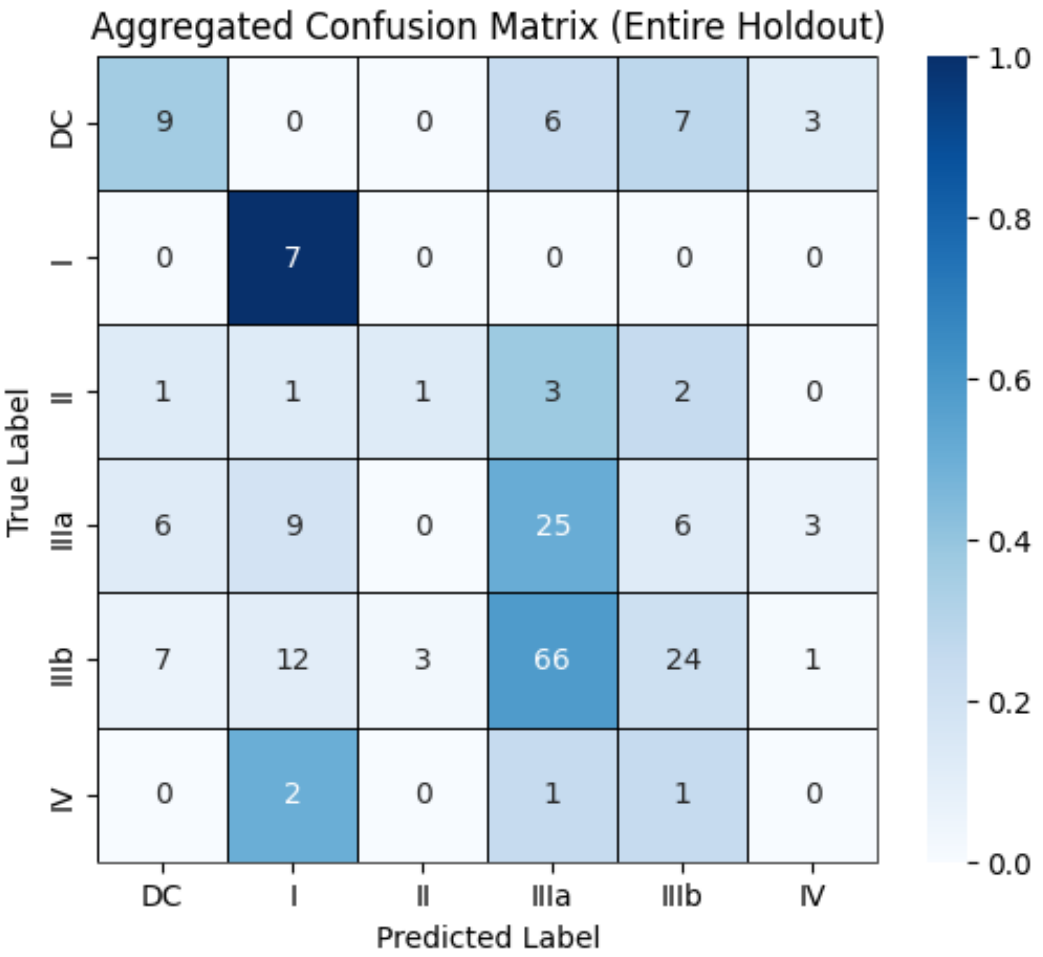


***Figure** Errore. Nel documento non esiste testo dello stile specificato.**.18: Confusion matrix which achieved the mean balanced accuracy for the flat multiclass classifier during testing.** In the matrix, IIIa corresponds to IUIS class III.1-3 and III.4 to IUIS class III.4. DC shows precision 0.391 and recall 0.360. Class I reaches precision 0.226 and recall 1.000. Class II has precision 0.250 and recall 0.125. Class IIIa records precision 0.248 and recall 0.510. Class III.4 attains precision 0.600 and recall 0.212. The model was unable to correctly identify any class IV sample in this run.*