Artificial Intelligence in Life Sciences - Challenge 1: QSAR

Benjamin Pommer k1525693

Description

The aim is to predict a molecule's biological activity or toxicity based on its chemical structure. You are given a dataset of molecules together with activities, and you should train a machine-learning model to predict the activities of new molecules based on the selected model.

Notes:
- In the training data: +1=active, 0=unknown, -1=inactive
- If you do any molecule standardization, make sure not to delete any test set molecules.

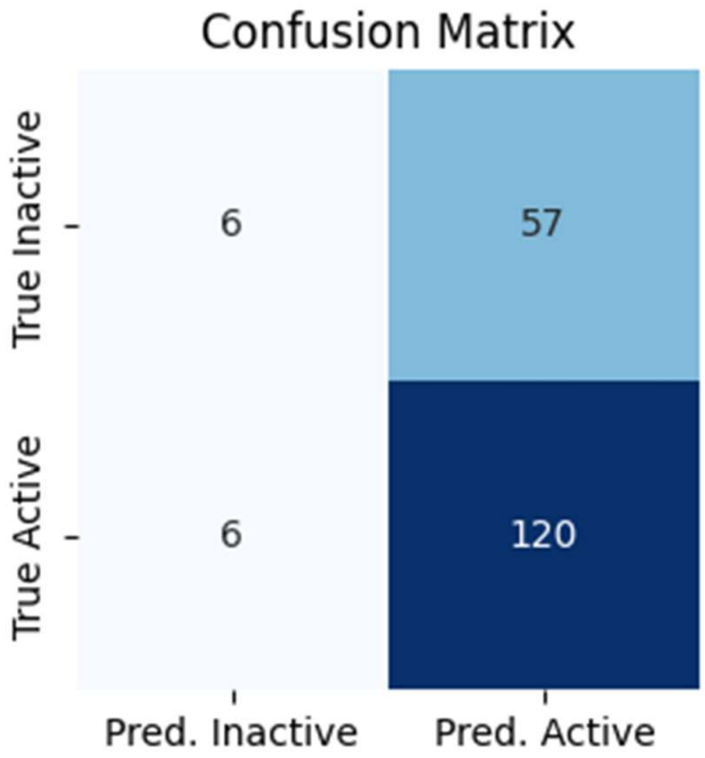
Workflow

- Load dataset with SMILES representation of the biomolecules
- Determine their fingerprints
- Replace 0 with NaN values for unknown values
- Train test split
- Create prediction with Random Forest for every task
- Evaluate the results
  - Confusion matrix for each task
  - Classification metrics for each task
  - ROC Curve / AUC value
- Readjust the hyperparameters for the Random Forest

Final Hyperparameter setting for Random Forest

- n_estimators=500
- max_depth = 30
- min_samples_split=2
- min_samples_leaf=1
- max_features='log2'

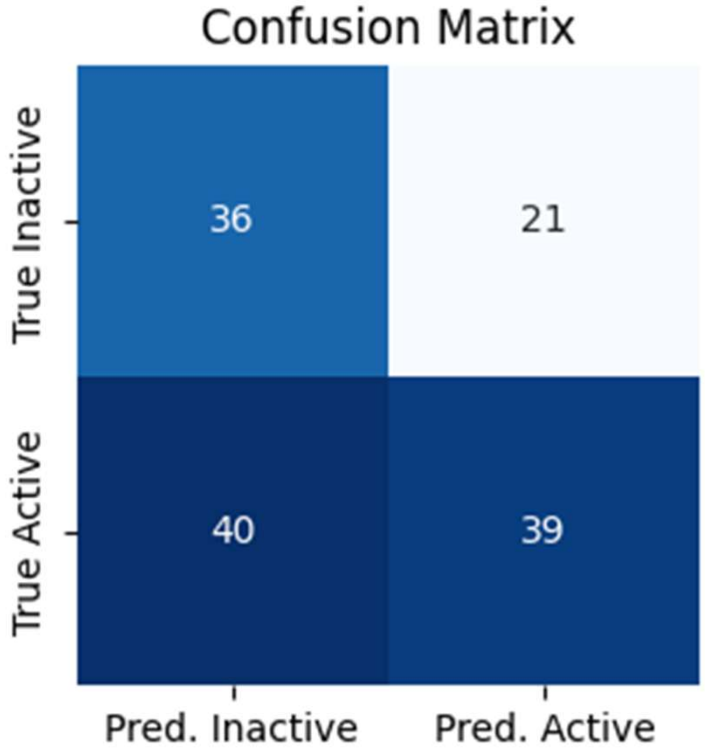
Confusion Matrix – Task 1



Confusion Matrix

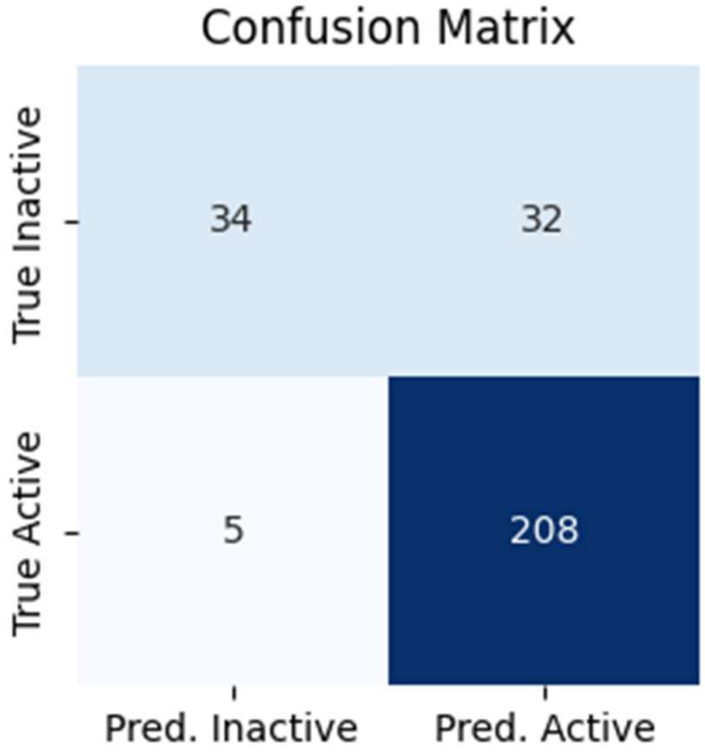
	Pred. Inactive	Pred. Active
True Inactive	6	57
True Active	6	120

Confusion Matrix – Task 2



	Pred. Inactive	Pred. Active
True Inactive	36	21
True Active	40	39

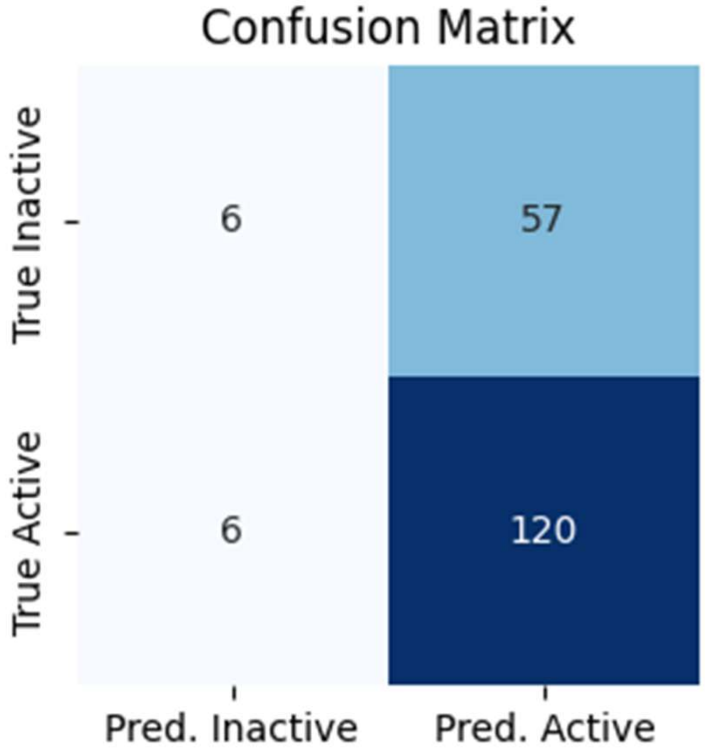
Confusion Matrix – Task 3



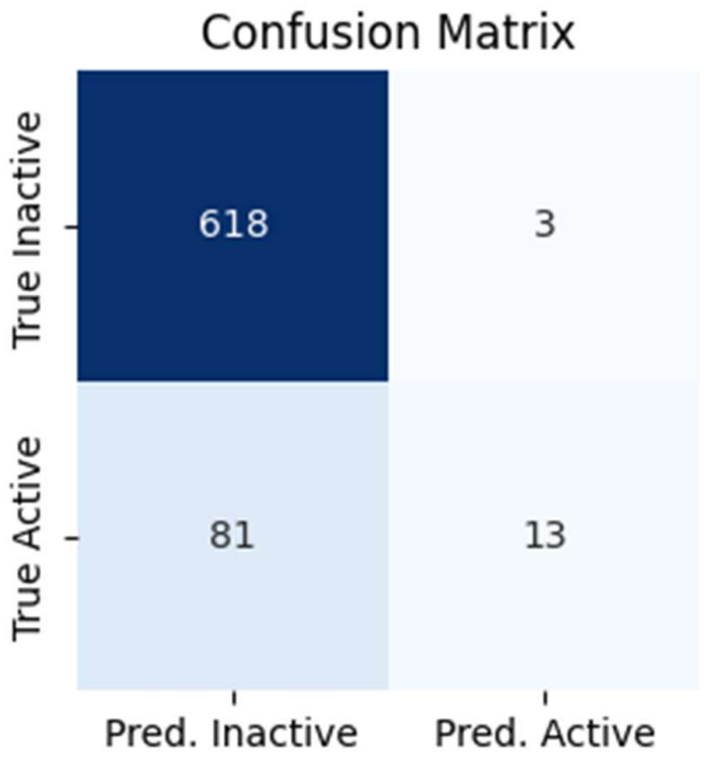
Confusion Matrix

	Pred. Inactive	Pred. Active
True Inactive	34	32
True Active	5	208

Confusion Matrix – Task 4



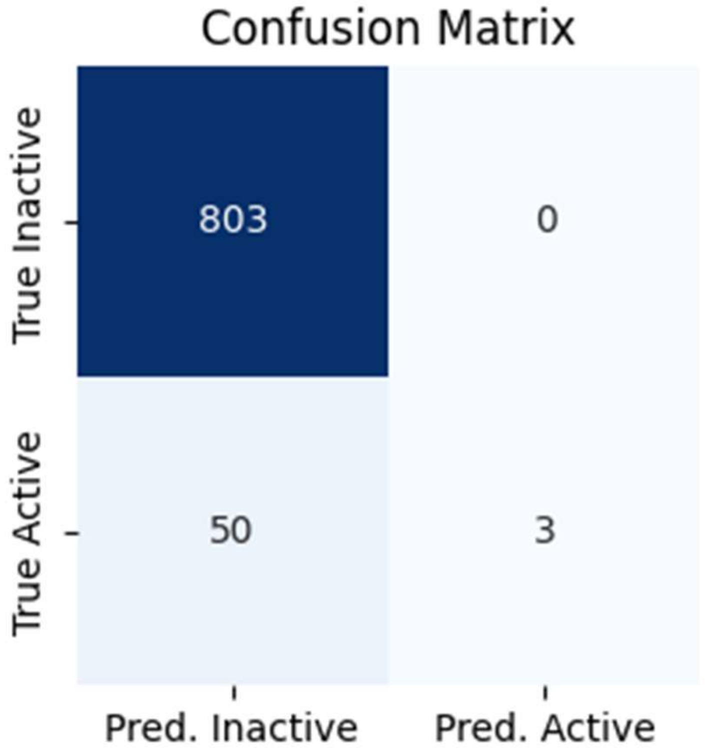
Confusion Matrix – Task 5



Confusion Matrix

	Pred. Inactive	Pred. Active
True Inactive	618	3
True Active	81	13

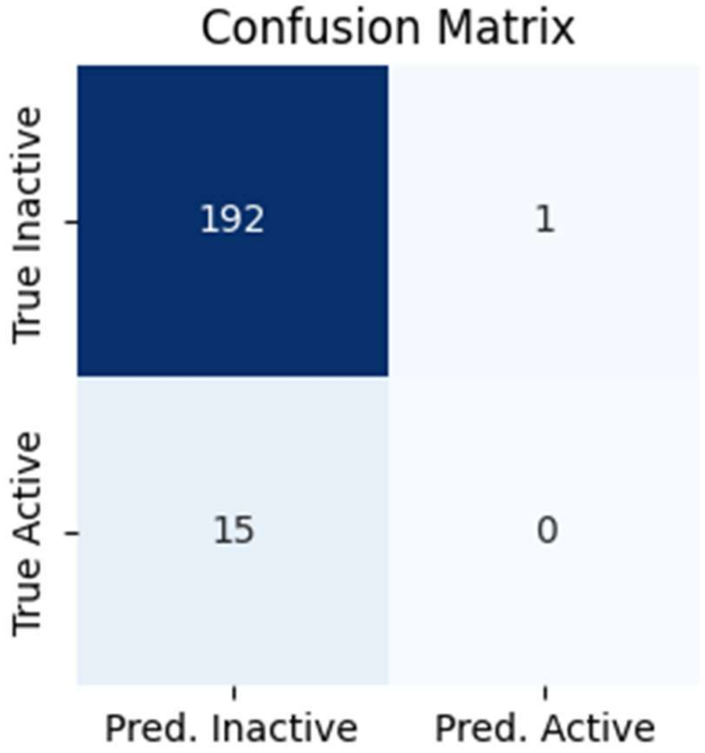
Confusion Matrix – Task 6



Confusion Matrix

	Pred. Inactive	Pred. Active
True Inactive	803	0
True Active	50	3

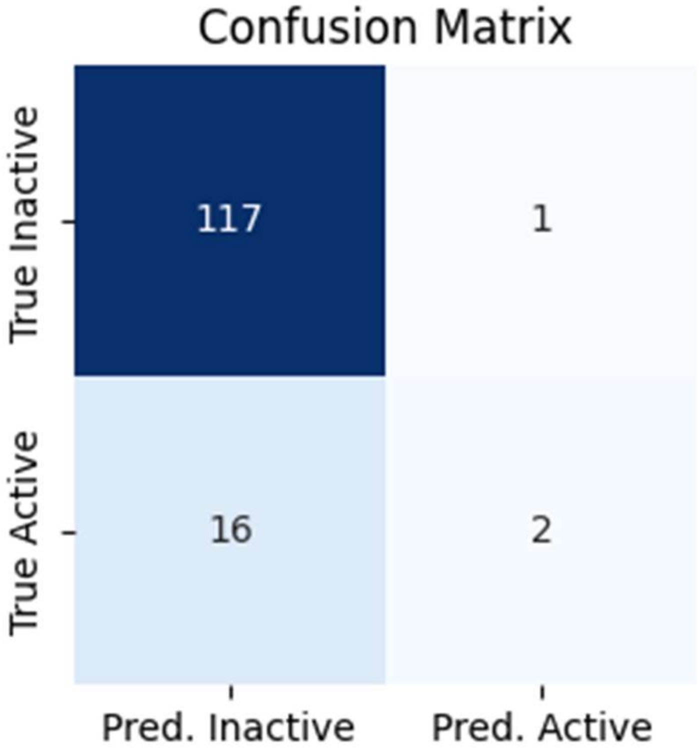
Confusion Matrix – Task 7



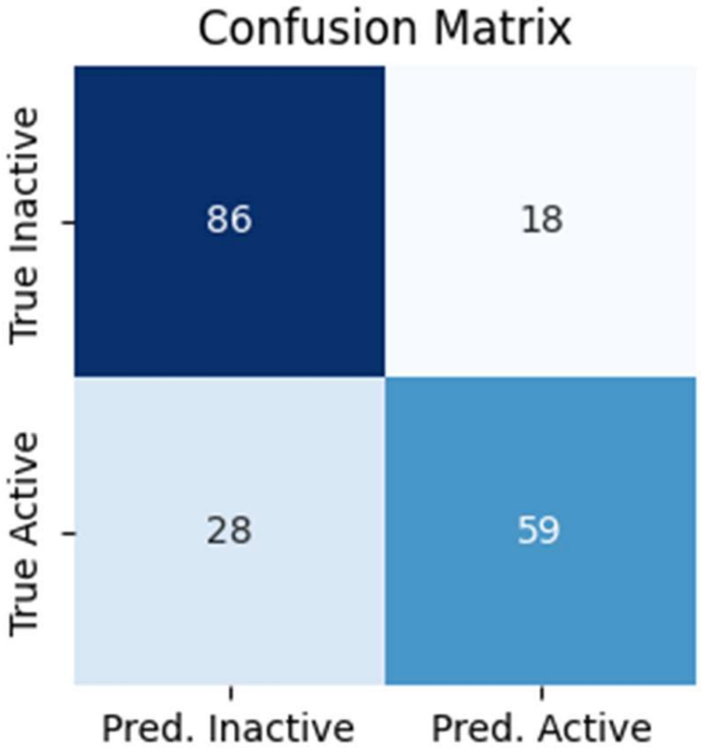
Confusion Matrix

	Pred. Inactive	Pred. Active
True Inactive	192	1
True Active	15	0

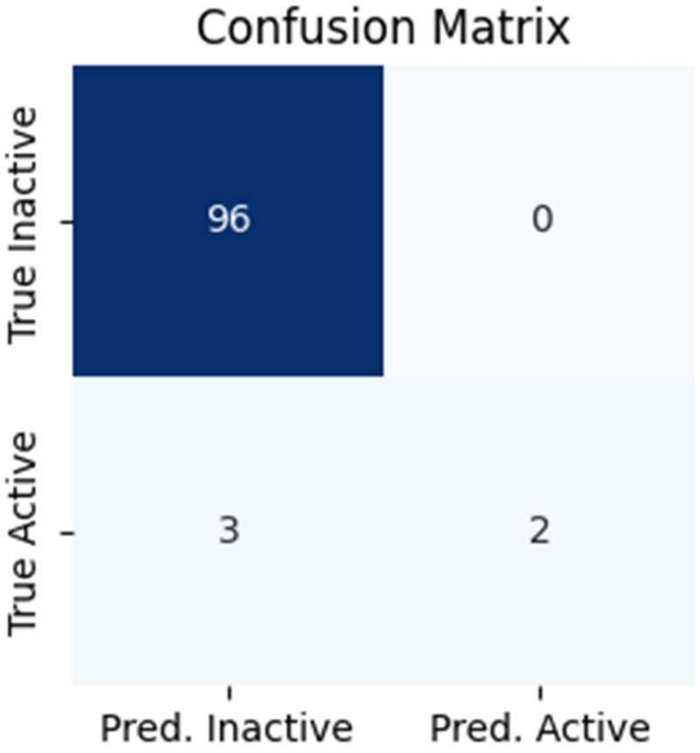
Confusion Matrix – Task 8



Confusion Matrix – Task 9

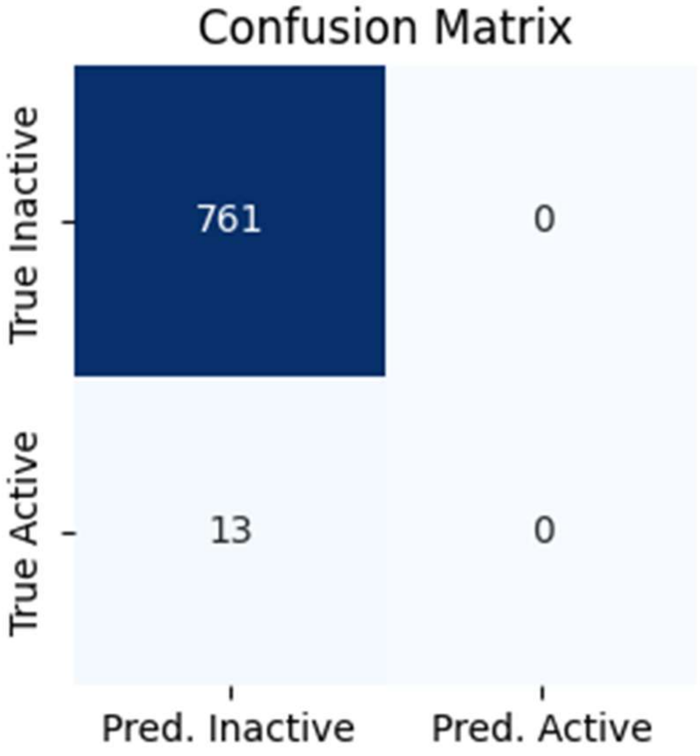


Confusion Matrix – Task 10



	Pred. Inactive	Pred. Active
True Inactive	96	0
True Active	3	2

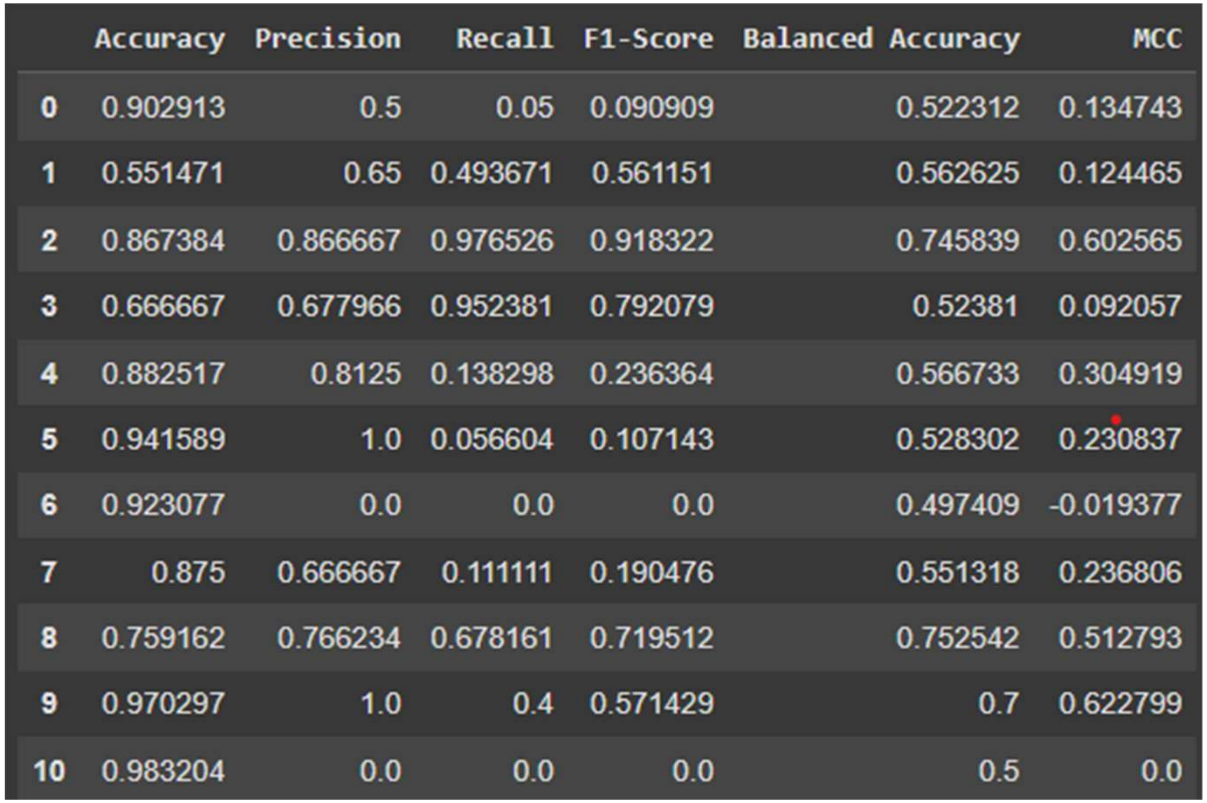
Confusion Matrix – Task 11



Confusion Matrix

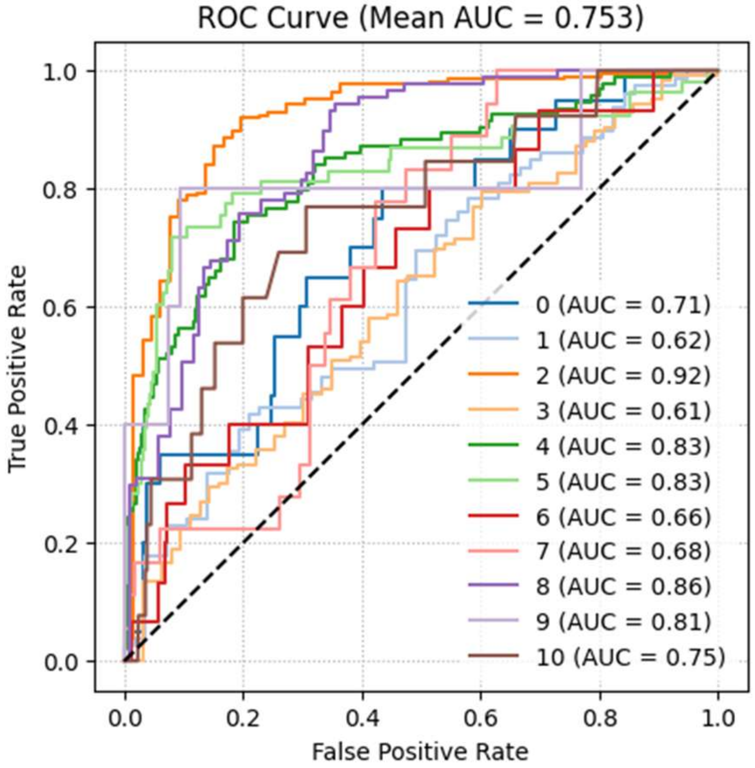
	Pred. Inactive	Pred. Active
True Inactive	761	0
True Active	13	0

Classification Metrics for each task



	Accuracy	Precision	Recall	F1-Score	Balanced Accuracy	MCC
0	0.902913	0.5	0.05	0.090909	0.522312	0.134743
1	0.551471	0.65	0.493671	0.561151	0.562625	0.124465
2	0.867384	0.866667	0.976526	0.918322	0.745839	0.602565
3	0.666667	0.677966	0.952381	0.792079	0.52381	0.092057
4	0.882517	0.8125	0.138298	0.236364	0.566733	0.304919
5	0.941589	1.0	0.056604	0.107143	0.528302	0.230837
6	0.923077	0.0	0.0	0.0	0.497409	-0.019377
7	0.875	0.666667	0.111111	0.190476	0.551318	0.236806
8	0.759162	0.766234	0.678161	0.719512	0.752542	0.512793
9	0.970297	1.0	0.4	0.571429	0.7	0.622799
10	0.983204	0.0	0.0	0.0	0.5	0.0

ROC Curve/AUC


ROC Curve (Mean AUC = 0.753)
True Positive Rate
False Positive Rate
0 (AUC = 0.71)
1 (AUC = 0.62)
2 (AUC = 0.92)
3 (AUC = 0.61)
4 (AUC = 0.83)
5 (AUC = 0.83)
6 (AUC = 0.66)
7 (AUC = 0.68)
8 (AUC = 0.86)
9 (AUC = 0.81)
10 (AUC = 0.75)