Import dataset and preprocess

```
from google.colab import files
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import pylab
from sklearn.model_selection import train_test_split
```

```
def import_dataset():
    uploaded = files.upload()

    for fn in uploaded.keys():
        print('User uploaded file "{name}" with length {length} bytes'.format(
        name=fn, length=len(uploaded[fn])))

    with open(fn, 'r') as opened_file:
        txt_lines = opened_file.readlines()

    sequences = []
    label_array = np.zeros(len(txt_lines))
    for i in range (len(txt_lines)):
        line_elements = txt_lines[i].split('\t')
        label_array[i] = int(line_elements[0])
        seq = list(line_elements[1][:-1])
        sequences.append(seq)

    raw_dataset = pd.DataFrame(data=sequences, columns=range(1,201))
    raw_dataset['label'] = label_array
    return raw_dataset
```

```
raw_dataset = import_dataset()
[raw_train, raw_test] = train_test_split(raw_dataset,train_size=0.8, random_state=103)
```

Train SVM model

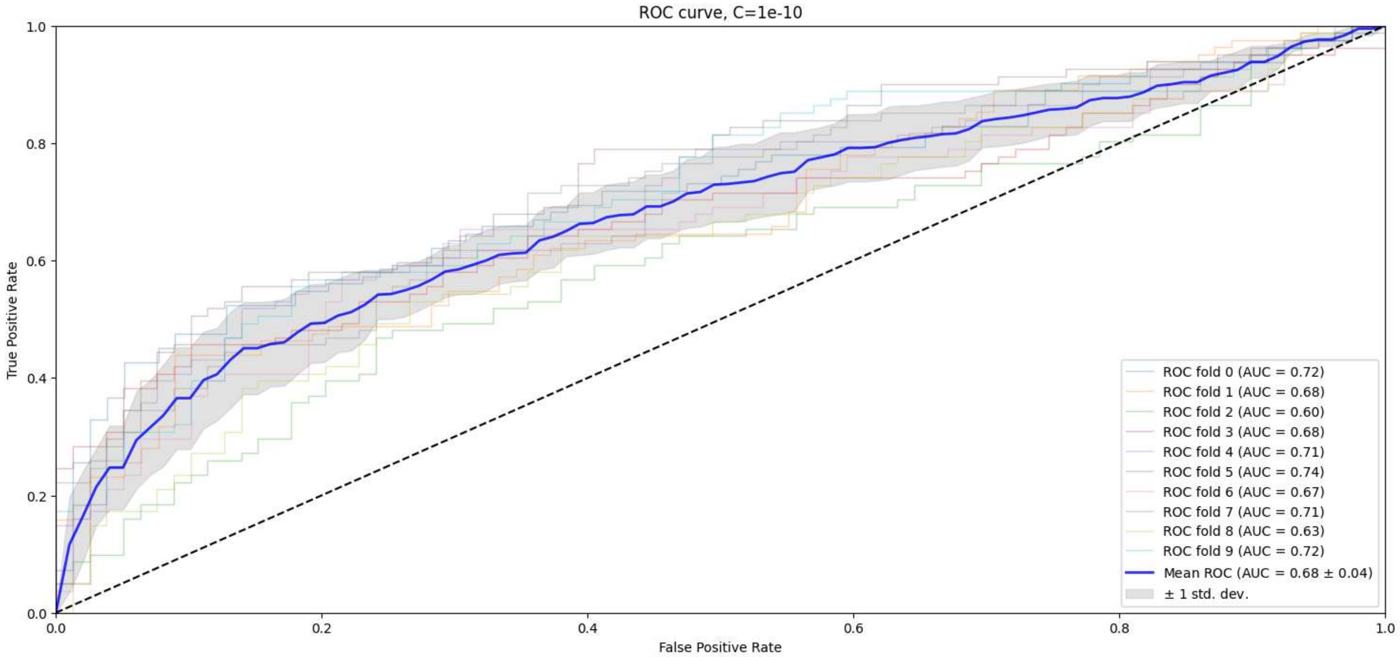
Analyze results with ROC curve

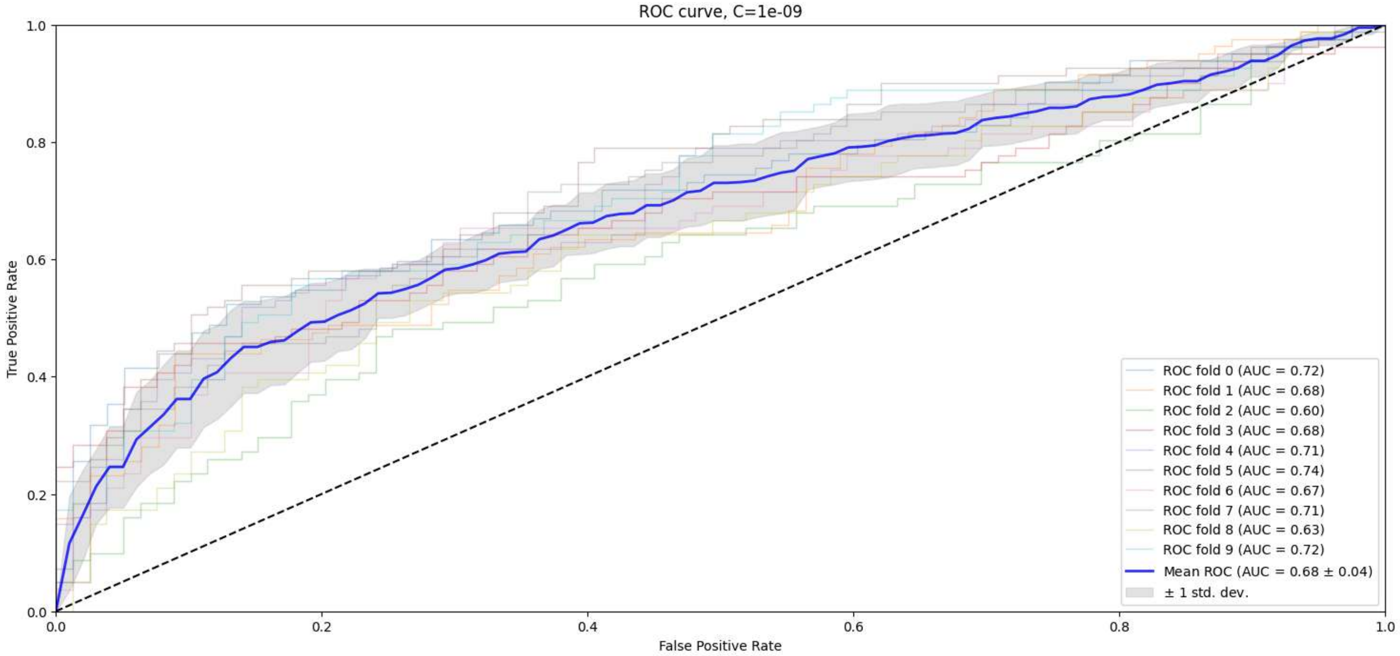
[ ] ↳ *1 cell hidden*

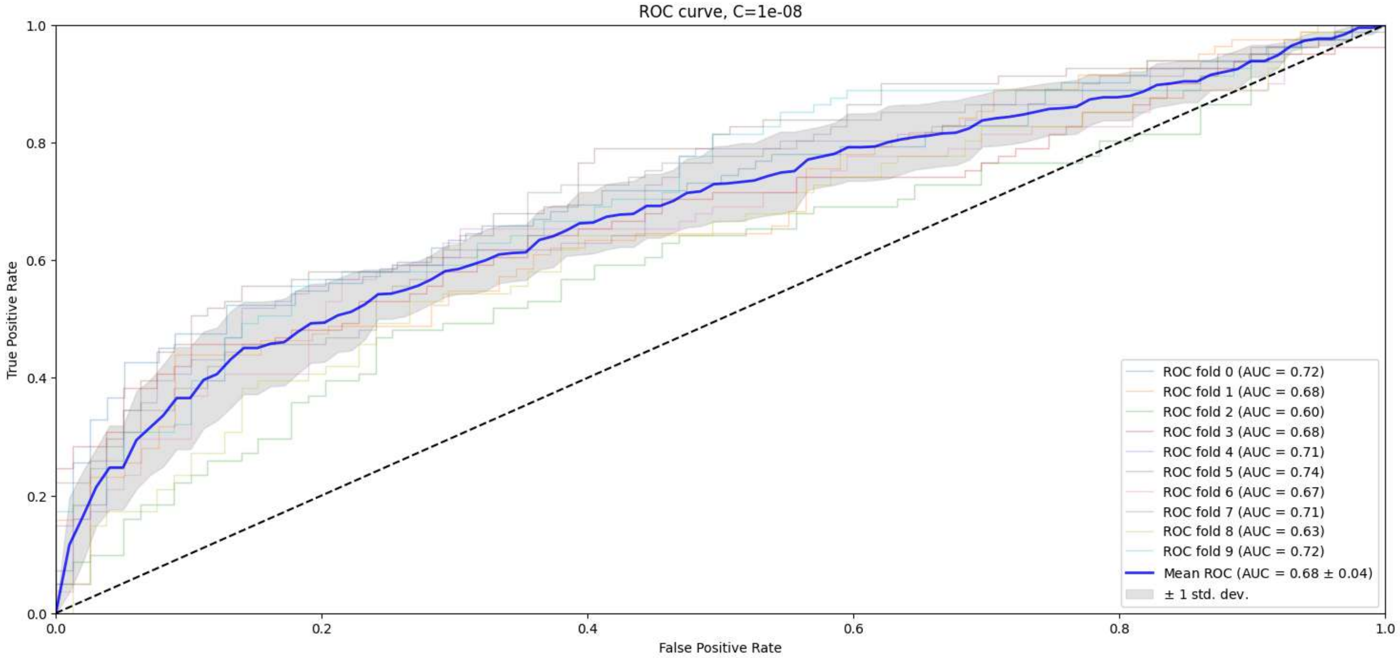
R^2 representation

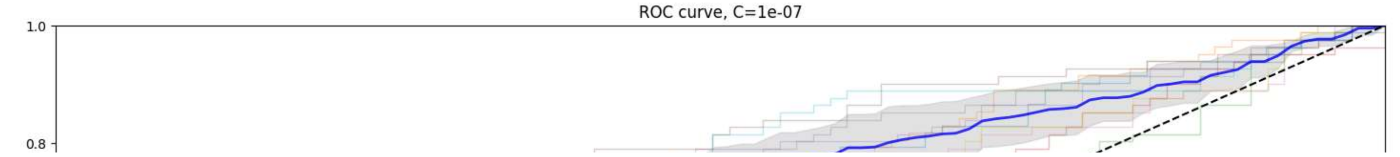
```
chars = ['A', 'a', 'C', 'c', 'G', 'g', 'T', 't']
real_values = [1, 1/(2 ** 0.5), 0, -1/(2 ** 0.5), -1, -1/(2 ** 0.5), 0, 1/(2 ** 0.5)]
imag_values = [0, 1/(2 ** 0.5), 1, 1/(2 ** 0.5), 0, -1/(2 ** 0.5), -1, -1/(2 ** 0.5)]
new_column_names = range(201,402)

train_real_values = raw_train.replace(chars, real_values)
train_imag_values = raw_train.replace(chars, imag_values)
train_imag_values.rename(columns=dict(zip(train_imag_values.columns, new_column_names)), inplace=T
train_imag_values.rename(columns={401: 'label'}, inplace=True)
train = pd.concat([train_real_values.iloc[:, 0:200], train_imag_values], axis=1)
X_train = train.iloc[:, 0:400].to_numpy()
Y_train = train['label'].to_numpy()

test_real_values = raw_test.replace(chars, real_values)
test_imag_values = raw_test.replace(chars, imag_values)
test_imag_values.rename(columns=dict(zip(test_imag_values.columns, new_column_names)), inplace=Tru
test_imag_values.rename(columns={401: 'label'}, inplace=True)
test = pd.concat([test_real_values.iloc[:, 0:200], test_imag_values], axis=1)
X_test = test.iloc[:, 0:400].to_numpy()
Y_test = test['label'].to_numpy()

# print(X_train)
```

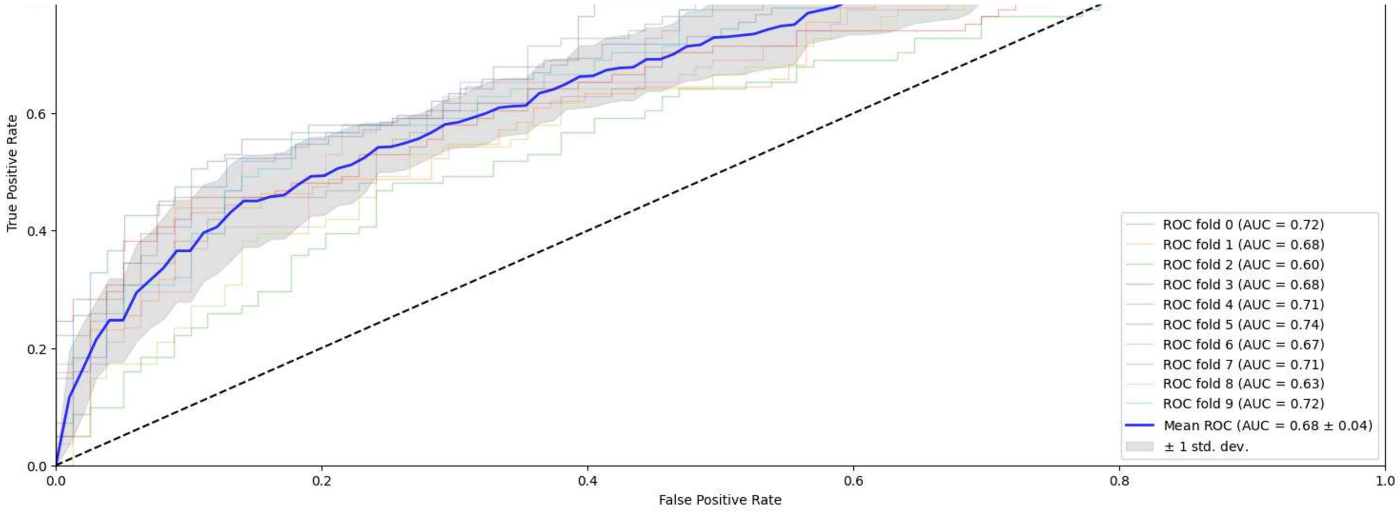
```
n_splits = 10
for c in np.logspace(-10,0,11):
    classifier = svm.SVC(C=c, kernel="linear")
    plot_roc_with_cv(classifier, X_train, Y_train, n_splits, f"ROC curve, C={c}")
```


ROC curve, C=1e-10
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.72)
ROC fold 1 (AUC = 0.68)
ROC fold 2 (AUC = 0.60)
ROC fold 3 (AUC = 0.68)
ROC fold 4 (AUC = 0.71)
ROC fold 5 (AUC = 0.74)
ROC fold 6 (AUC = 0.67)
ROC fold 7 (AUC = 0.71)
ROC fold 8 (AUC = 0.63)
ROC fold 9 (AUC = 0.72)
Mean ROC (AUC = 0.68 ± 0.04)
± 1 std. dev.

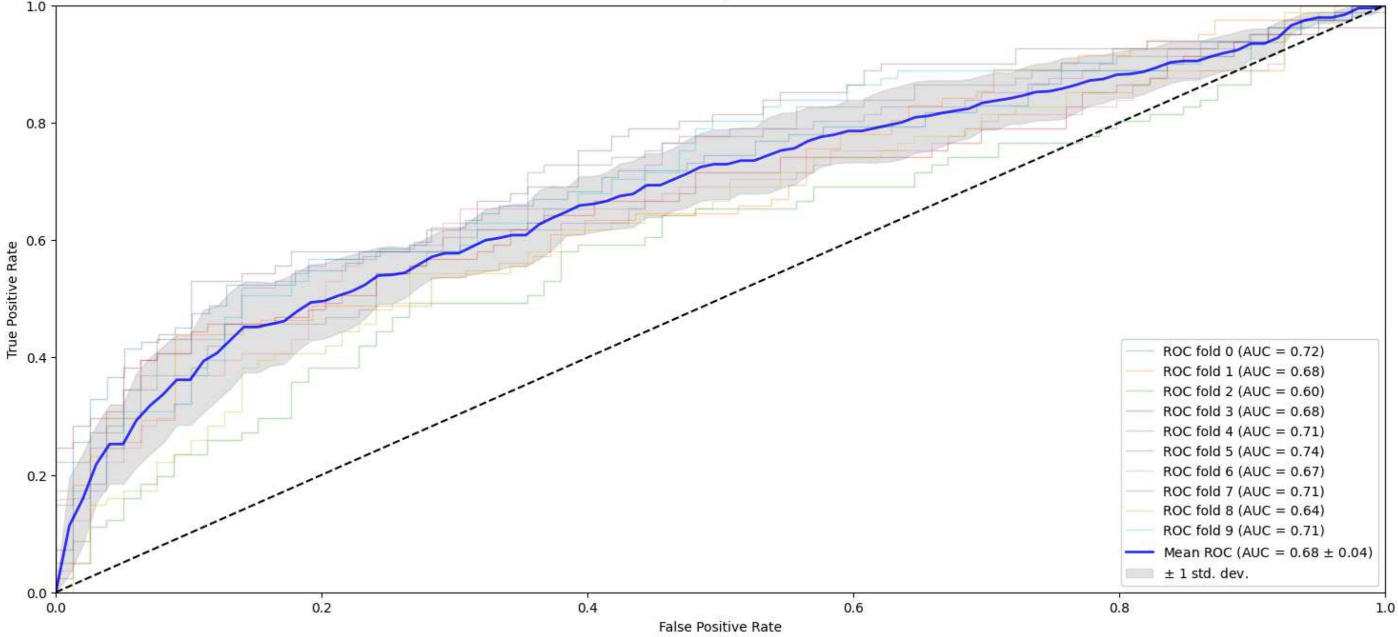

ROC curve, C=1e-09
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.72)
ROC fold 1 (AUC = 0.68)
ROC fold 2 (AUC = 0.60)
ROC fold 3 (AUC = 0.68)
ROC fold 4 (AUC = 0.71)
ROC fold 5 (AUC = 0.74)
ROC fold 6 (AUC = 0.67)
ROC fold 7 (AUC = 0.71)
ROC fold 8 (AUC = 0.63)
ROC fold 9 (AUC = 0.72)
Mean ROC (AUC = 0.68 ± 0.04)
± 1 std. dev.


ROC curve, C=1e-08
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.72)
ROC fold 1 (AUC = 0.68)
ROC fold 2 (AUC = 0.60)
ROC fold 3 (AUC = 0.68)
ROC fold 4 (AUC = 0.71)
ROC fold 5 (AUC = 0.74)
ROC fold 6 (AUC = 0.67)
ROC fold 7 (AUC = 0.71)
ROC fold 8 (AUC = 0.63)
ROC fold 9 (AUC = 0.72)
Mean ROC (AUC = 0.68 ± 0.04)
± 1 std. dev.

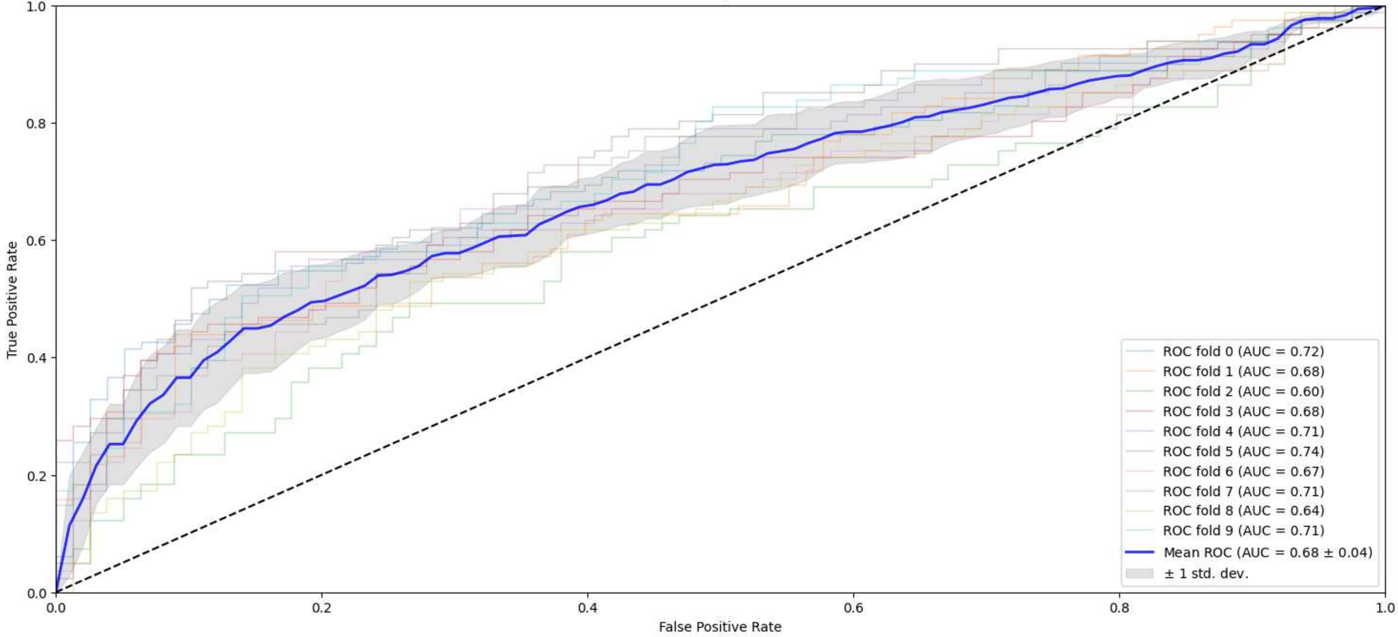

ROC curve, C=1e-07
1.0
0.8



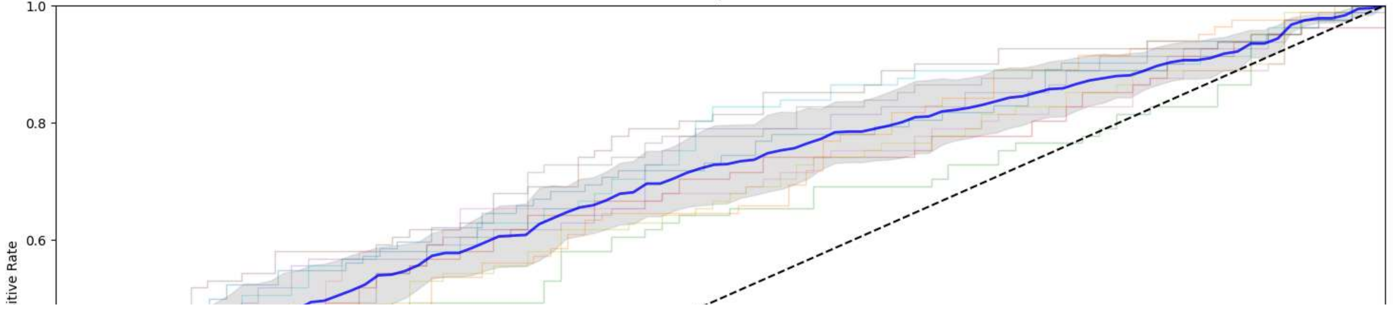
ROC curve, C=1e-06

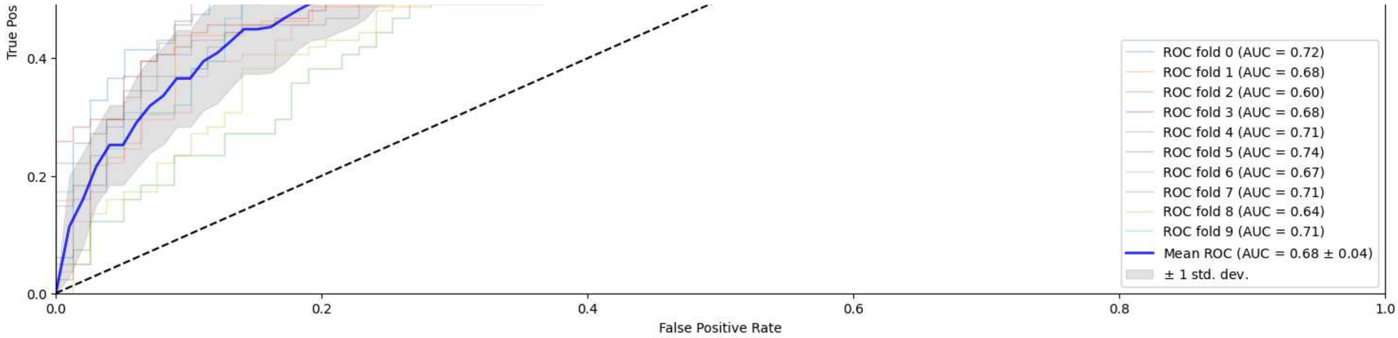


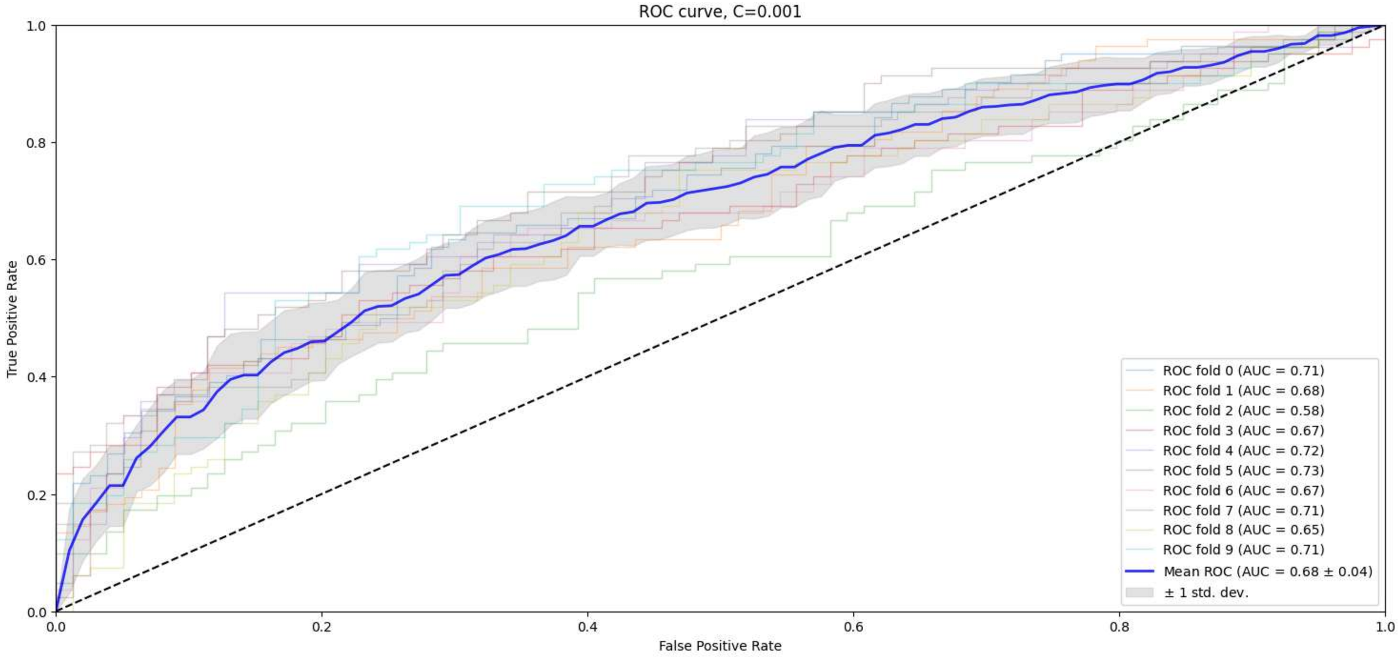
ROC curve, C=1e-05

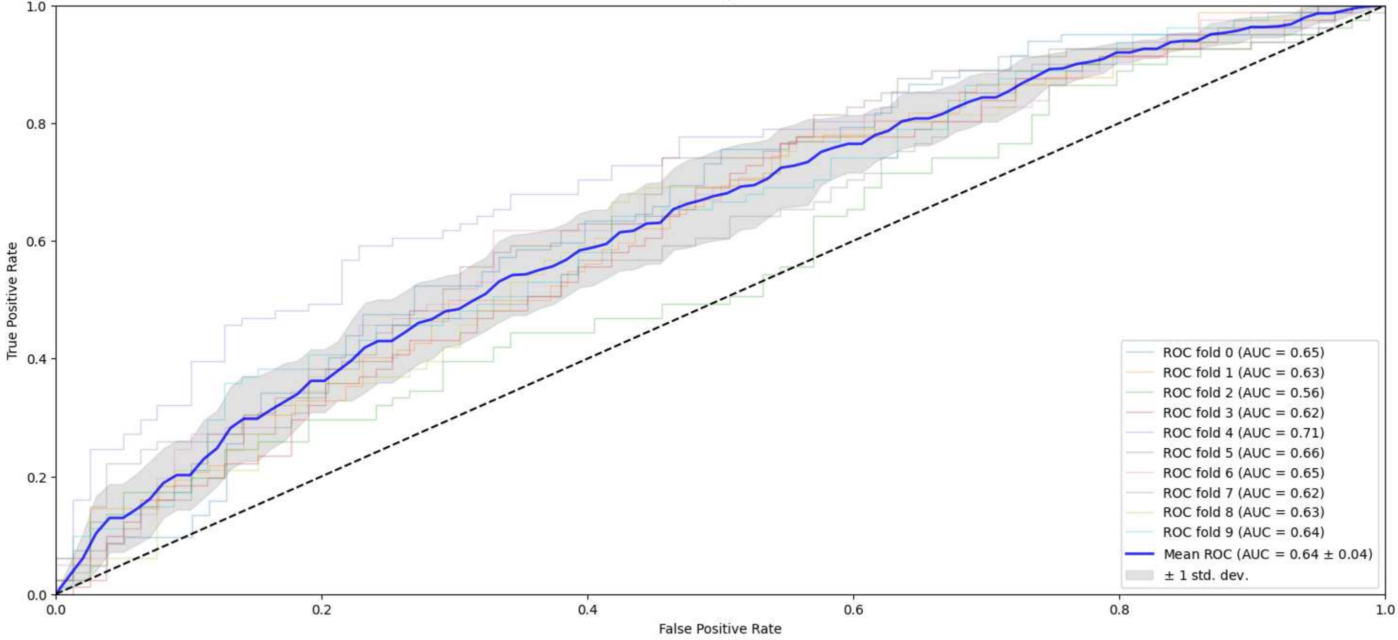


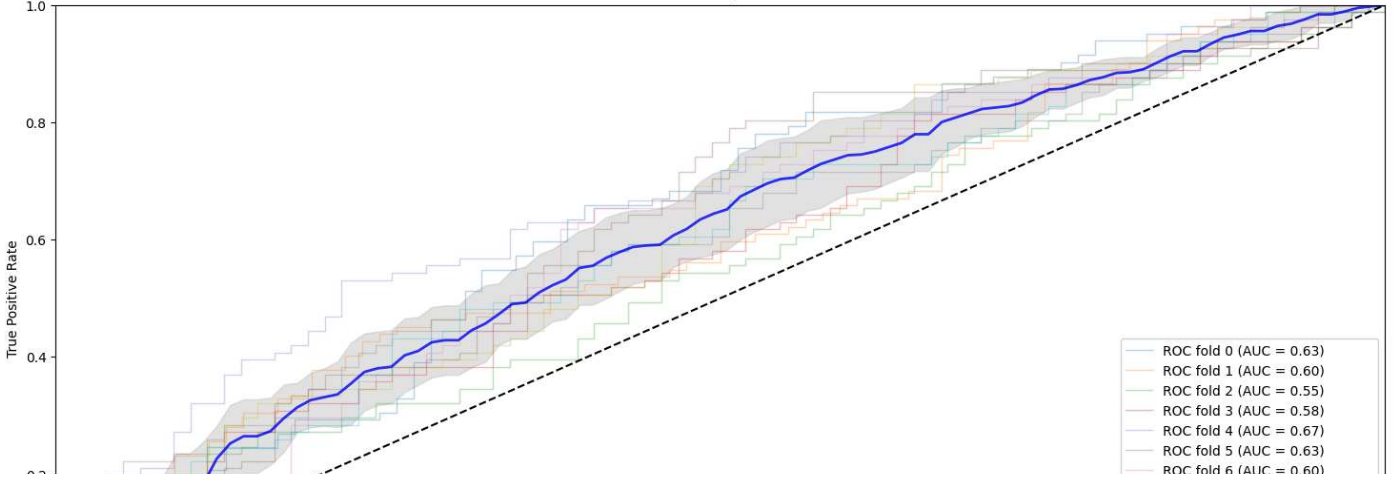
ROC curve, C=0.0001

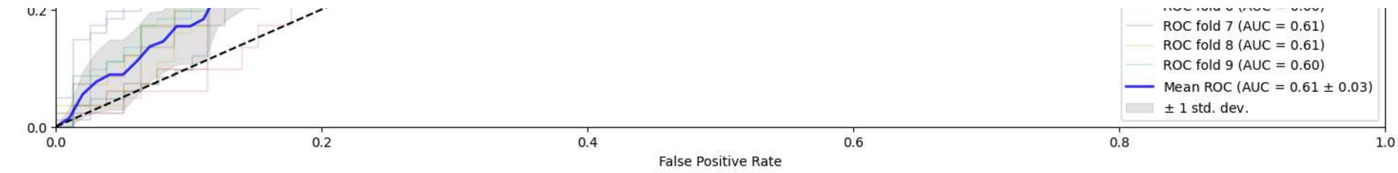


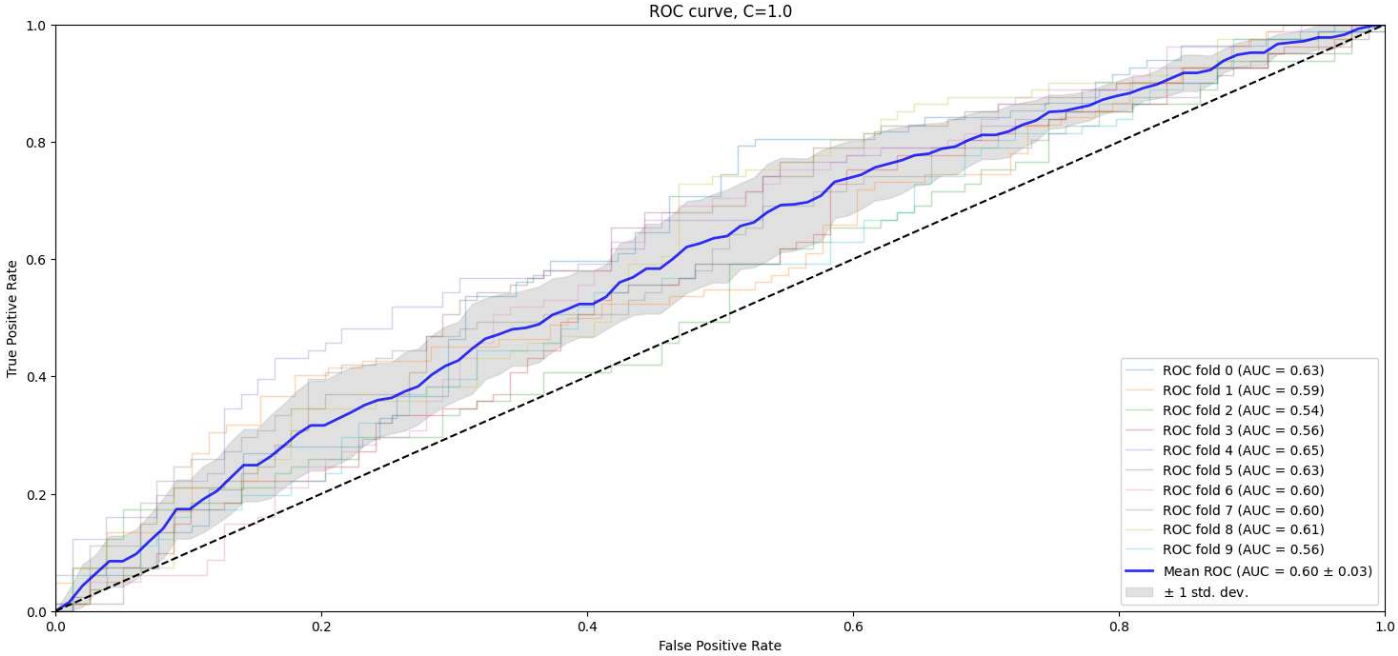



ROC curve, C=0.001


ROC curve, C=0.01


ROC curve, C=0.1

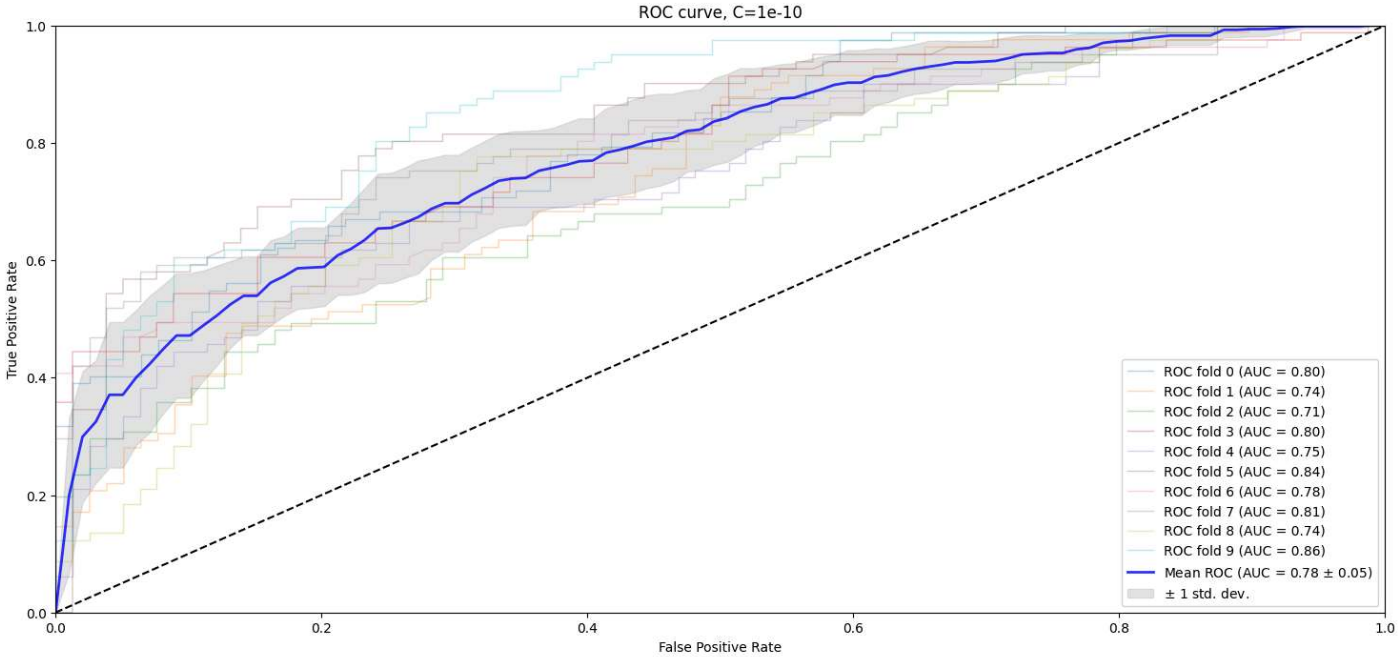

0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC fold 6 (AUC = 0.60)
ROC fold 7 (AUC = 0.61)
ROC fold 8 (AUC = 0.61)
ROC fold 9 (AUC = 0.60)
Mean ROC (AUC = 0.61 ± 0.03)
± 1 std. dev.

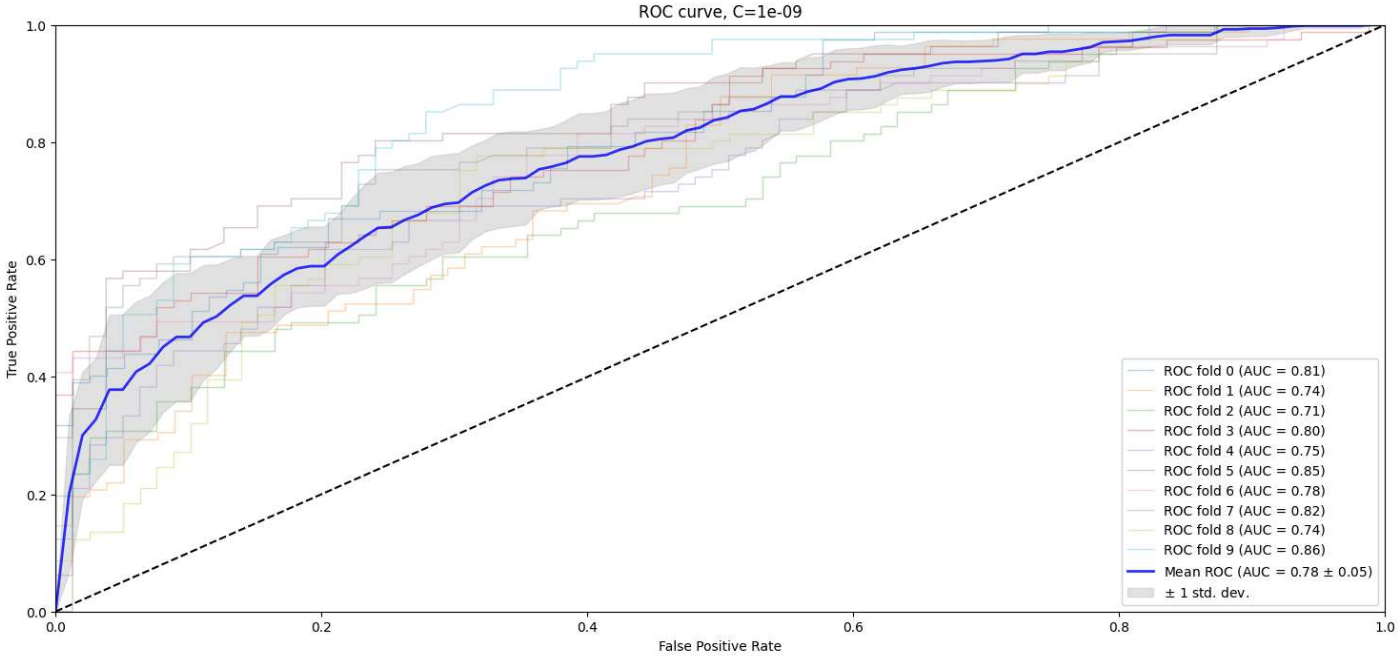

ROC curve, C=1.0
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC fold 0 (AUC = 0.63)
ROC fold 1 (AUC = 0.59)
ROC fold 2 (AUC = 0.54)
ROC fold 3 (AUC = 0.56)
ROC fold 4 (AUC = 0.65)
ROC fold 5 (AUC = 0.63)
ROC fold 6 (AUC = 0.60)
ROC fold 7 (AUC = 0.60)
ROC fold 8 (AUC = 0.61)
ROC fold 9 (AUC = 0.56)
Mean ROC (AUC = 0.60 ± 0.03)
± 1 std. dev.

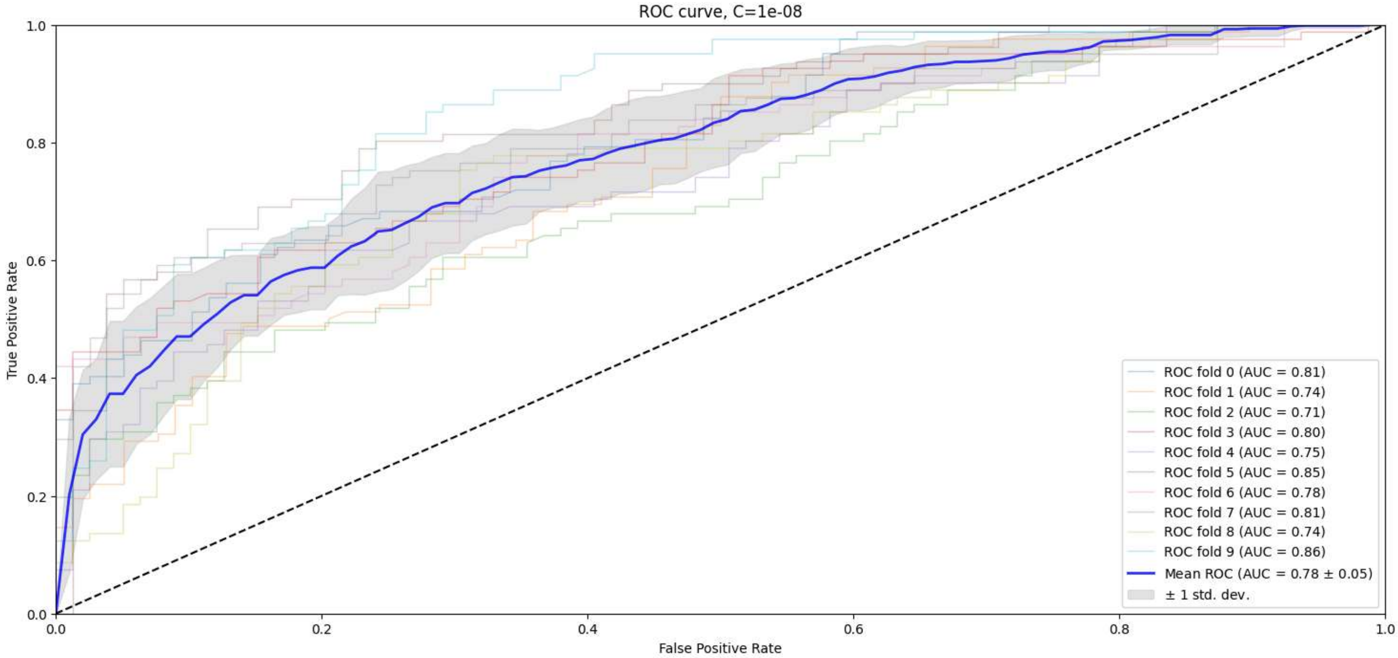
One-Hot Encoding

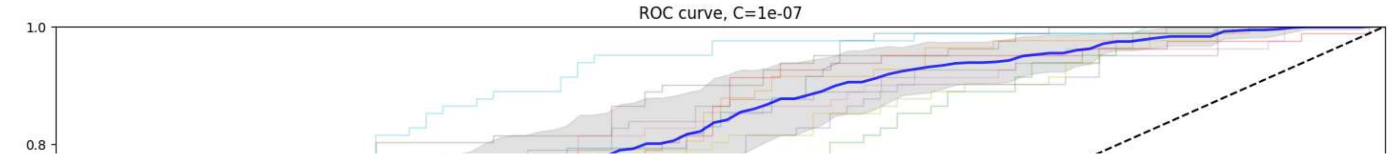
```
X_train = pd.get_dummies(raw_train.iloc[:, 0:200]).to_numpy()
Y_train = raw_train['label'].to_numpy()

X_test = pd.get_dummies(raw_test.iloc[:, 0:200]).to_numpy()
Y_test = raw_test['label'].to_numpy()
# print(X_train)
```

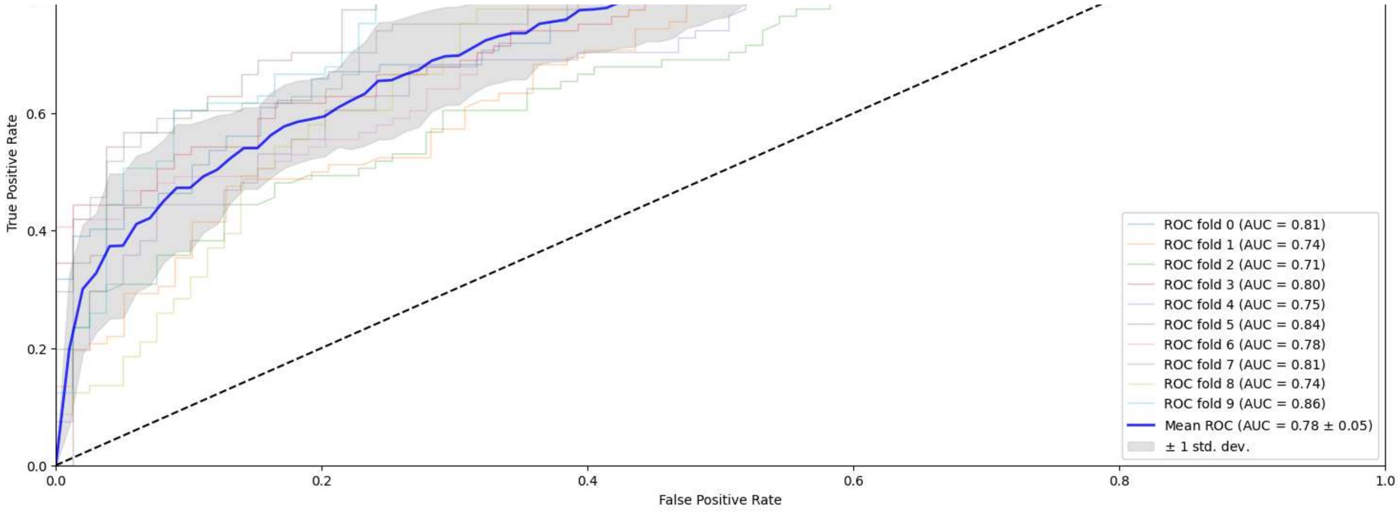
```
n_splits = 10
for c in np.logspace(-10,2,13):
    classifier = svm.SVC(C=c, kernel="linear")
    plot_roc_with_cv(classifier, X_train, Y_train, n_splits, f"ROC curve, C={c}")
```

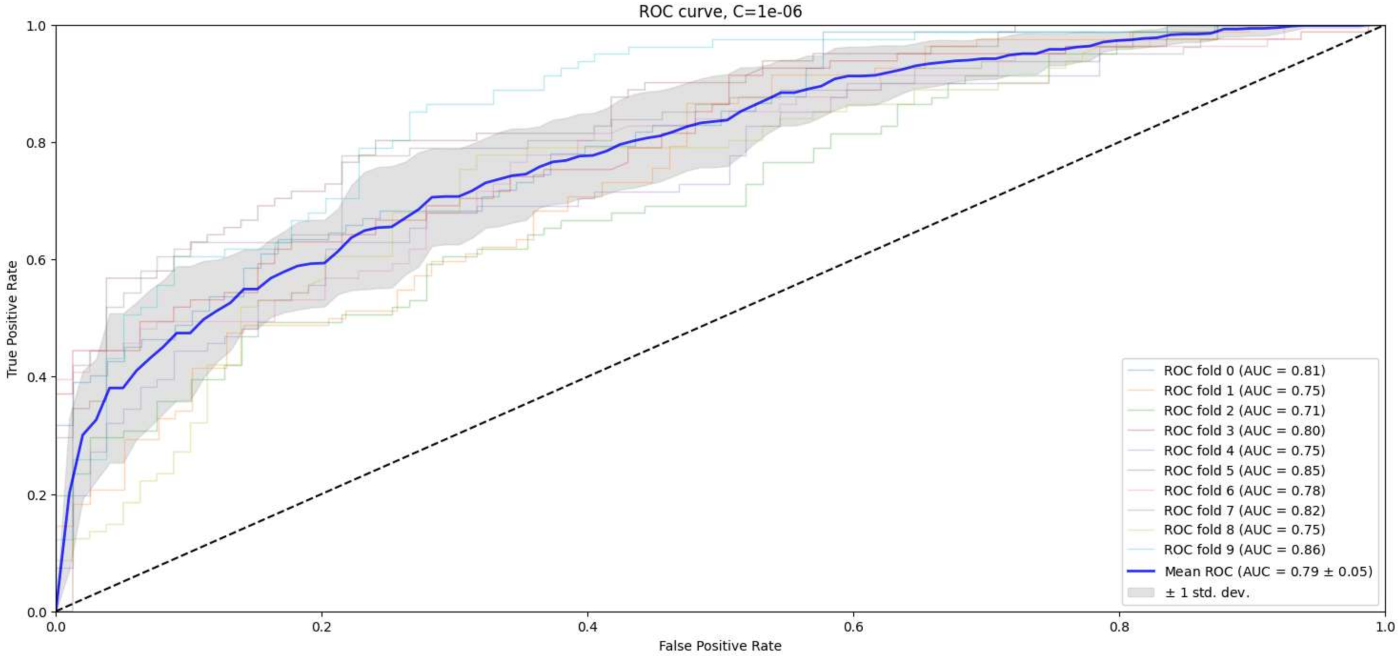

ROC curve, C=1e-10
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.80)
ROC fold 1 (AUC = 0.74)
ROC fold 2 (AUC = 0.71)
ROC fold 3 (AUC = 0.80)
ROC fold 4 (AUC = 0.75)
ROC fold 5 (AUC = 0.84)
ROC fold 6 (AUC = 0.78)
ROC fold 7 (AUC = 0.81)
ROC fold 8 (AUC = 0.74)
ROC fold 9 (AUC = 0.86)
Mean ROC (AUC = 0.78 ± 0.05)
± 1 std. dev.

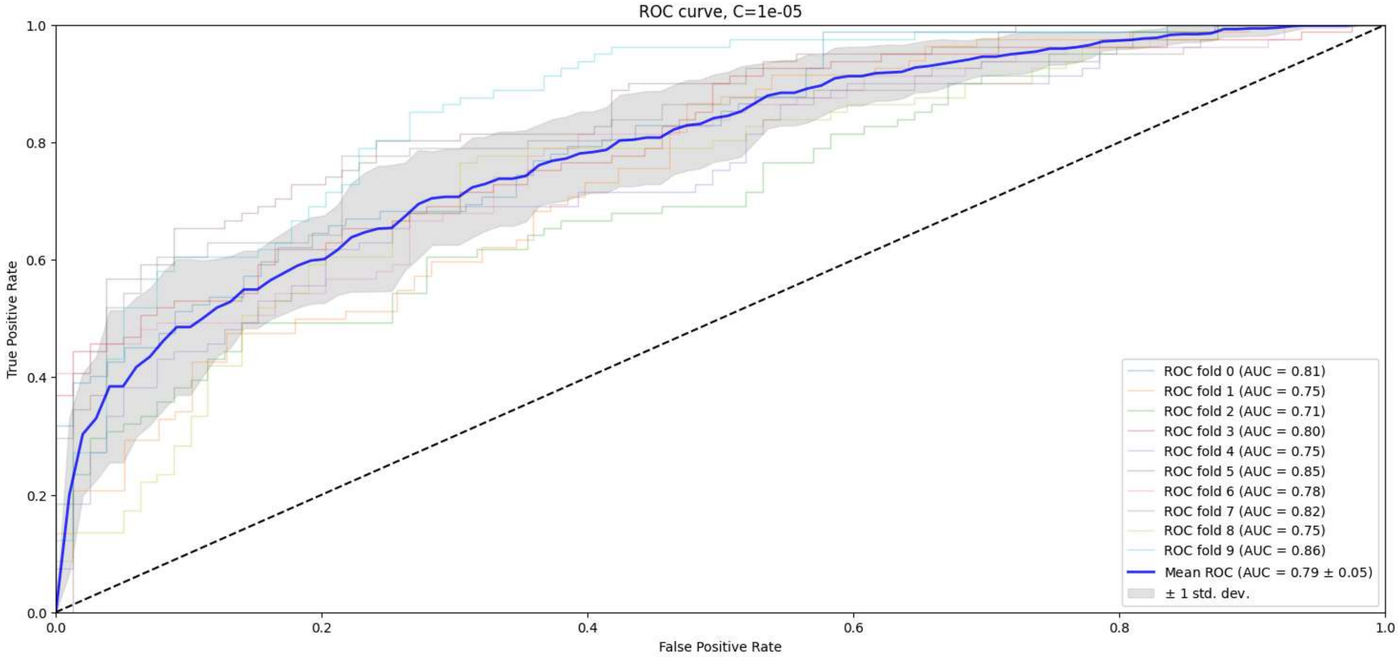

ROC curve, C=1e-09
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.81)
ROC fold 1 (AUC = 0.74)
ROC fold 2 (AUC = 0.71)
ROC fold 3 (AUC = 0.80)
ROC fold 4 (AUC = 0.75)
ROC fold 5 (AUC = 0.85)
ROC fold 6 (AUC = 0.78)
ROC fold 7 (AUC = 0.82)
ROC fold 8 (AUC = 0.74)
ROC fold 9 (AUC = 0.86)
Mean ROC (AUC = 0.78 ± 0.05)
± 1 std. dev.

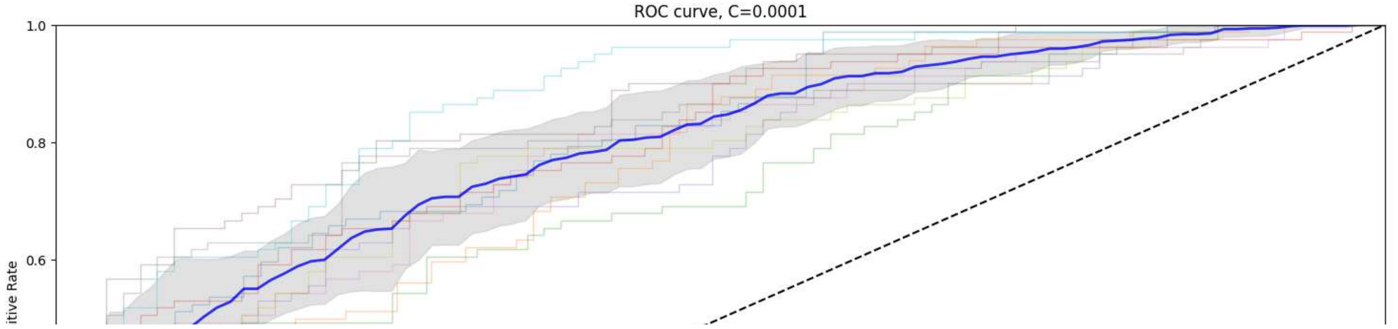

ROC curve, C=1e-08
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.81)
ROC fold 1 (AUC = 0.74)
ROC fold 2 (AUC = 0.71)
ROC fold 3 (AUC = 0.80)
ROC fold 4 (AUC = 0.75)
ROC fold 5 (AUC = 0.85)
ROC fold 6 (AUC = 0.78)
ROC fold 7 (AUC = 0.81)
ROC fold 8 (AUC = 0.74)
ROC fold 9 (AUC = 0.86)
Mean ROC (AUC = 0.78 ± 0.05)
± 1 std. dev.

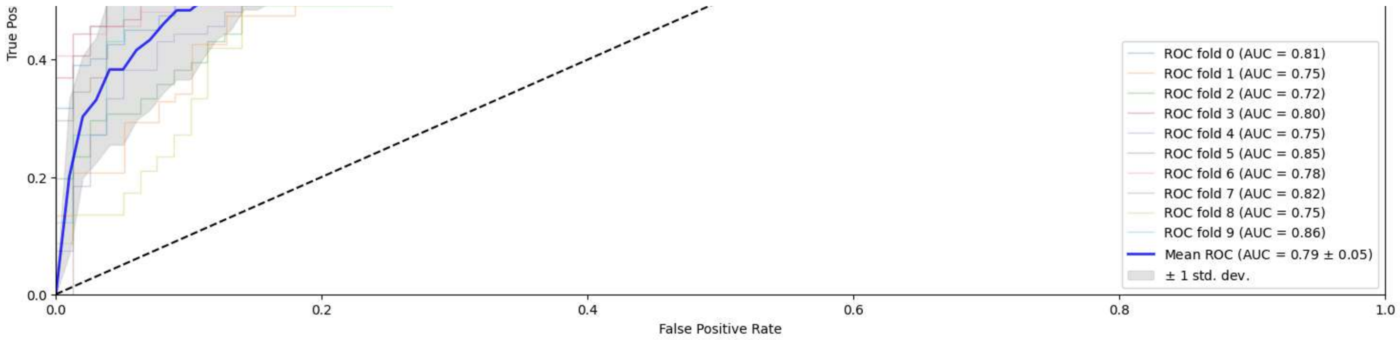

ROC curve, C=1e-07
1.0
0.8

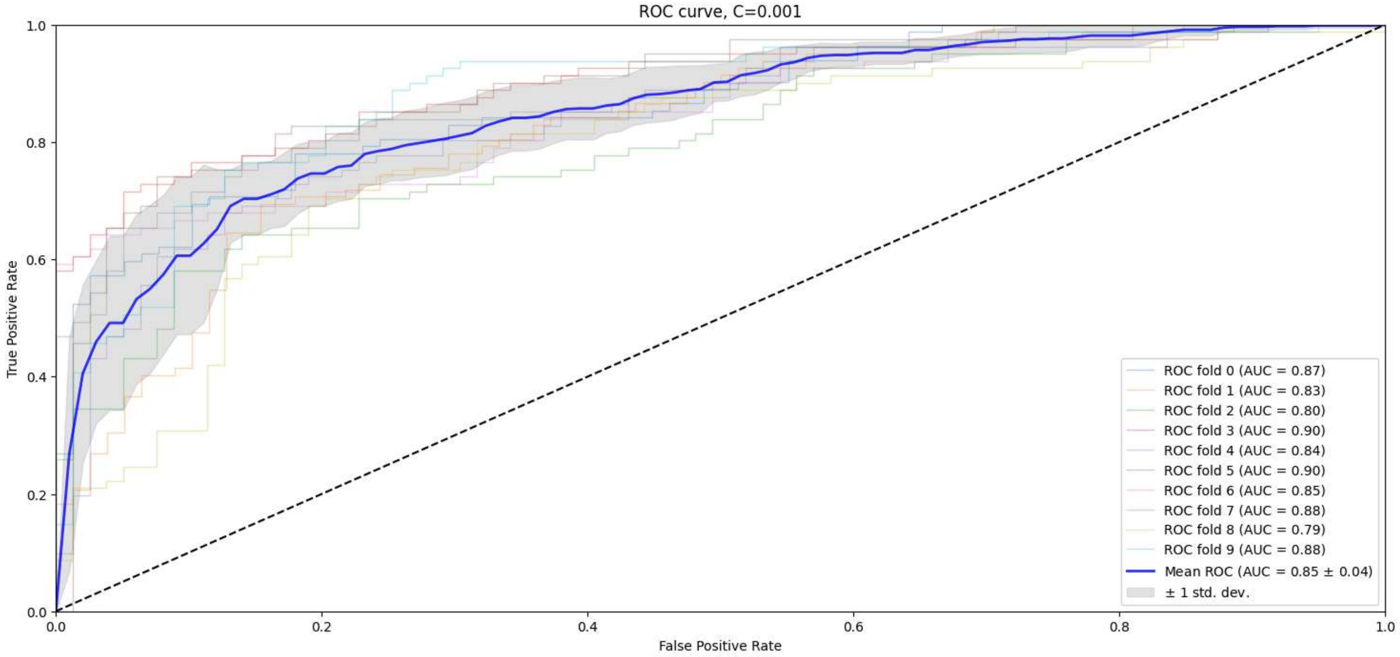


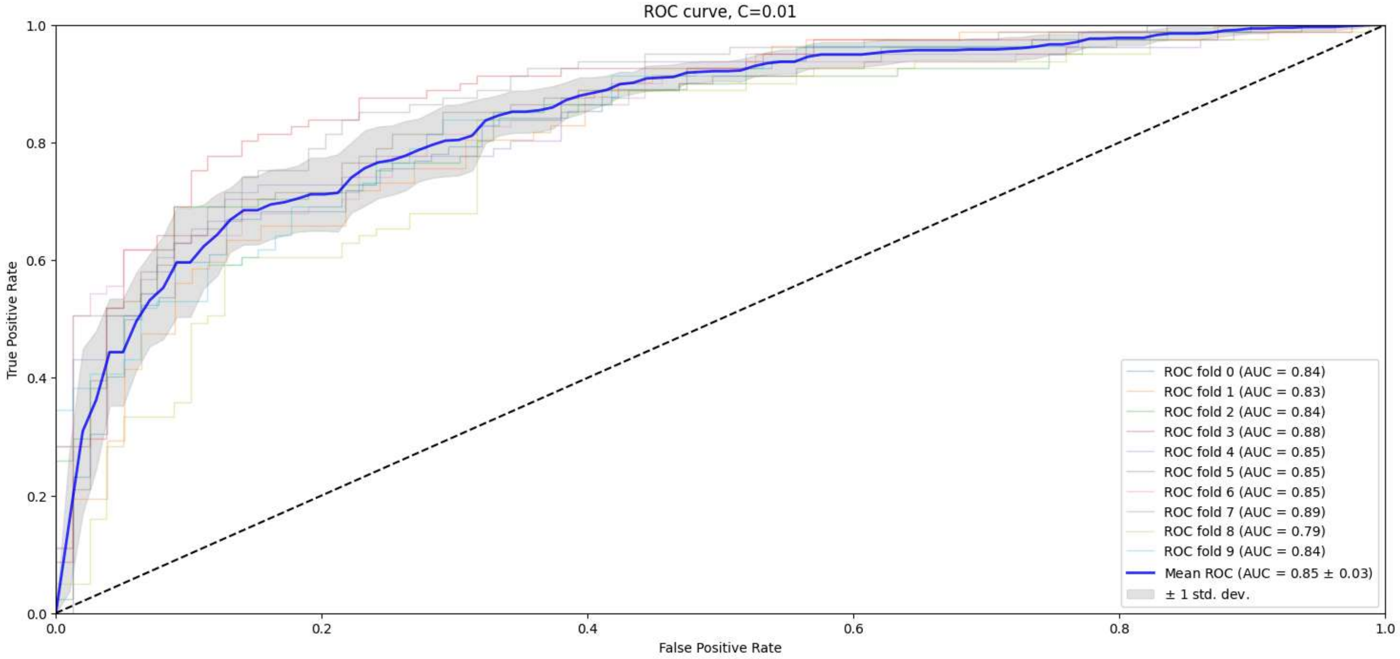

ROC curve, C=1e-06

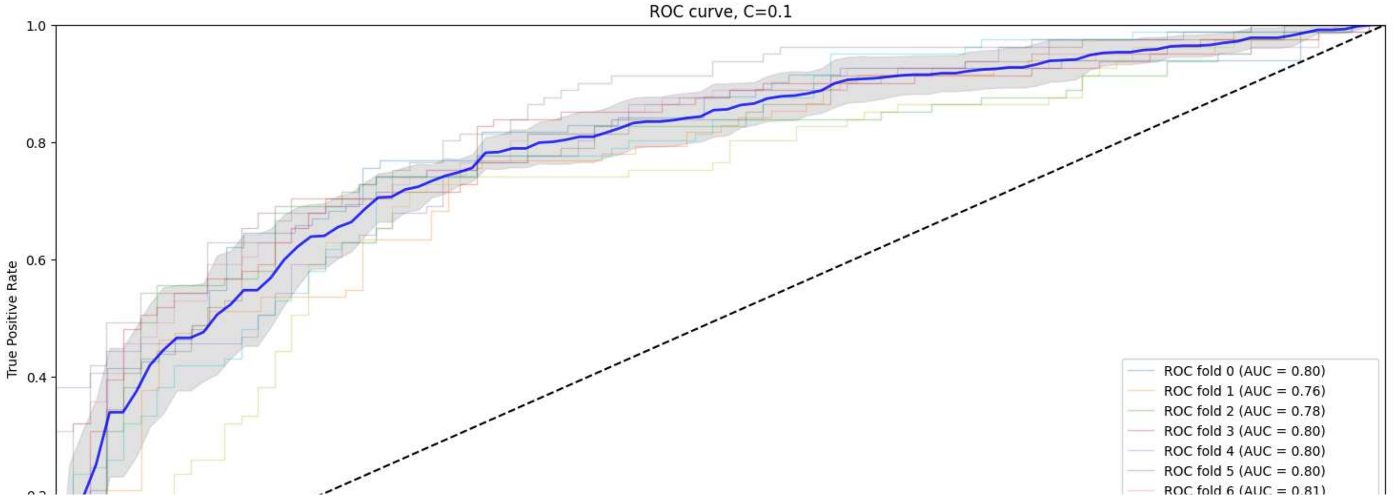

ROC curve, C=1e-05

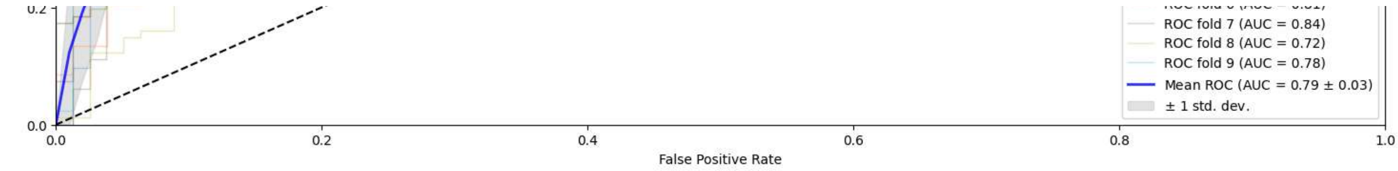

ROC curve, C=0.0001

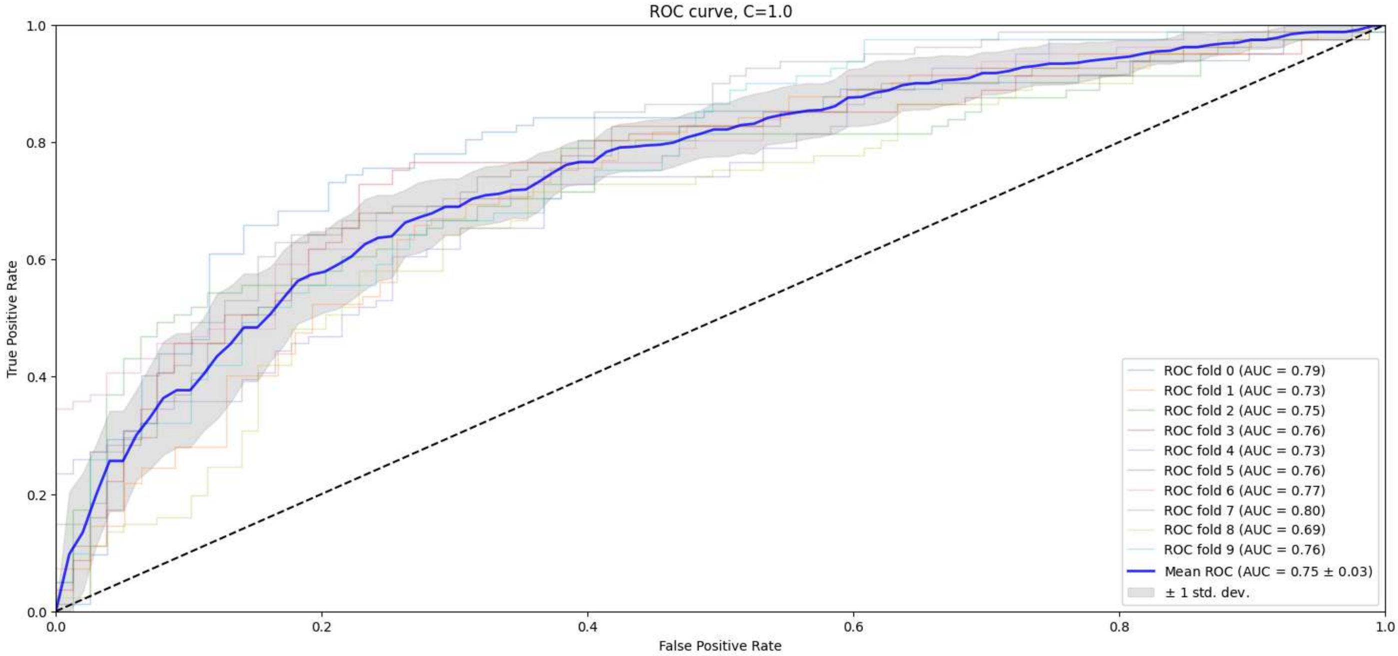

True Pos
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC fold 0 (AUC = 0.81)
ROC fold 1 (AUC = 0.75)
ROC fold 2 (AUC = 0.72)
ROC fold 3 (AUC = 0.80)
ROC fold 4 (AUC = 0.75)
ROC fold 5 (AUC = 0.85)
ROC fold 6 (AUC = 0.78)
ROC fold 7 (AUC = 0.82)
ROC fold 8 (AUC = 0.75)
ROC fold 9 (AUC = 0.86)
Mean ROC (AUC = 0.79 ± 0.05)
± 1 std. dev.

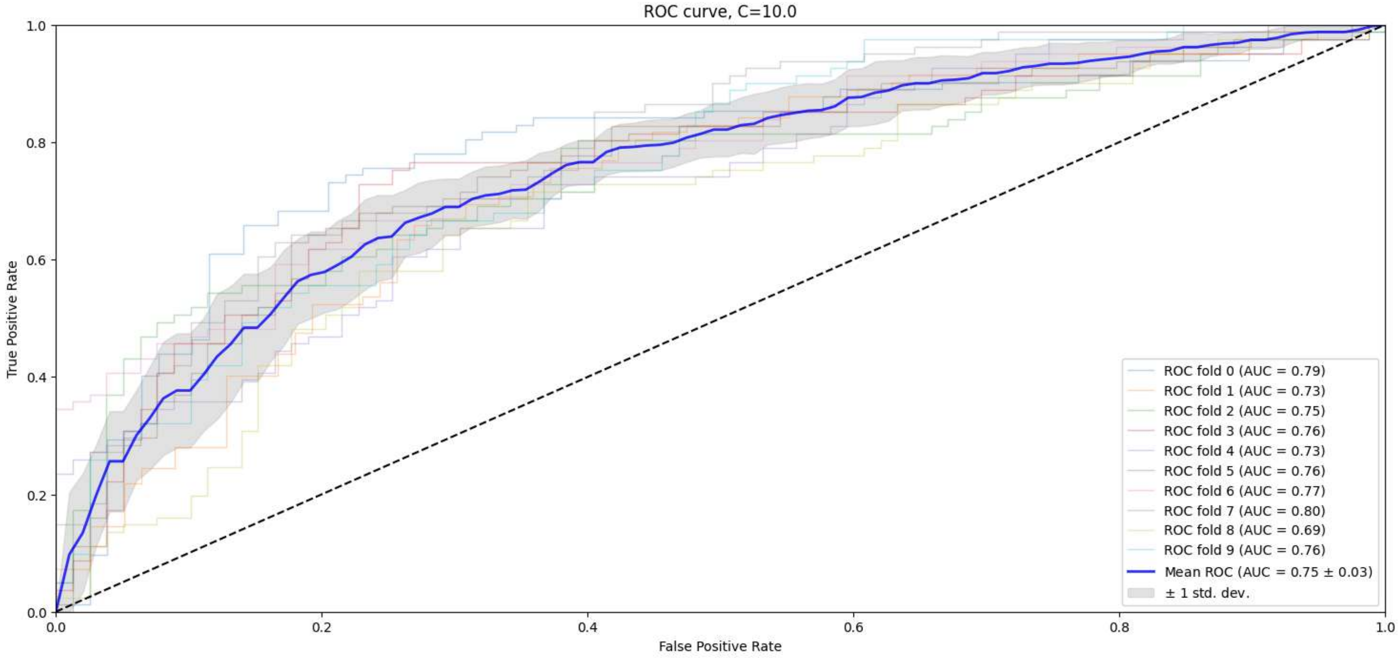

ROC curve, C=0.001
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC fold 0 (AUC = 0.87)
ROC fold 1 (AUC = 0.83)
ROC fold 2 (AUC = 0.80)
ROC fold 3 (AUC = 0.90)
ROC fold 4 (AUC = 0.84)
ROC fold 5 (AUC = 0.90)
ROC fold 6 (AUC = 0.85)
ROC fold 7 (AUC = 0.88)
ROC fold 8 (AUC = 0.79)
ROC fold 9 (AUC = 0.88)
Mean ROC (AUC = 0.85 ± 0.04)
± 1 std. dev.

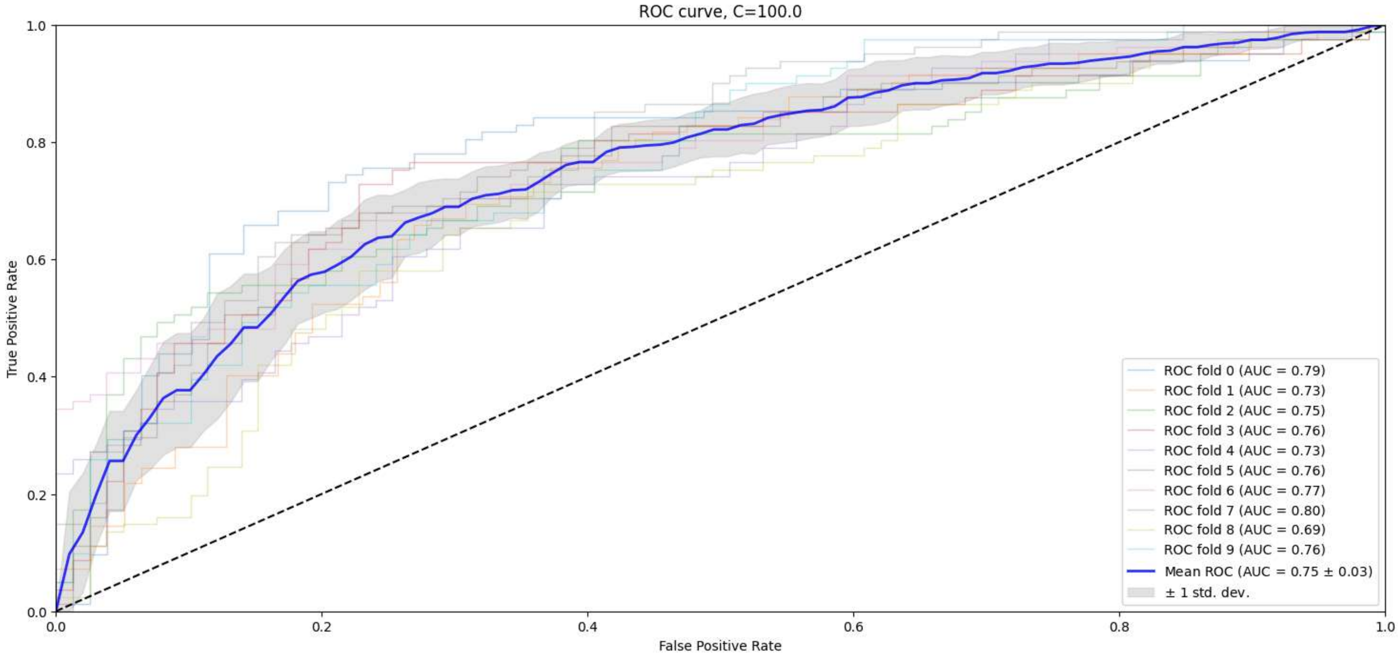

ROC curve, C=0.01
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC fold 0 (AUC = 0.84)
ROC fold 1 (AUC = 0.83)
ROC fold 2 (AUC = 0.84)
ROC fold 3 (AUC = 0.88)
ROC fold 4 (AUC = 0.85)
ROC fold 5 (AUC = 0.85)
ROC fold 6 (AUC = 0.85)
ROC fold 7 (AUC = 0.89)
ROC fold 8 (AUC = 0.79)
ROC fold 9 (AUC = 0.84)
Mean ROC (AUC = 0.85 ± 0.03)
± 1 std. dev.


ROC curve, C=0.1
True Positive Rate
1.0
0.8
0.6
0.4
ROC fold 0 (AUC = 0.80)
ROC fold 1 (AUC = 0.76)
ROC fold 2 (AUC = 0.78)
ROC fold 3 (AUC = 0.80)
ROC fold 4 (AUC = 0.80)
ROC fold 5 (AUC = 0.80)
ROC fold 6 (AUC = 0.81)

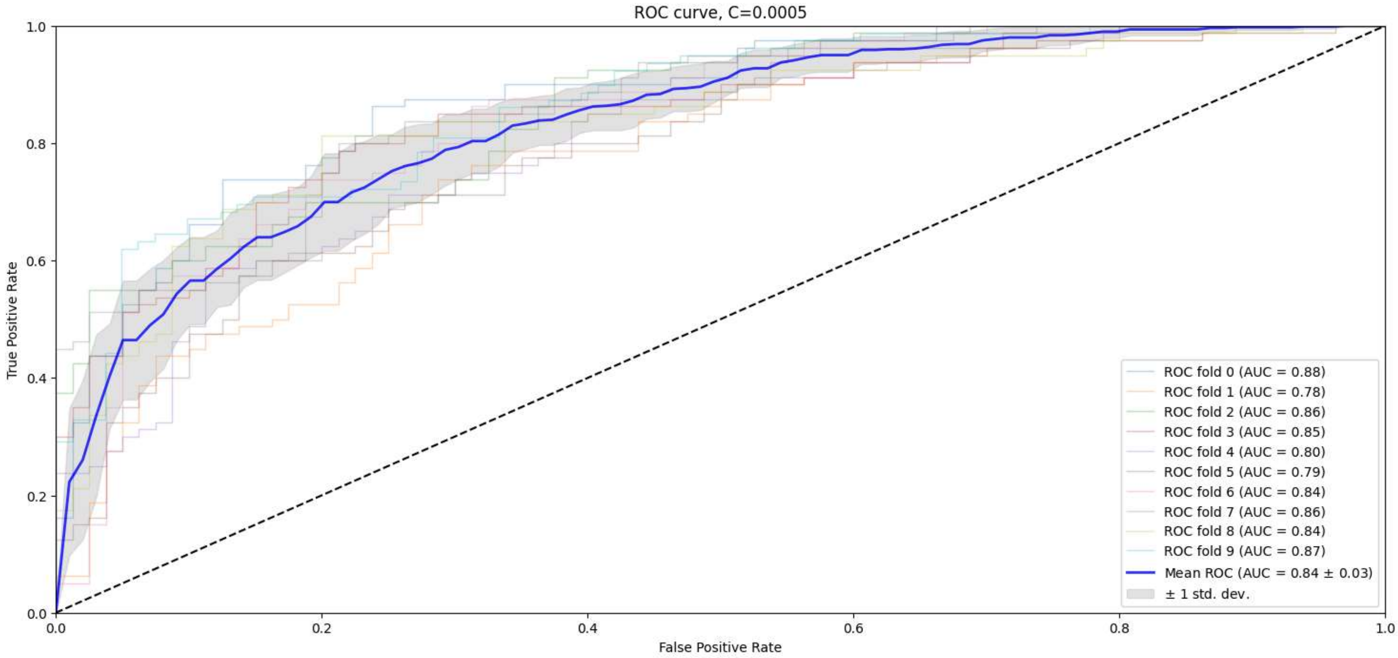

0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC fold 7 (AUC = 0.84)
ROC fold 8 (AUC = 0.72)
ROC fold 9 (AUC = 0.78)
Mean ROC (AUC = 0.79 ± 0.03)
± 1 std. dev.

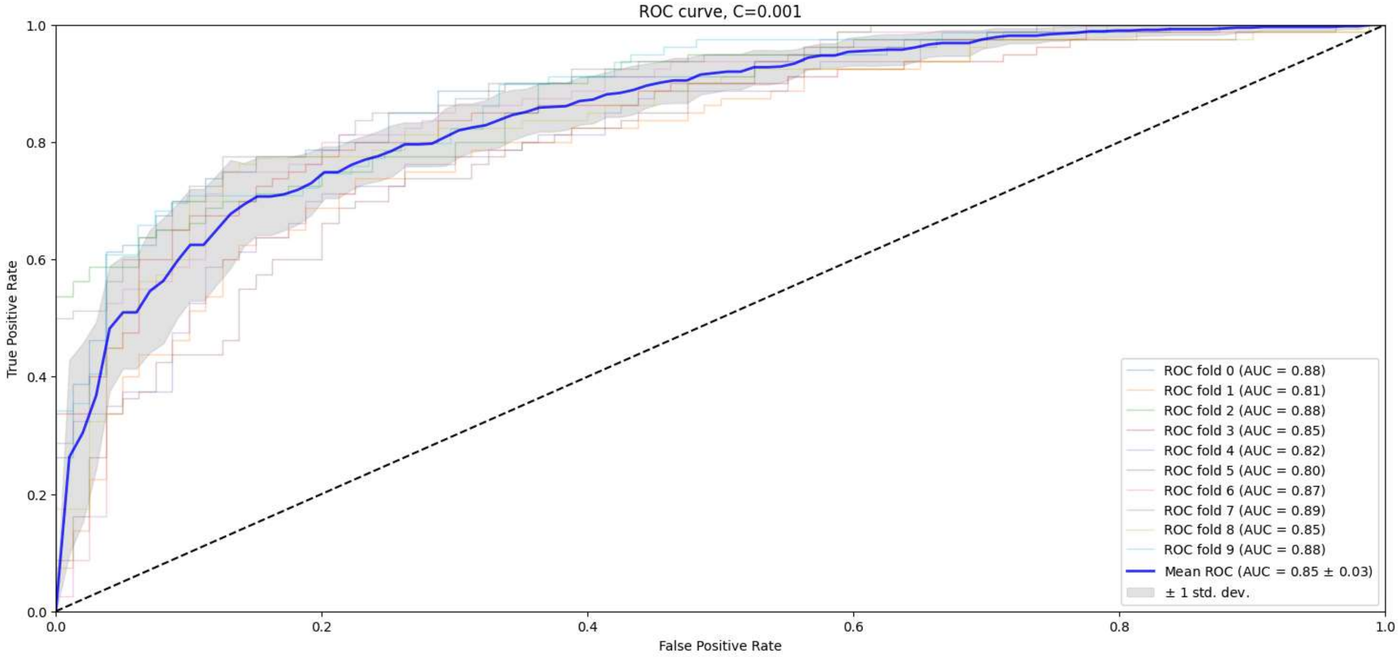

ROC curve, C=1.0
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.79)
ROC fold 1 (AUC = 0.73)
ROC fold 2 (AUC = 0.75)
ROC fold 3 (AUC = 0.76)
ROC fold 4 (AUC = 0.73)
ROC fold 5 (AUC = 0.76)
ROC fold 6 (AUC = 0.77)
ROC fold 7 (AUC = 0.80)
ROC fold 8 (AUC = 0.69)
ROC fold 9 (AUC = 0.76)
Mean ROC (AUC = 0.75 ± 0.03)
± 1 std. dev.

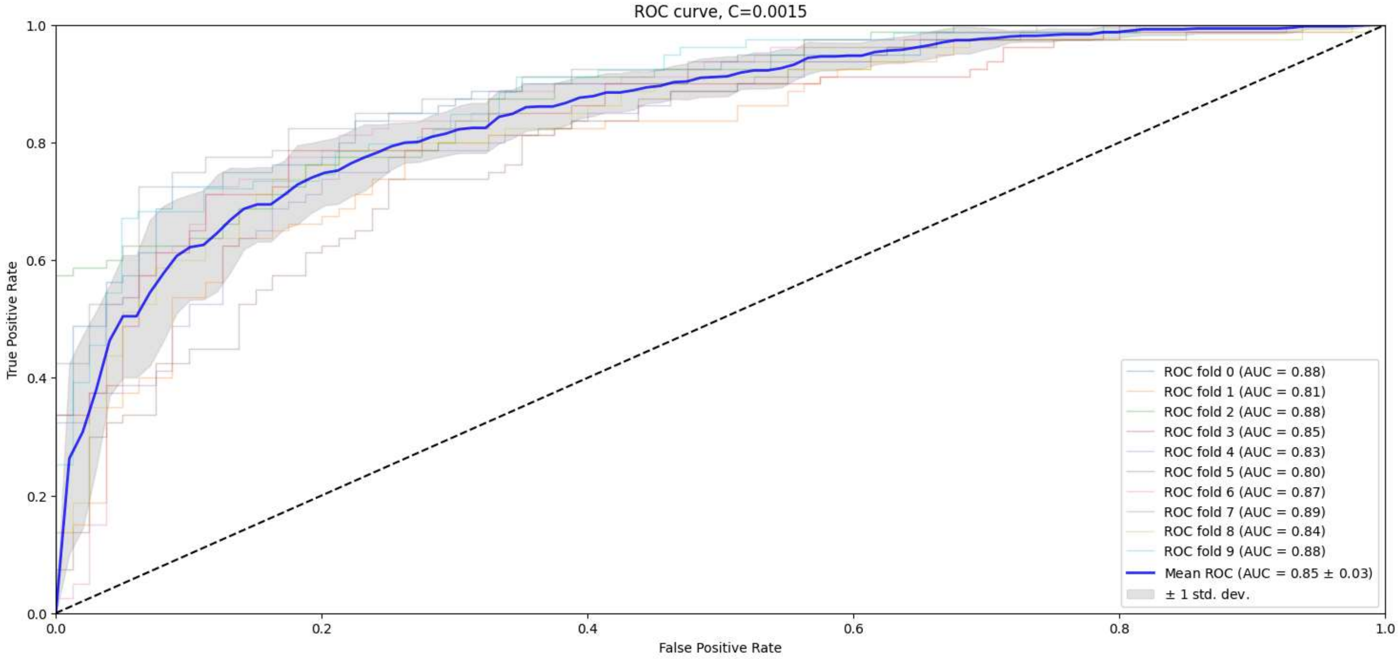

ROC curve, C=10.0
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.79)
ROC fold 1 (AUC = 0.73)
ROC fold 2 (AUC = 0.75)
ROC fold 3 (AUC = 0.76)
ROC fold 4 (AUC = 0.73)
ROC fold 5 (AUC = 0.76)
ROC fold 6 (AUC = 0.77)
ROC fold 7 (AUC = 0.80)
ROC fold 8 (AUC = 0.69)
ROC fold 9 (AUC = 0.76)
Mean ROC (AUC = 0.75 ± 0.03)
± 1 std. dev.

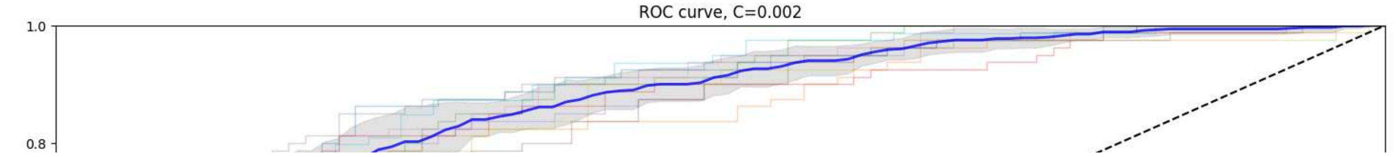

ROC curve, C=100.0
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.79)
ROC fold 1 (AUC = 0.73)
ROC fold 2 (AUC = 0.75)
ROC fold 3 (AUC = 0.76)
ROC fold 4 (AUC = 0.73)
ROC fold 5 (AUC = 0.76)
ROC fold 6 (AUC = 0.77)
ROC fold 7 (AUC = 0.80)
ROC fold 8 (AUC = 0.69)
ROC fold 9 (AUC = 0.76)
Mean ROC (AUC = 0.75 ± 0.03)
± 1 std. dev.

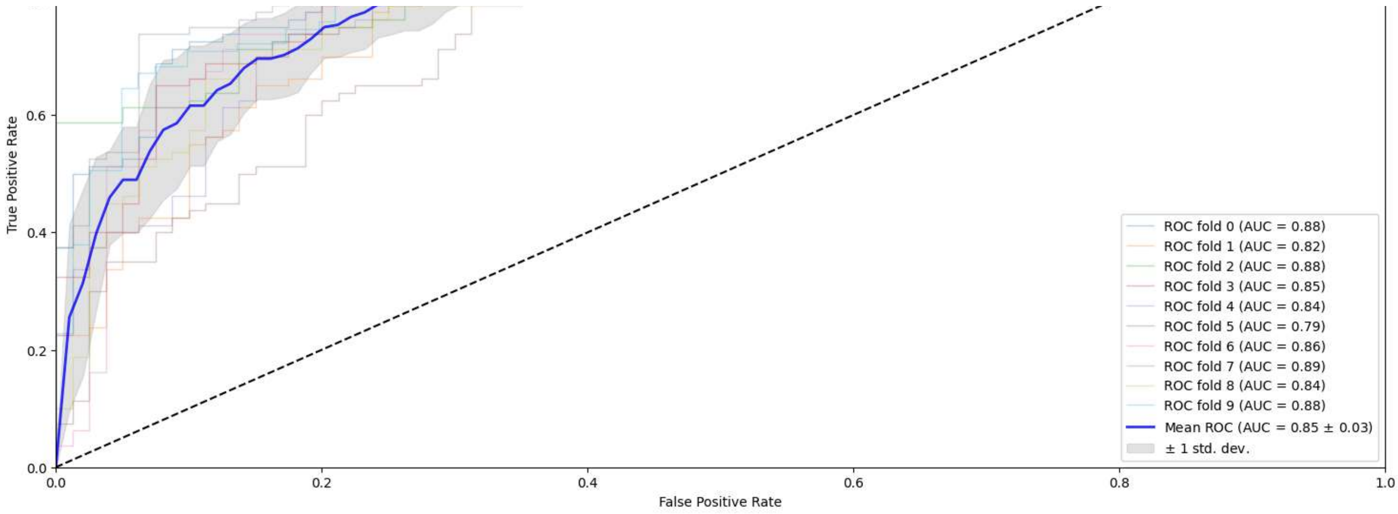
```
n_splits = 10
for c in np.linspace(1e-3 * 0.5,1e-2,20):
    classifier = svm.SVC(C=c, kernel="linear")
    plot_roc_with_cv(classifier, X_train, Y_train, n_splits, f"ROC curve, C={c}")
```


ROC curve, C=0.0005
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.88)
ROC fold 1 (AUC = 0.78)
ROC fold 2 (AUC = 0.86)
ROC fold 3 (AUC = 0.85)
ROC fold 4 (AUC = 0.80)
ROC fold 5 (AUC = 0.79)
ROC fold 6 (AUC = 0.84)
ROC fold 7 (AUC = 0.86)
ROC fold 8 (AUC = 0.84)
ROC fold 9 (AUC = 0.87)
Mean ROC (AUC = 0.84 ± 0.03)
± 1 std. dev.

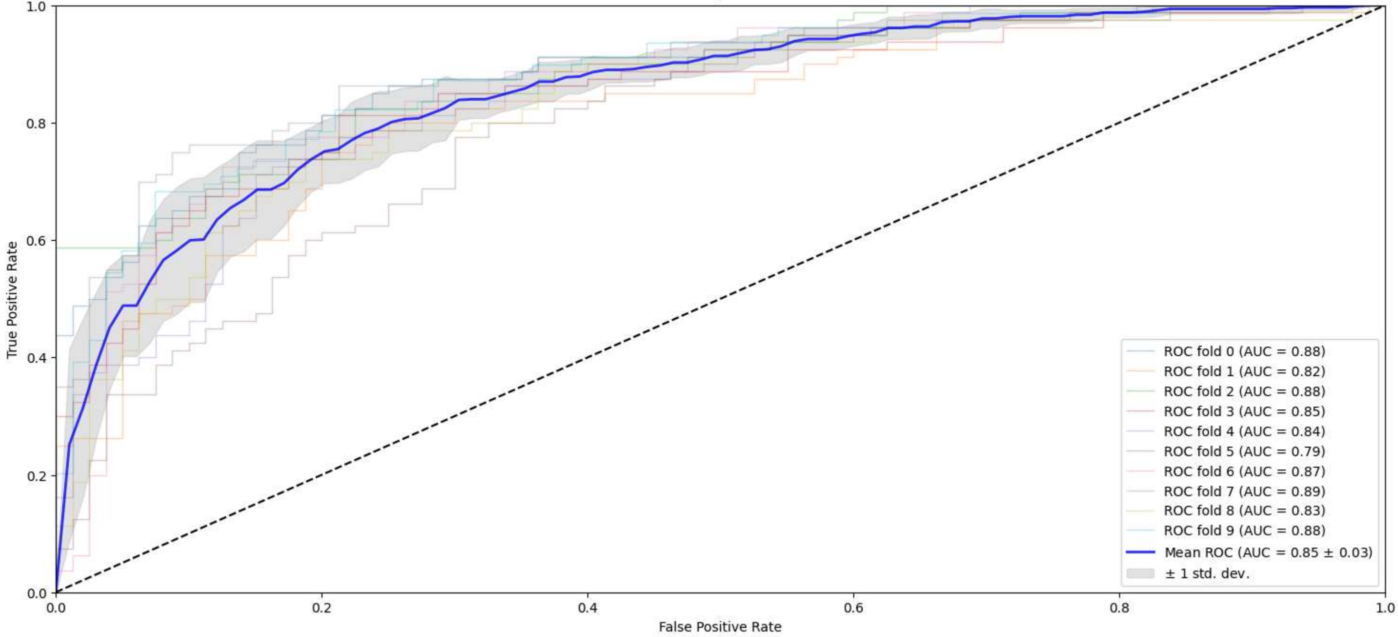

ROC curve, C=0.001
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.88)
ROC fold 1 (AUC = 0.81)
ROC fold 2 (AUC = 0.88)
ROC fold 3 (AUC = 0.85)
ROC fold 4 (AUC = 0.82)
ROC fold 5 (AUC = 0.80)
ROC fold 6 (AUC = 0.87)
ROC fold 7 (AUC = 0.89)
ROC fold 8 (AUC = 0.85)
ROC fold 9 (AUC = 0.88)
Mean ROC (AUC = 0.85 ± 0.03)
± 1 std. dev.


ROC curve, C=0.0015
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.88)
ROC fold 1 (AUC = 0.81)
ROC fold 2 (AUC = 0.88)
ROC fold 3 (AUC = 0.85)
ROC fold 4 (AUC = 0.83)
ROC fold 5 (AUC = 0.80)
ROC fold 6 (AUC = 0.87)
ROC fold 7 (AUC = 0.89)
ROC fold 8 (AUC = 0.84)
ROC fold 9 (AUC = 0.88)
Mean ROC (AUC = 0.85 ± 0.03)
± 1 std. dev.

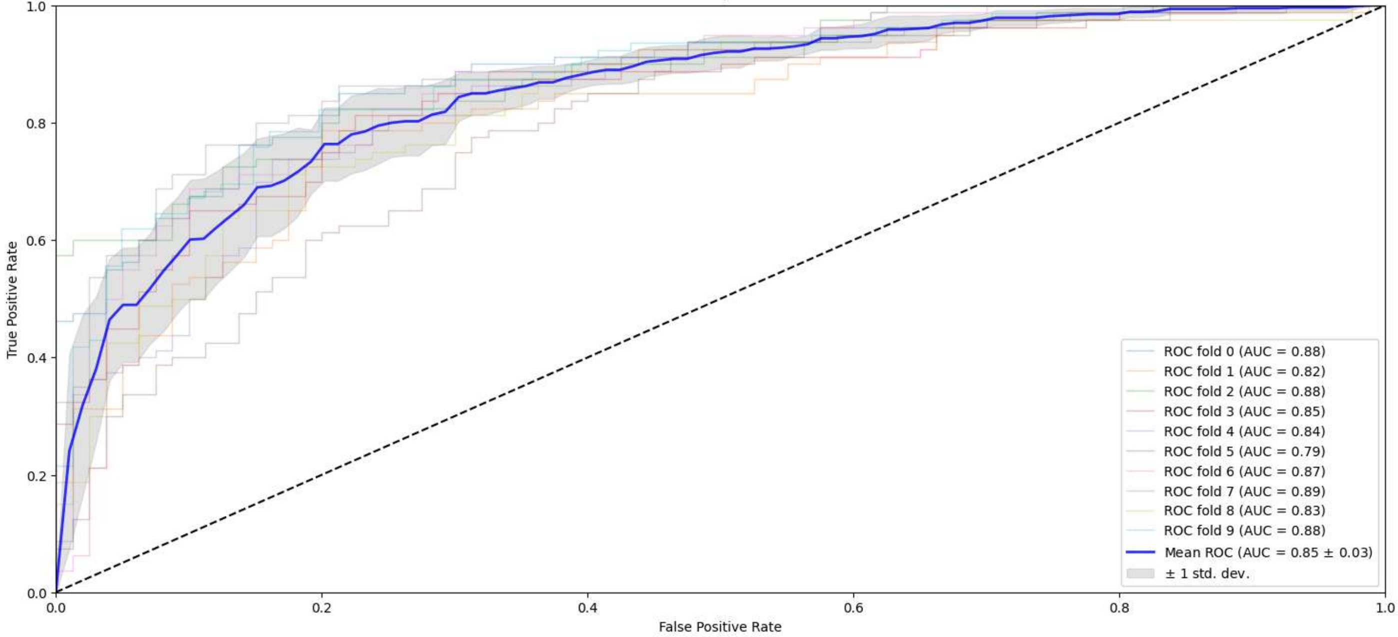

ROC curve, C=0.002
1.0
0.8



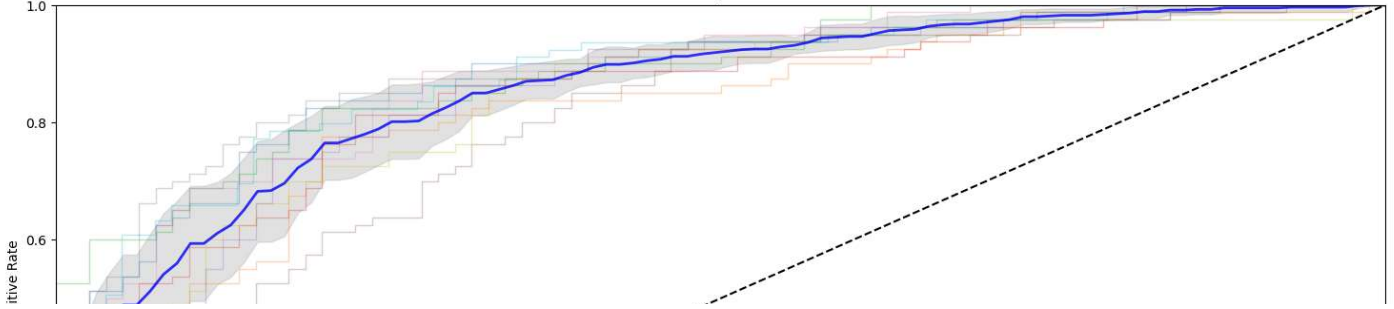
ROC curve, C=0.0025

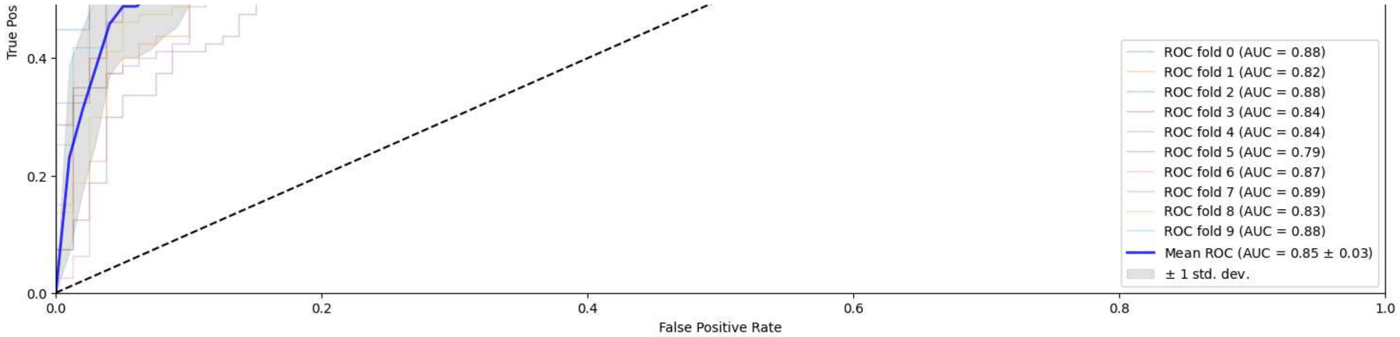


ROC curve, C=0.003

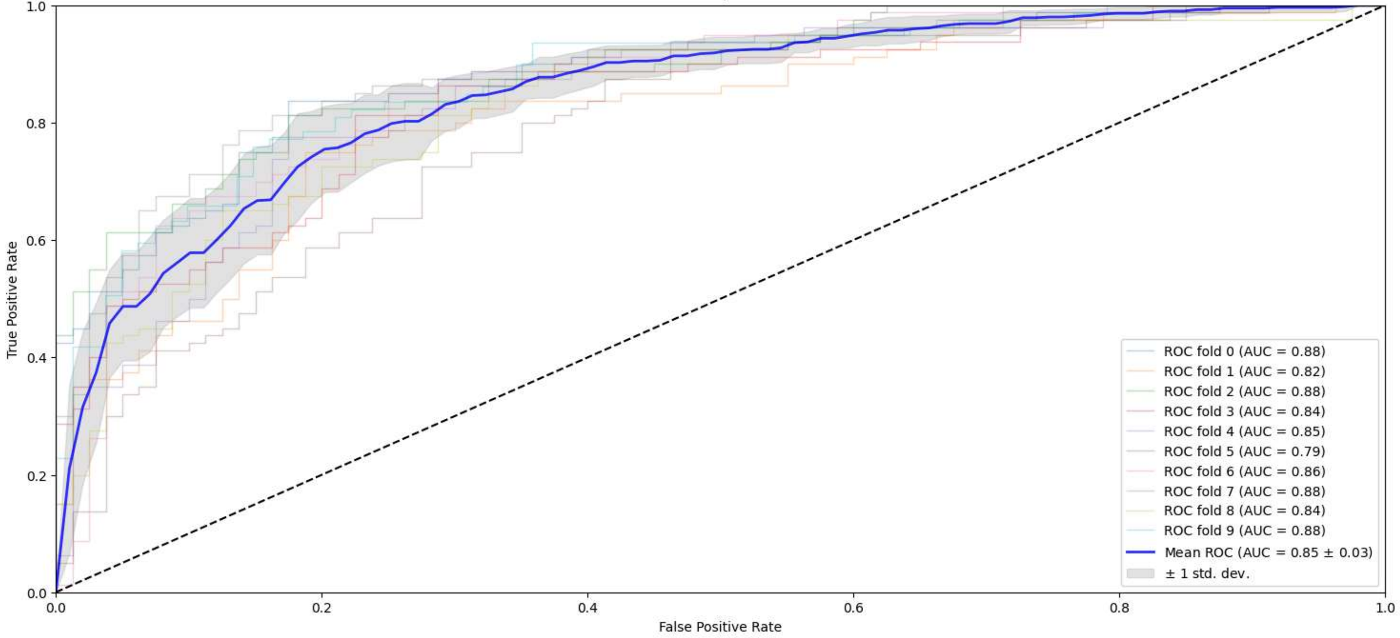


ROC curve, C=0.0035

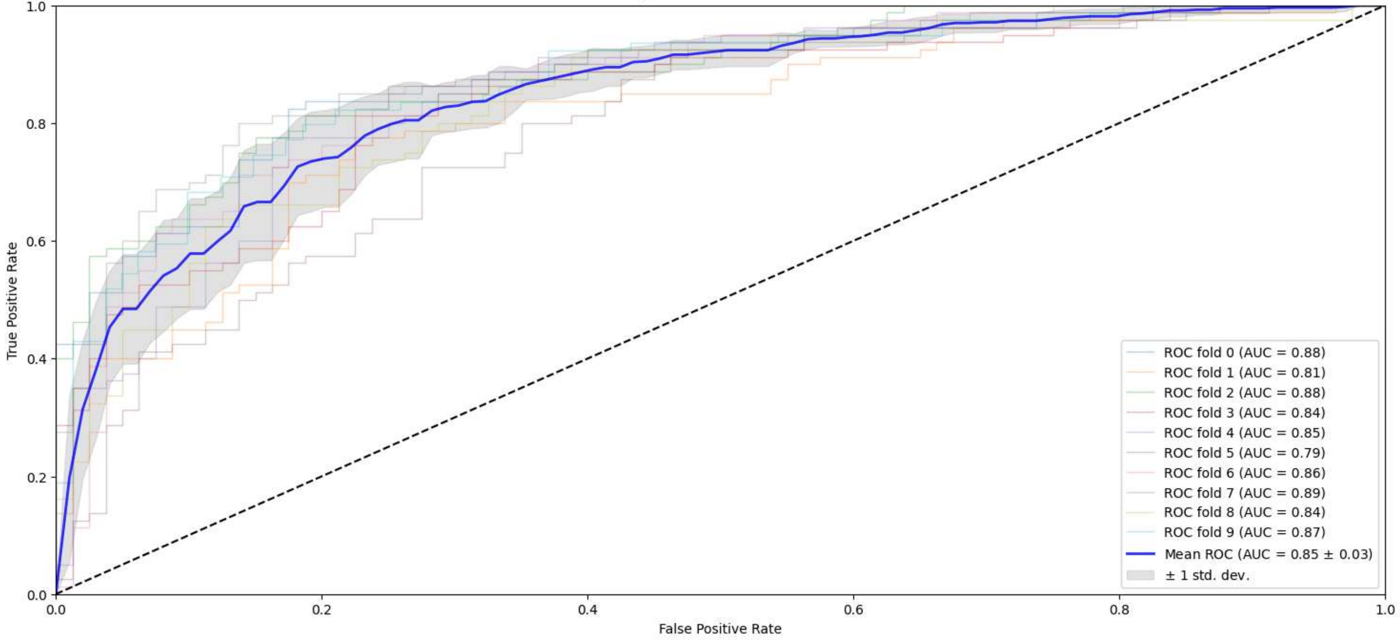




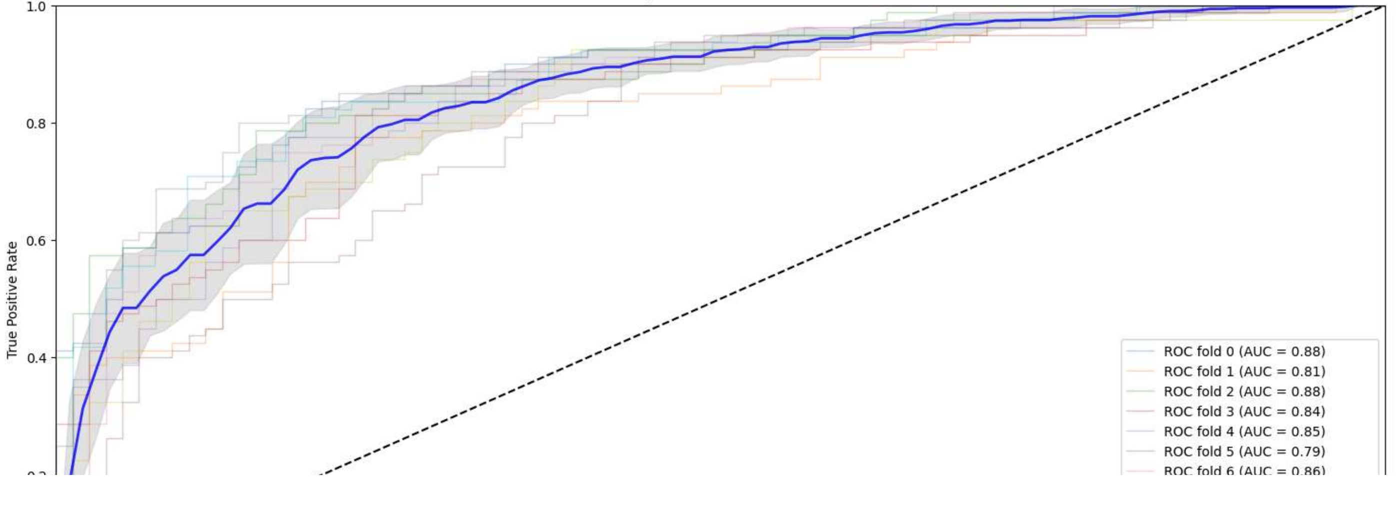
ROC curve, C=0.004



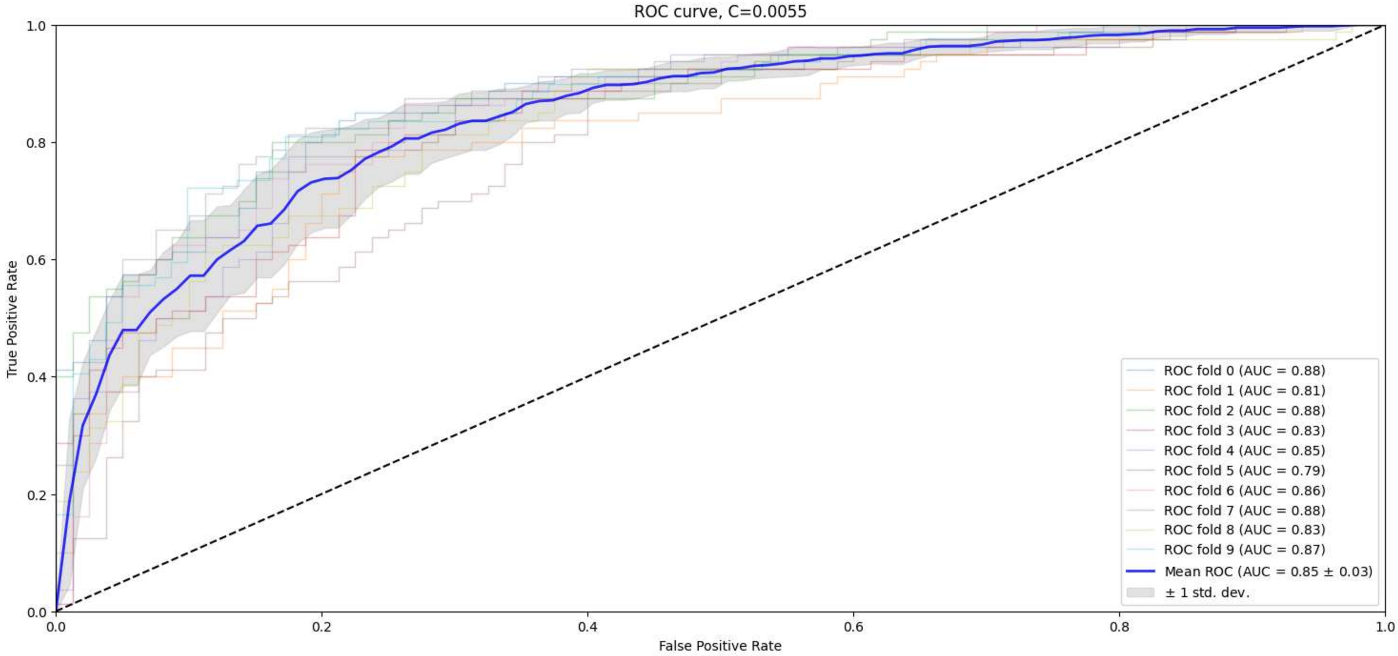
ROC curve, C=0.0045000000000000005

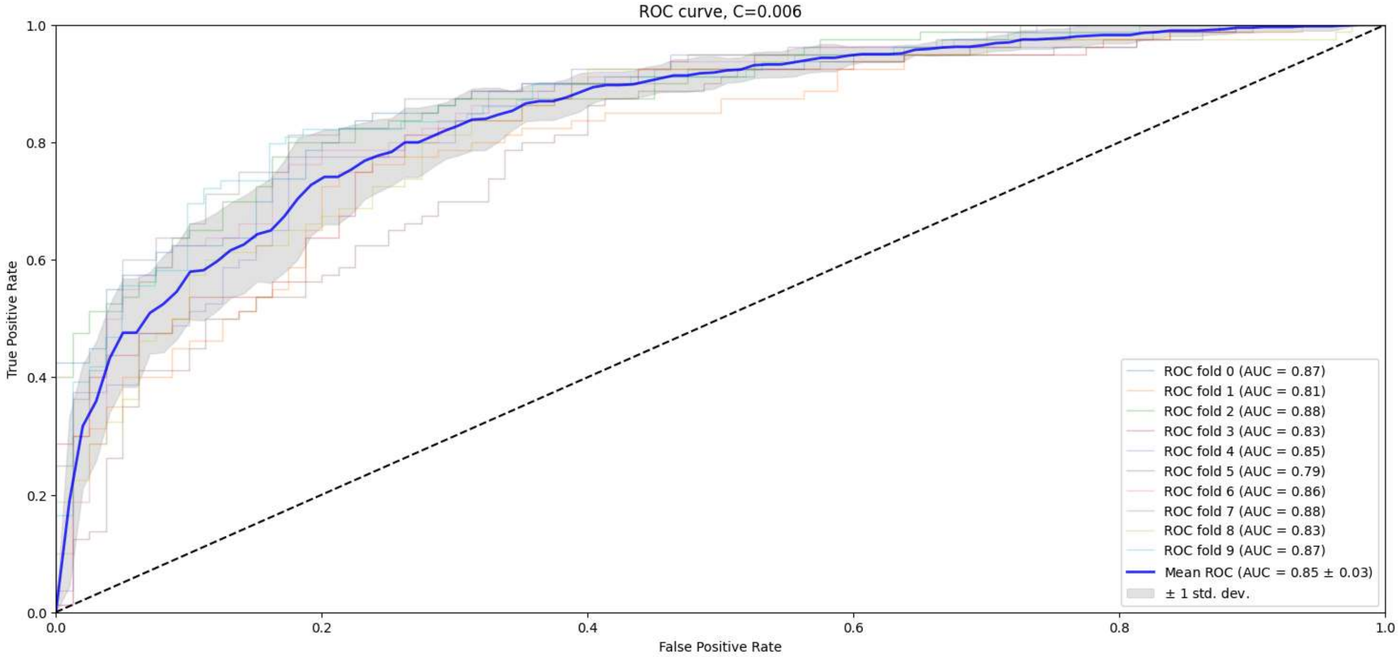


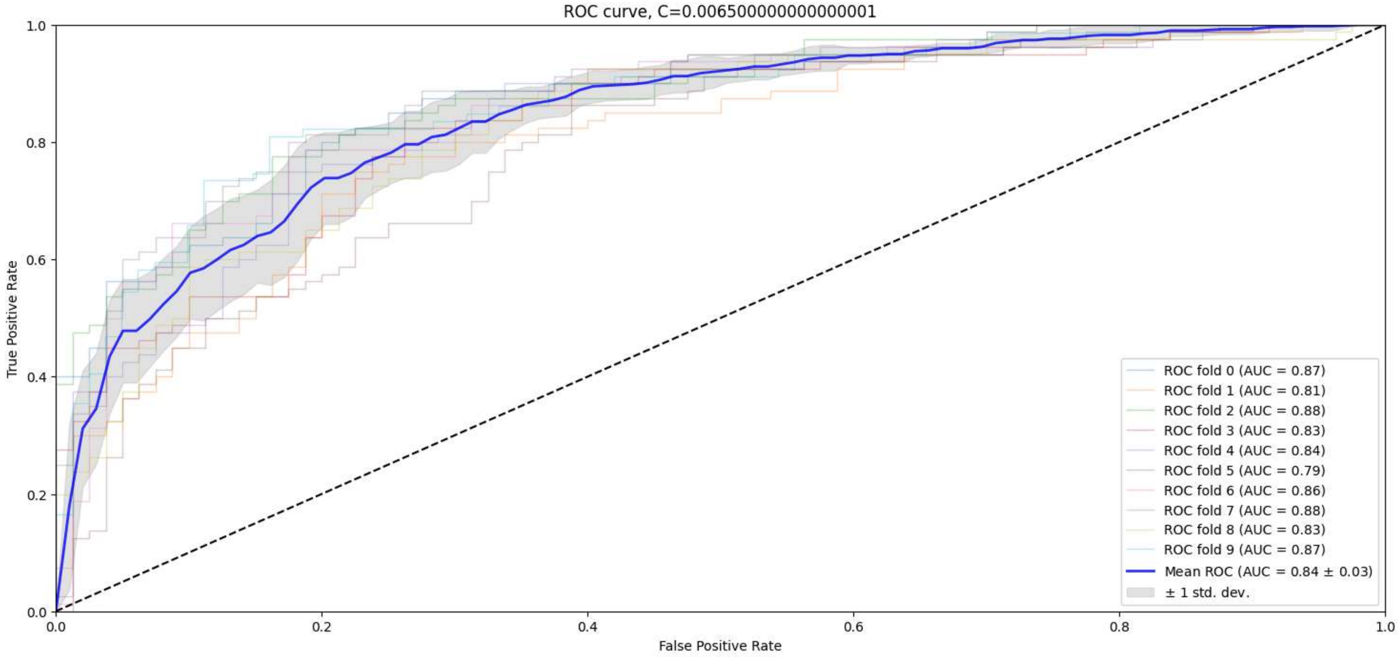
ROC curve, C=0.005000000000000001

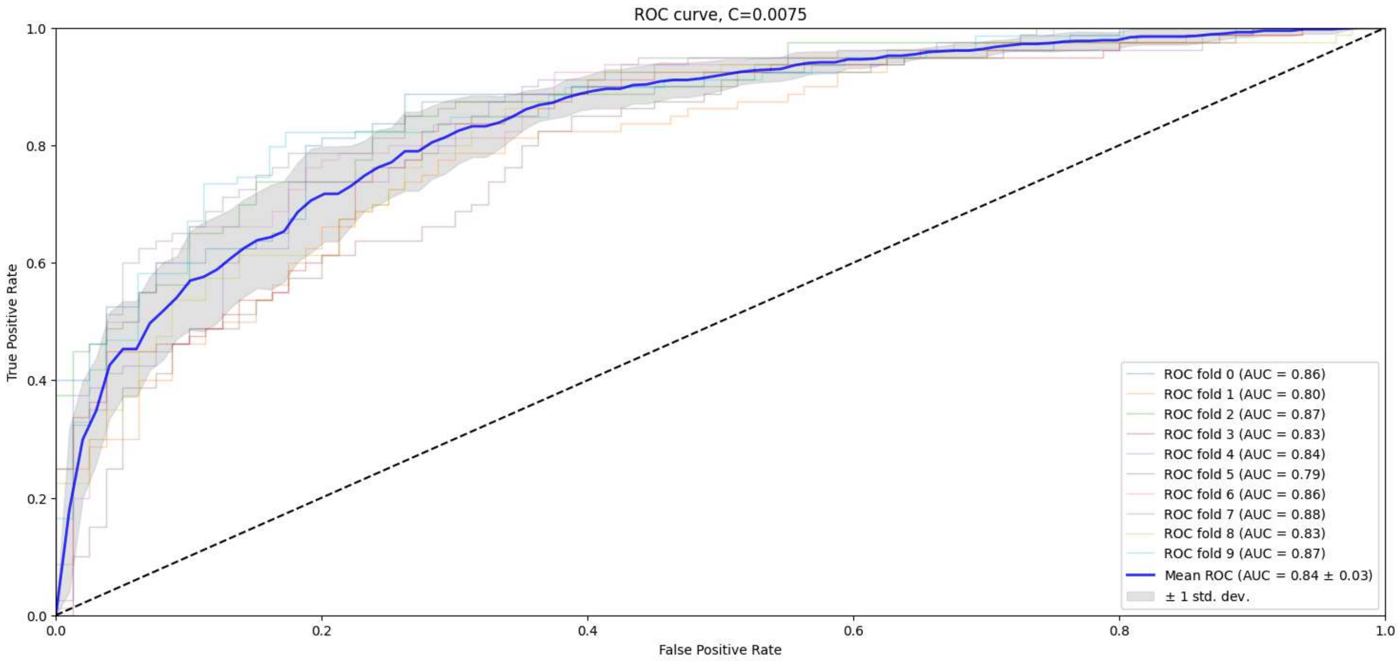


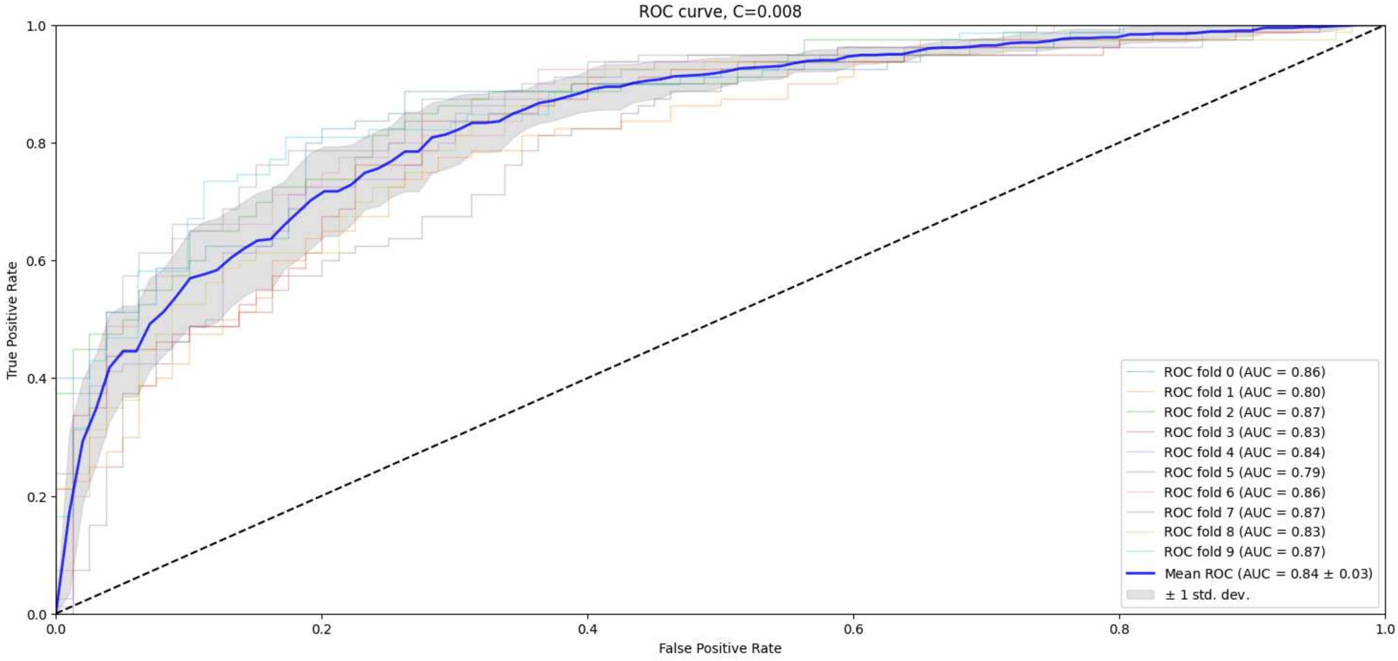
0.2
0.0
ROC fold 7 (AUC = 0.88)
ROC fold 8 (AUC = 0.83)
ROC fold 9 (AUC = 0.87)
Mean ROC (AUC = 0.85 ± 0.03)
± 1 std. dev.
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

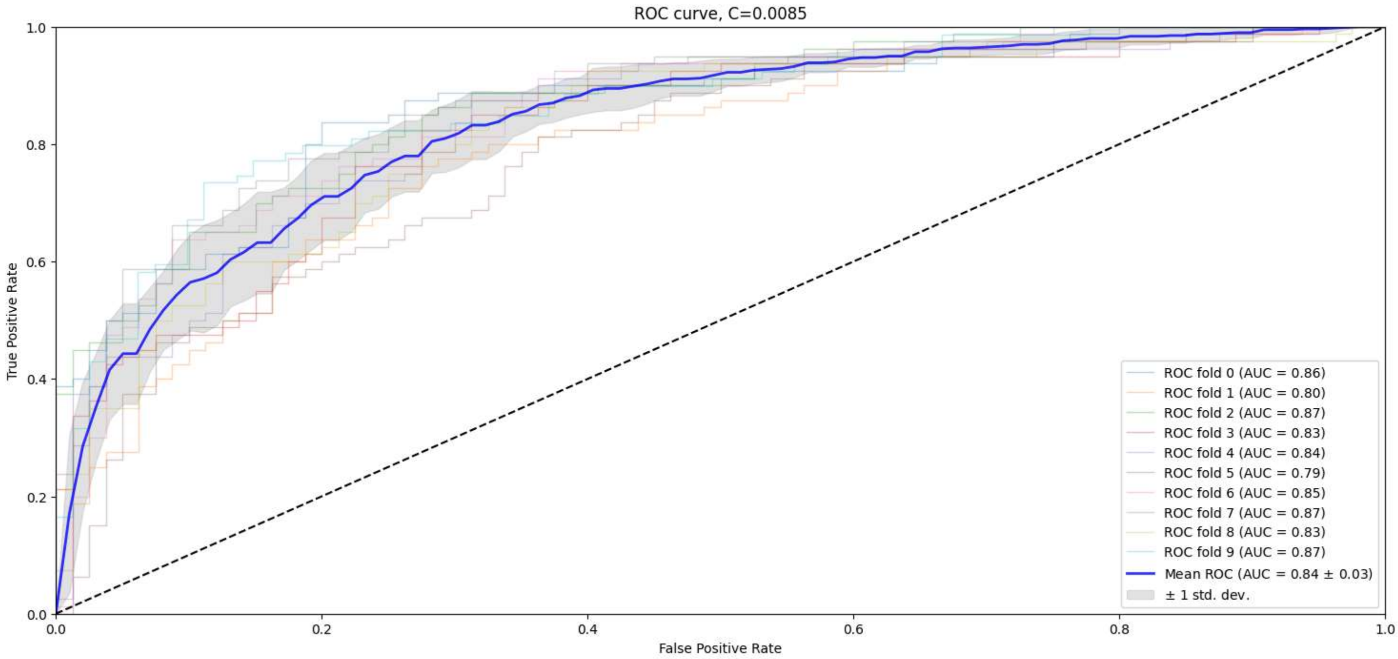

ROC curve, C=0.0055

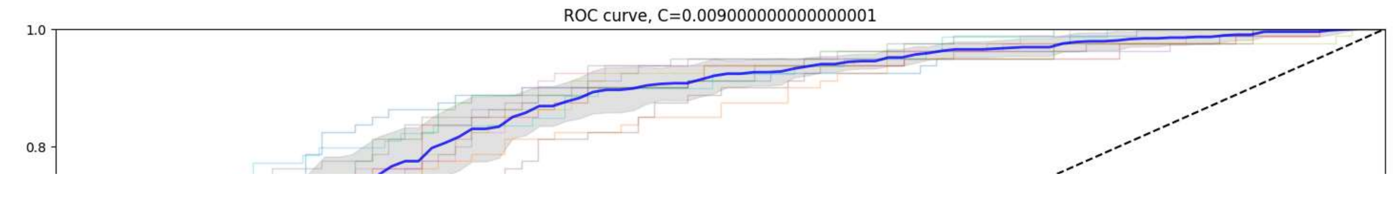

ROC curve, C=0.006

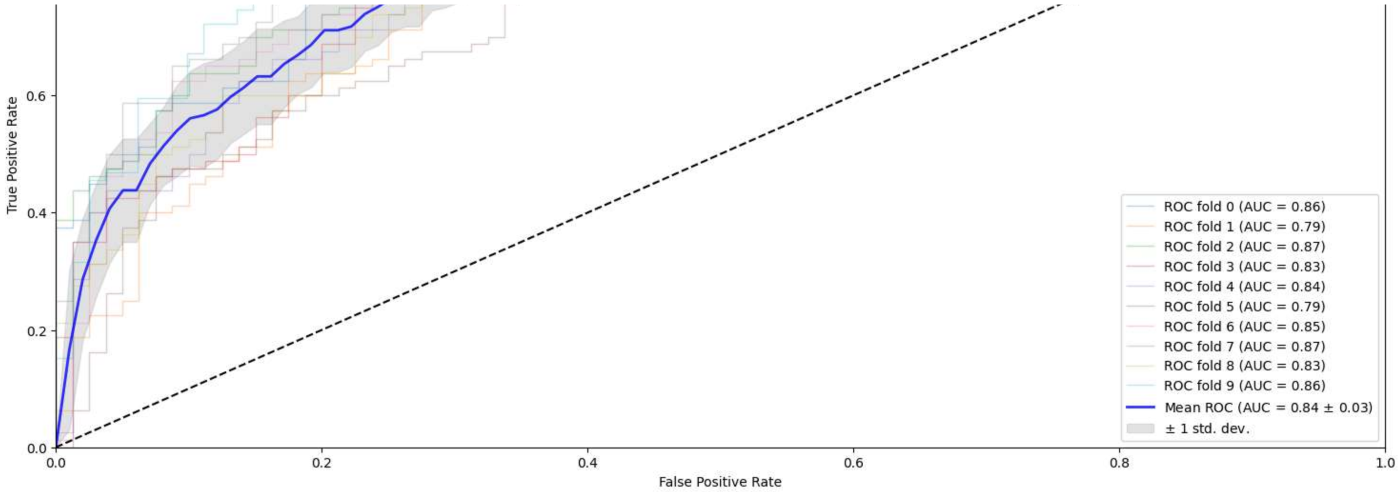

ROC curve, C=0.006500000000000001

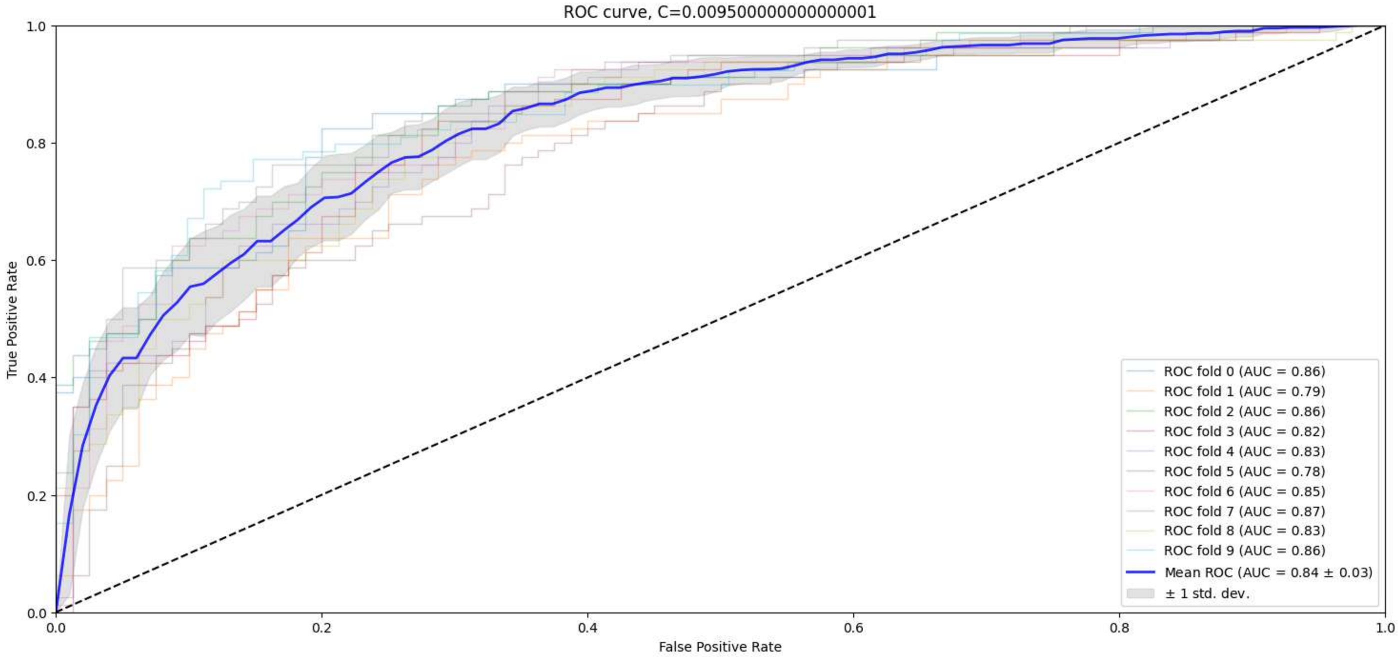

ROC curve, C=0.0075
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.86)
ROC fold 1 (AUC = 0.80)
ROC fold 2 (AUC = 0.87)
ROC fold 3 (AUC = 0.83)
ROC fold 4 (AUC = 0.84)
ROC fold 5 (AUC = 0.79)
ROC fold 6 (AUC = 0.86)
ROC fold 7 (AUC = 0.88)
ROC fold 8 (AUC = 0.83)
ROC fold 9 (AUC = 0.87)
Mean ROC (AUC = 0.84 ± 0.03)
± 1 std. dev.

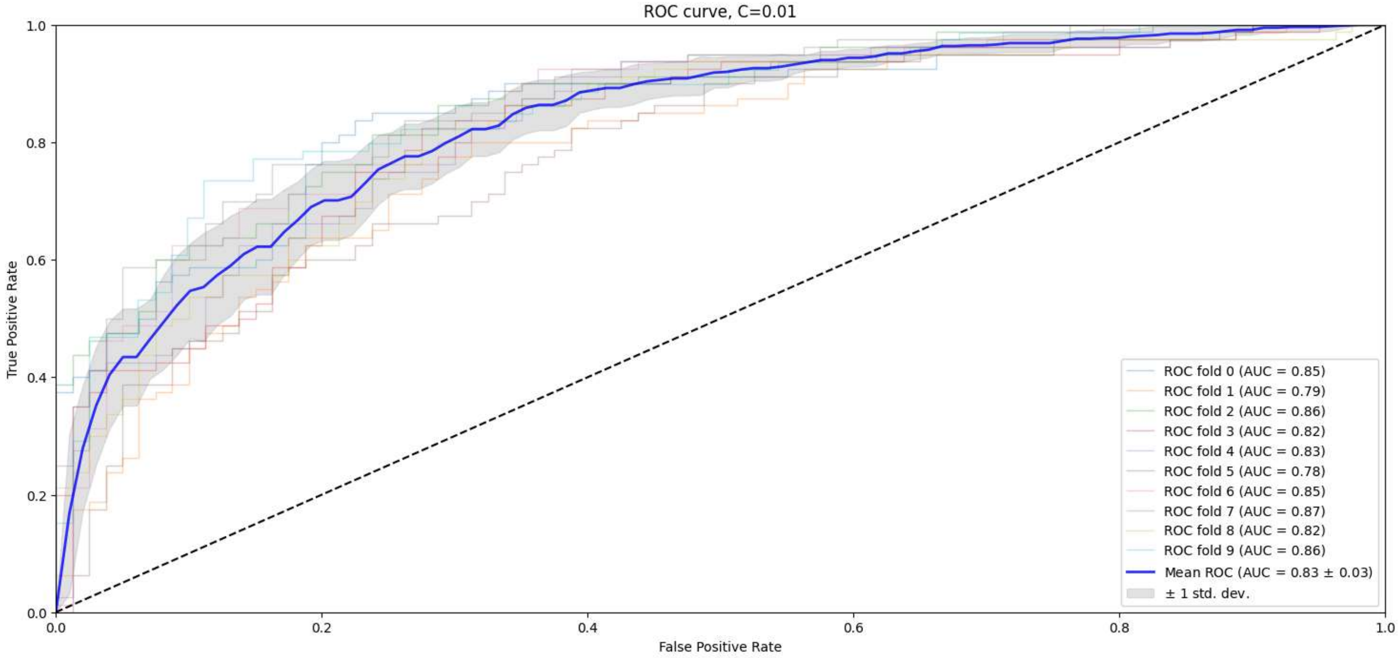

ROC curve, C=0.008
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.86)
ROC fold 1 (AUC = 0.80)
ROC fold 2 (AUC = 0.87)
ROC fold 3 (AUC = 0.83)
ROC fold 4 (AUC = 0.84)
ROC fold 5 (AUC = 0.79)
ROC fold 6 (AUC = 0.86)
ROC fold 7 (AUC = 0.87)
ROC fold 8 (AUC = 0.83)
ROC fold 9 (AUC = 0.87)
Mean ROC (AUC = 0.84 ± 0.03)
± 1 std. dev.


ROC curve, C=0.0085
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.86)
ROC fold 1 (AUC = 0.80)
ROC fold 2 (AUC = 0.87)
ROC fold 3 (AUC = 0.83)
ROC fold 4 (AUC = 0.84)
ROC fold 5 (AUC = 0.79)
ROC fold 6 (AUC = 0.85)
ROC fold 7 (AUC = 0.87)
ROC fold 8 (AUC = 0.83)
ROC fold 9 (AUC = 0.87)
Mean ROC (AUC = 0.84 ± 0.03)
± 1 std. dev.


ROC curve, C=0.009000000000000001
1.0
0.8




ROC curve, C=0.009500000000000001


ROC curve, C=0.01