

UROP

MULTI-CLASS ESG CLASSIFICATION

SHE, FONG WING (CINNIE)

Research Topic

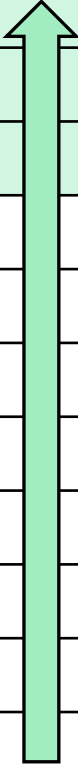
Evaluate the performance of FinBERT on multi-class ESG classification task based on comparison with other Deep Learning and machine learning algorithms

Methodology

- Dataset:
 - 4,000 ESG sentences, 500 for each label
- Dataset Split (randomly split using sklearn's train_test_split method):
 - 81% Training (3,240 sentences)
 - 9% Validation (360 sentences)
 - 10% Test (400 sentences)
- Determining the prediction label:
 - Assign each sentence to label with the highest likelihood predicted by the algorithms
- Fine-Tuning:
 - Based on validation accuracy
- Evaluation Metrics:
 - Accuracy, Macro-F1 score, Confusion matrix
 - Same test dataset is used for evaluating final test performance on all models

Results

Model	Test Accuracy	Test Macro-f1 Score
FinBERT	0.8978	0.8981
BERT	0.8978	0.8975
MLP	0.8800	0.8803
Linear SVM	0.8489	0.8483
Logistic Regression	0.8400	0.8409
CNN	0.8378	0.8396
Bi-directional LSTM	0.8311	0.8292
Naïve Bayes	0.8289	0.8299
Random Forest	0.8133	0.8150
LSTM	0.7644	0.7672
GRU	0.7489	0.7543



Observations

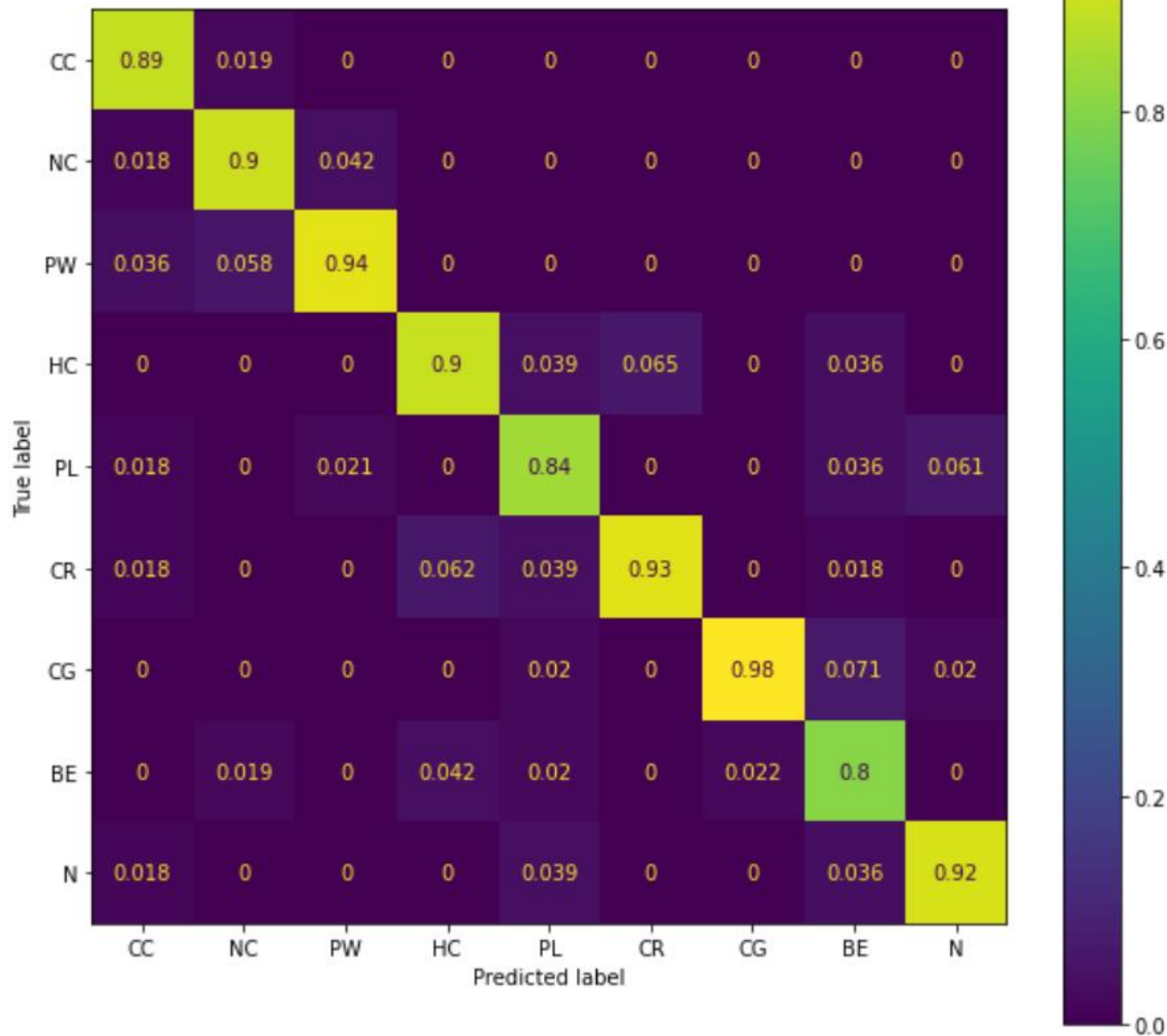
- FinBERT outperform all other tested models based on accuracy and macro-f1 score
 - However, performance of FinBERT and BERT is super close, with same accuracy and a smaller than 0.001 difference in macro-f1 score
 - Top 3 models (FinBERT, BERT, and MLP) outperform other models with at least 0.0311 accuracy difference
 - Not all deep learning models (MLP, CNN, LSTM, GRU, Bi-LSTM) outperform machine learning algorithms (NB, SVM, LR, RF)
- Comparing FinBERT and BERT:
 - Overall:
 - FinBERT gives a less balanced distribution of precision on the predicting the 9 different classes, while BERT have similar precision performance on all classes (e.g., FinBERT is particular weak in predicting PL and BE sentences)
 - Good at:
 - FinBERT performs better in classifying 'environmental' ('CC', 'NC', 'PW') related, and 'CR', 'CG' related sentences
 - BERT performs better in classifying 'HC', 'PL', 'BE', and 'N' related sentences
 - Poor at:
 - FinBERT's most common wrong predictions are predicting 'CG' as 'BE' (7.1%), 'HC' as 'CR' (6.5%), or 'CR' as 'HC' (6.2%)
 - BERT's most common wrong predictions are predicting 'HC' as 'CR' (8%), or 'NC' as 'PW' (7.7%)

FinBERT

Test Accuracy: 0.8978

Test Macro-F1 Score: 0.8981

	precision	recall	f1-score	support
CC	0.89	0.98	0.93	50
NC	0.90	0.94	0.92	50
PW	0.94	0.90	0.92	50
HC	0.90	0.86	0.88	50
PL	0.84	0.86	0.85	50
CR	0.93	0.86	0.90	50
CG	0.98	0.88	0.93	50
BE	0.80	0.90	0.85	50
N	0.92	0.90	0.91	50
accuracy			0.90	450
macro avg	0.90	0.90	0.90	450
weighted avg	0.90	0.90	0.90	450

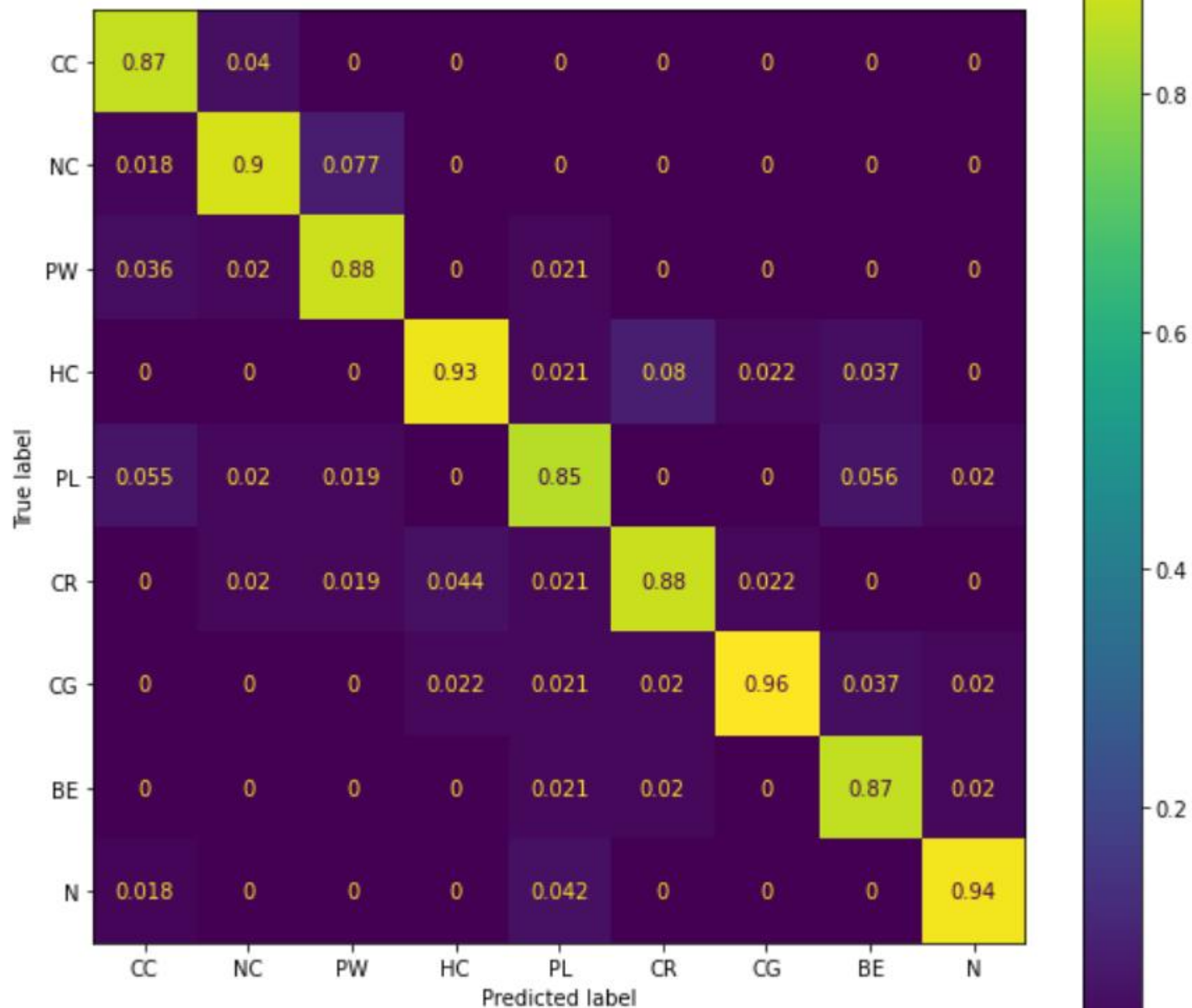


BERT

Test Accuracy: 0.8978

Test Macro-F1 Score: 0.8975

	precision	recall	f1-score	support
CC	0.87	0.96	0.91	50
NC	0.90	0.90	0.90	50
PW	0.88	0.92	0.90	50
HC	0.93	0.84	0.88	50
PL	0.85	0.82	0.84	50
CR	0.88	0.88	0.88	50
CG	0.96	0.88	0.92	50
BE	0.87	0.94	0.90	50
N	0.94	0.94	0.94	50
accuracy			0.90	450
macro avg	0.90	0.90	0.90	450
weighted avg	0.90	0.90	0.90	450

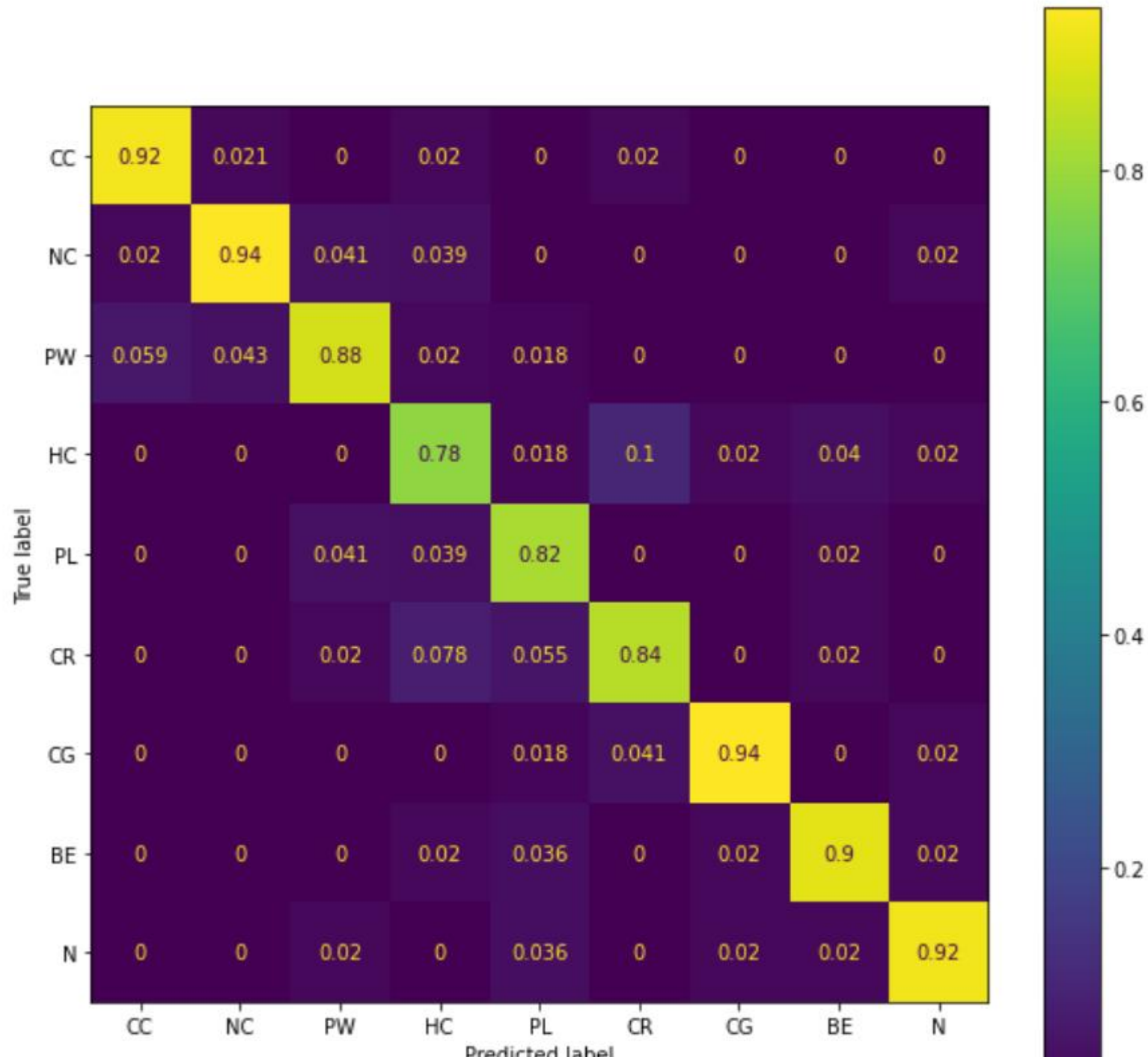


MLP

Test Accuracy: 0.8800

Test Macro-F1 Score: 0.8803

	precision	recall	f1-score	support
CC	0.92	0.94	0.93	50
NC	0.94	0.88	0.91	50
PW	0.88	0.86	0.87	50
HC	0.78	0.80	0.79	50
PL	0.82	0.90	0.86	50
CR	0.84	0.82	0.83	50
CG	0.94	0.92	0.93	50
BE	0.90	0.90	0.90	50
N	0.92	0.90	0.91	50
accuracy			0.88	450
macro avg	0.88	0.88	0.88	450
weighted avg	0.88	0.88	0.88	450

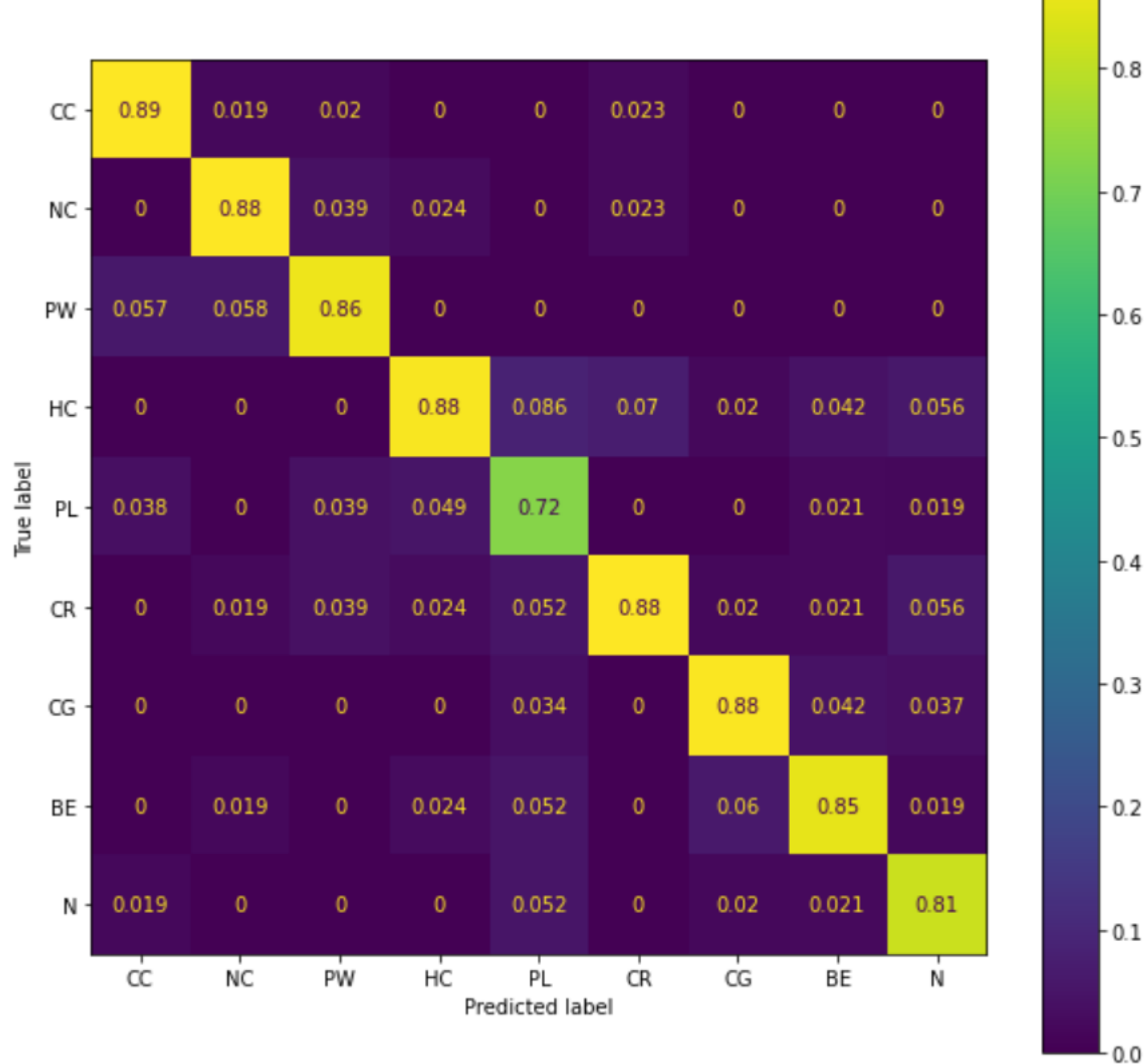


Linear SVM

Test Accuracy: 0.8489

Test Macro-F1 Score: 0.8483

	precision	recall	f1-score	support
CC	0.89	0.94	0.91	50
NC	0.88	0.92	0.90	50
PW	0.86	0.88	0.87	50
HC	0.88	0.72	0.79	50
PL	0.72	0.84	0.78	50
CR	0.88	0.76	0.82	50
CG	0.88	0.88	0.88	50
BE	0.85	0.82	0.84	50
N	0.81	0.88	0.85	50
accuracy			0.85	450
macro avg	0.85	0.85	0.85	450
weighted avg	0.85	0.85	0.85	450

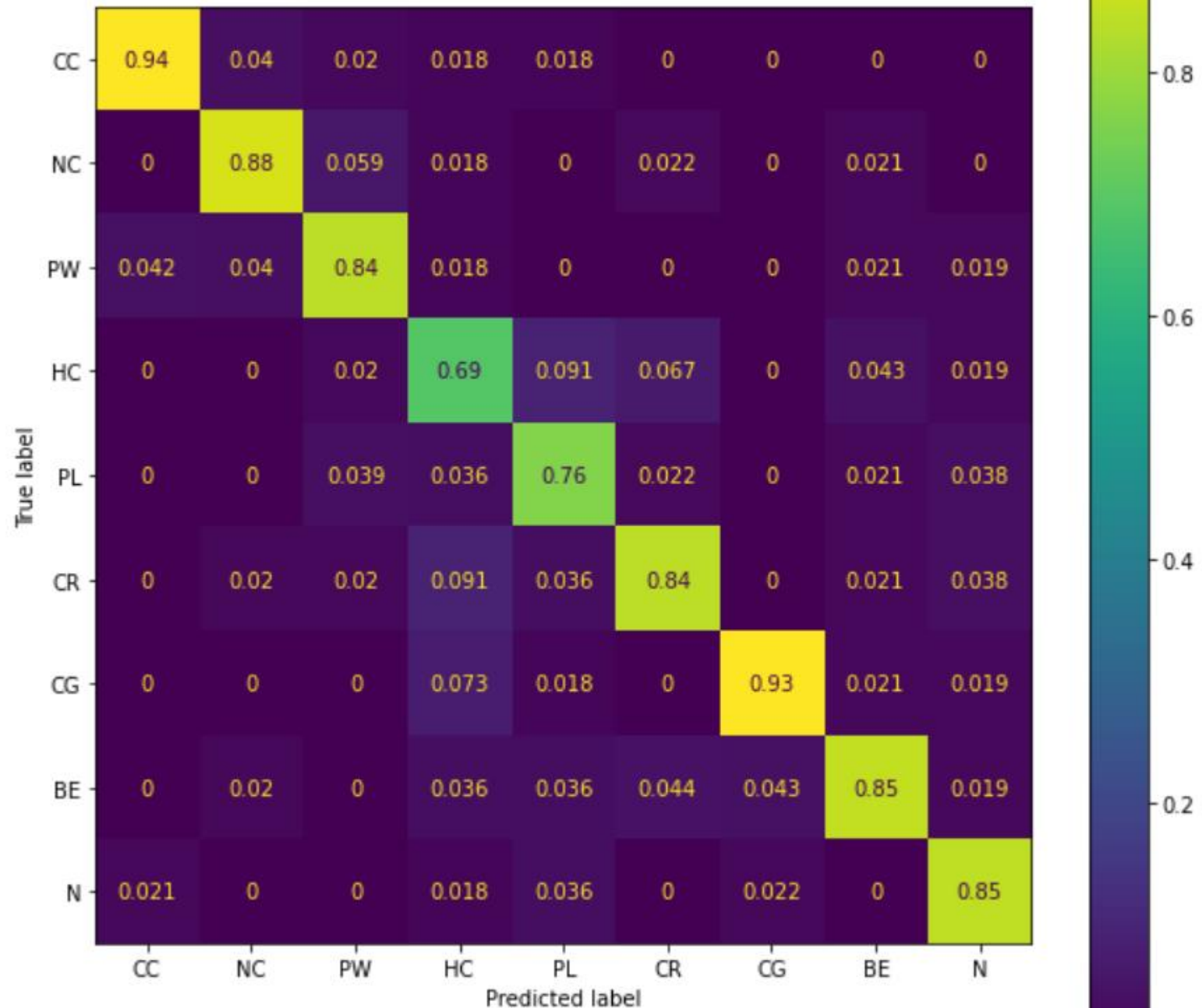


Logistic Regression

Test Accuracy: 0.8400

Test Macro-F1 Score: 0.8409

	precision	recall	f1-score	support
CC	0.94	0.90	0.92	50
NC	0.88	0.88	0.88	50
PW	0.84	0.86	0.85	50
HC	0.69	0.76	0.72	50
PL	0.76	0.84	0.80	50
CR	0.84	0.76	0.80	50
CG	0.93	0.86	0.90	50
BE	0.85	0.80	0.82	50
N	0.85	0.90	0.87	50
accuracy			0.84	450
macro avg	0.84	0.84	0.84	450
weighted avg	0.84	0.84	0.84	450

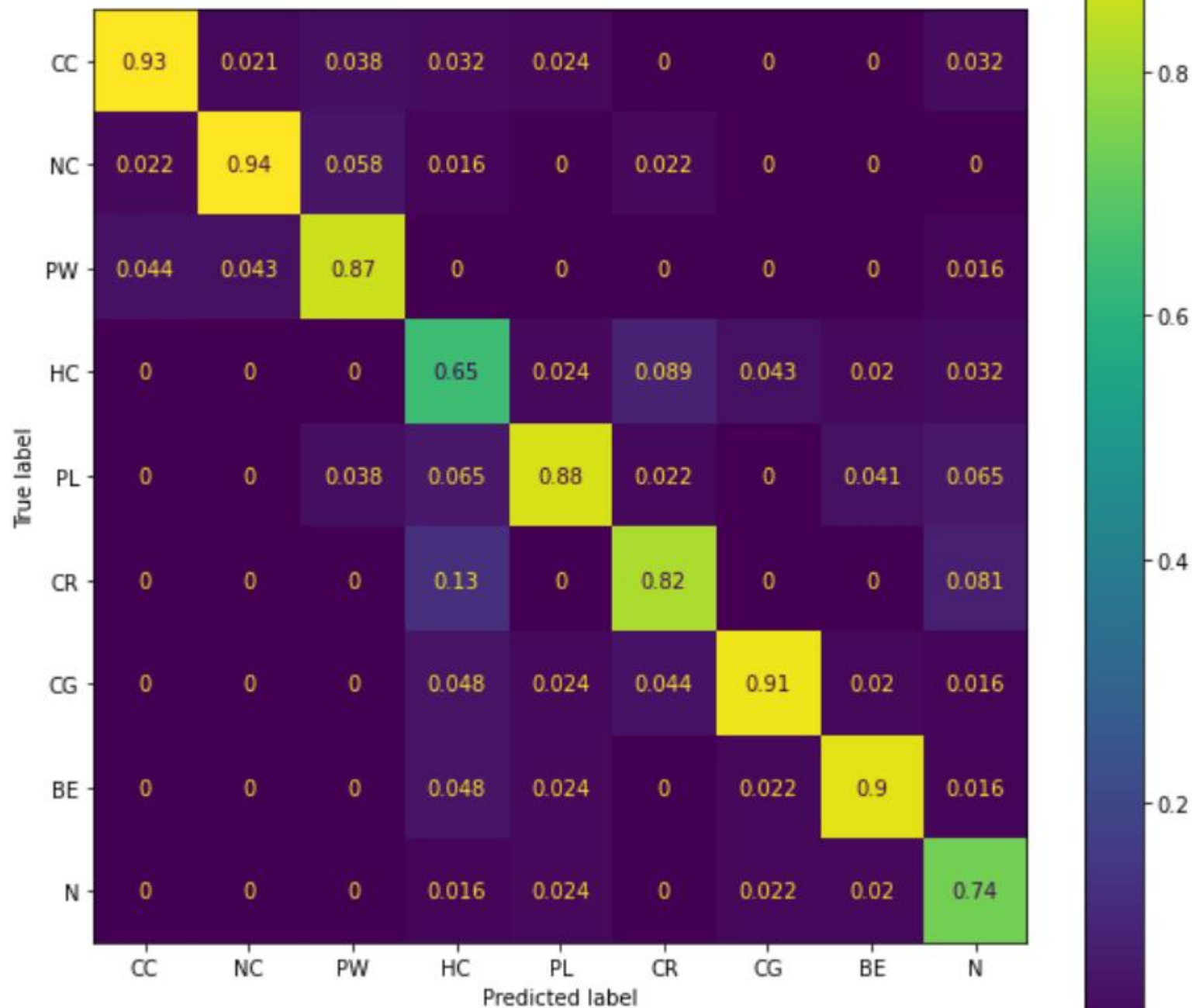


CNN

Test Accuracy: 0.8378

Test Macro-F1 Score: 0.8396

	precision	recall	f1-score	support
CC	0.93	0.84	0.88	50
NC	0.94	0.88	0.91	50
PW	0.87	0.90	0.88	50
HC	0.65	0.80	0.71	50
PL	0.88	0.74	0.80	50
CR	0.82	0.74	0.78	50
CG	0.91	0.84	0.87	50
BE	0.90	0.88	0.89	50
N	0.74	0.92	0.82	50
accuracy			0.84	450
macro avg	0.85	0.84	0.84	450
weighted avg	0.85	0.84	0.84	450

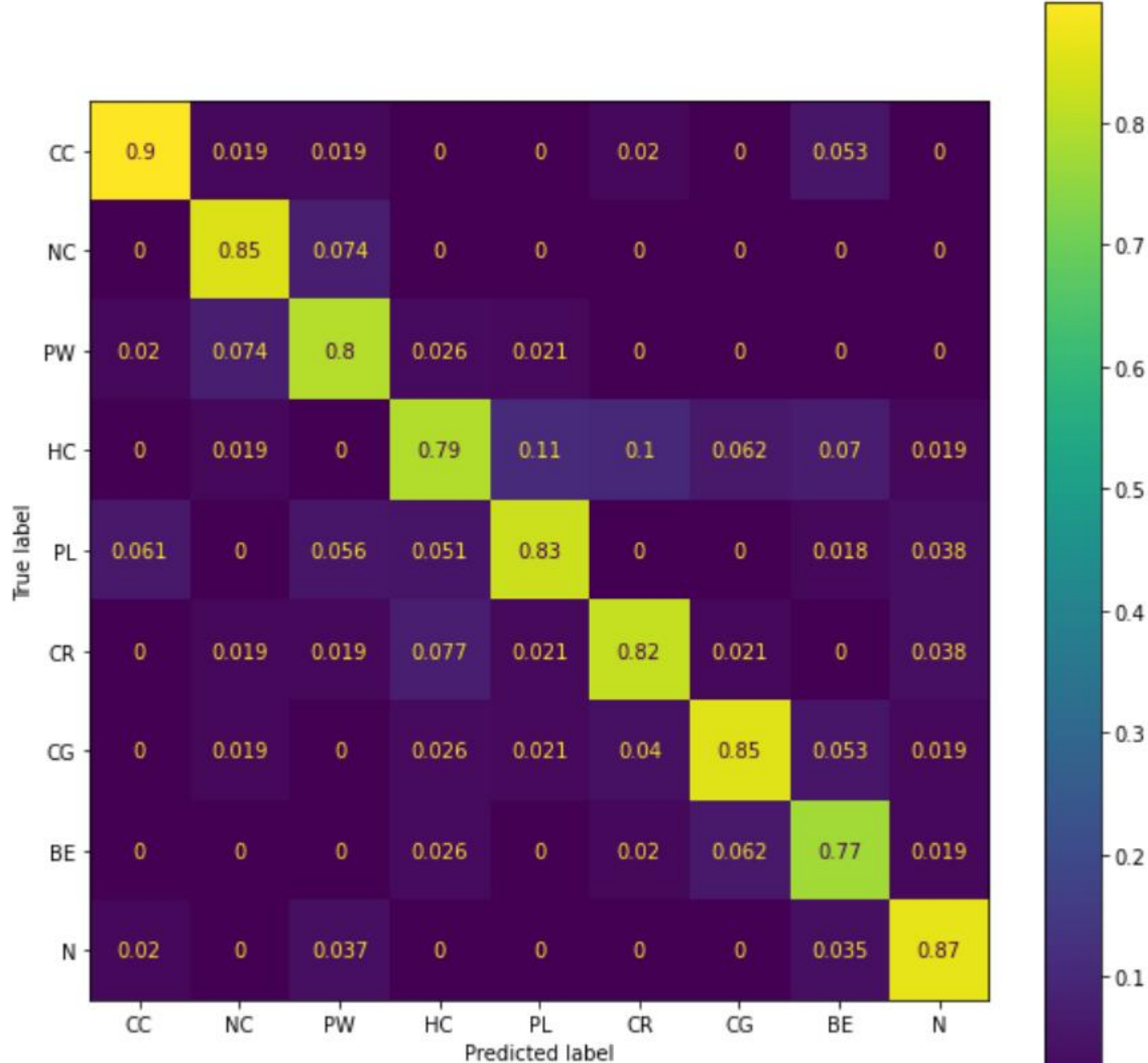


Bi-Directional LSTM

Test Accuracy: 0.8378

Test Macro-F1 Score: 0.8396

	precision	recall	f1-score	support
CC	0.90	0.88	0.89	50
NC	0.85	0.92	0.88	50
PW	0.80	0.86	0.83	50
HC	0.79	0.62	0.70	50
PL	0.83	0.78	0.80	50
CR	0.82	0.82	0.82	50
CG	0.85	0.82	0.84	50
BE	0.77	0.88	0.82	50
N	0.87	0.90	0.88	50
accuracy			0.83	450
macro avg	0.83	0.83	0.83	450
weighted avg	0.83	0.83	0.83	450

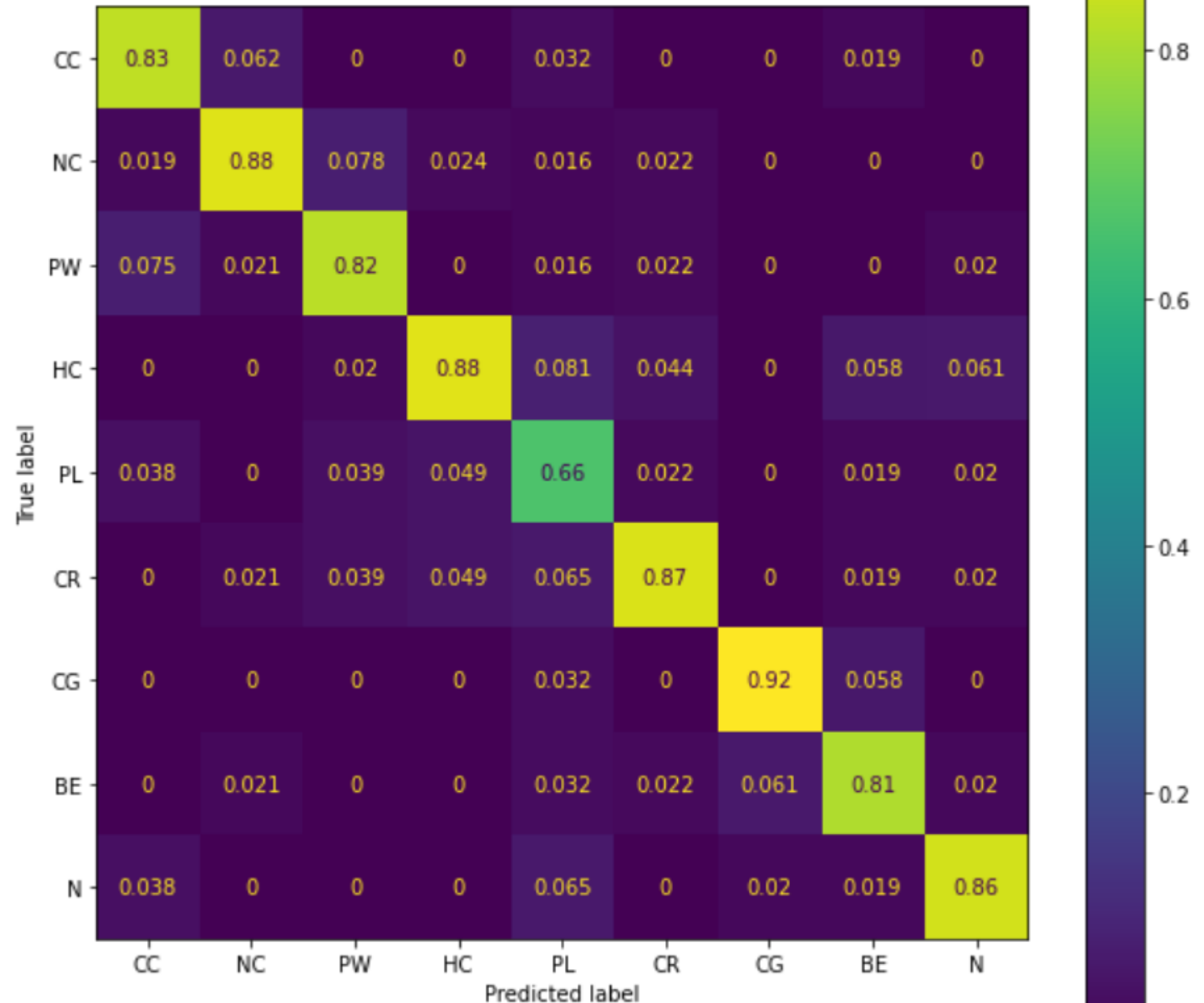


Naïve Bayes

Test Accuracy: 0.8289

Test Macro-F1 Score: 0.8299

	precision	recall	f1-score	support
CC	0.83	0.88	0.85	50
NC	0.88	0.84	0.86	50
PW	0.82	0.84	0.83	50
HC	0.88	0.72	0.79	50
PL	0.66	0.82	0.73	50
CR	0.87	0.78	0.82	50
CG	0.92	0.90	0.91	50
BE	0.81	0.84	0.82	50
N	0.86	0.84	0.85	50
accuracy			0.83	450
macro avg	0.84	0.83	0.83	450
weighted avg	0.84	0.83	0.83	450

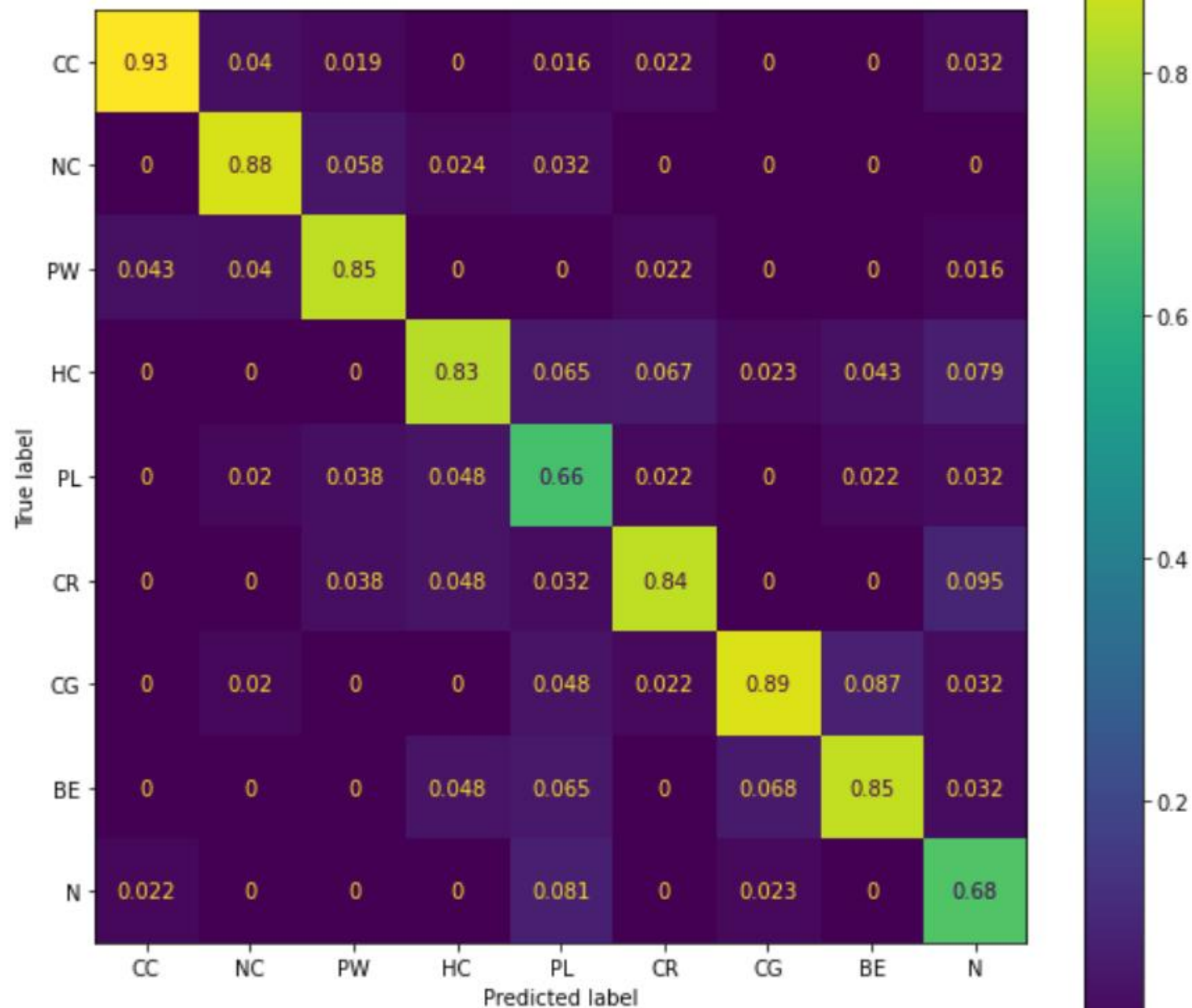


Random Forest

Test Accuracy: 0.8133

Test Macro-F1 Score: 0.8150

	precision	recall	f1-score	support
CC	0.93	0.86	0.90	50
NC	0.88	0.88	0.88	50
PW	0.85	0.88	0.86	50
HC	0.83	0.70	0.76	50
PL	0.66	0.82	0.73	50
CR	0.84	0.76	0.80	50
CG	0.89	0.78	0.83	50
BE	0.85	0.78	0.81	50
N	0.68	0.86	0.76	50
accuracy			0.81	450
macro avg	0.82	0.81	0.81	450
weighted avg	0.82	0.81	0.81	450

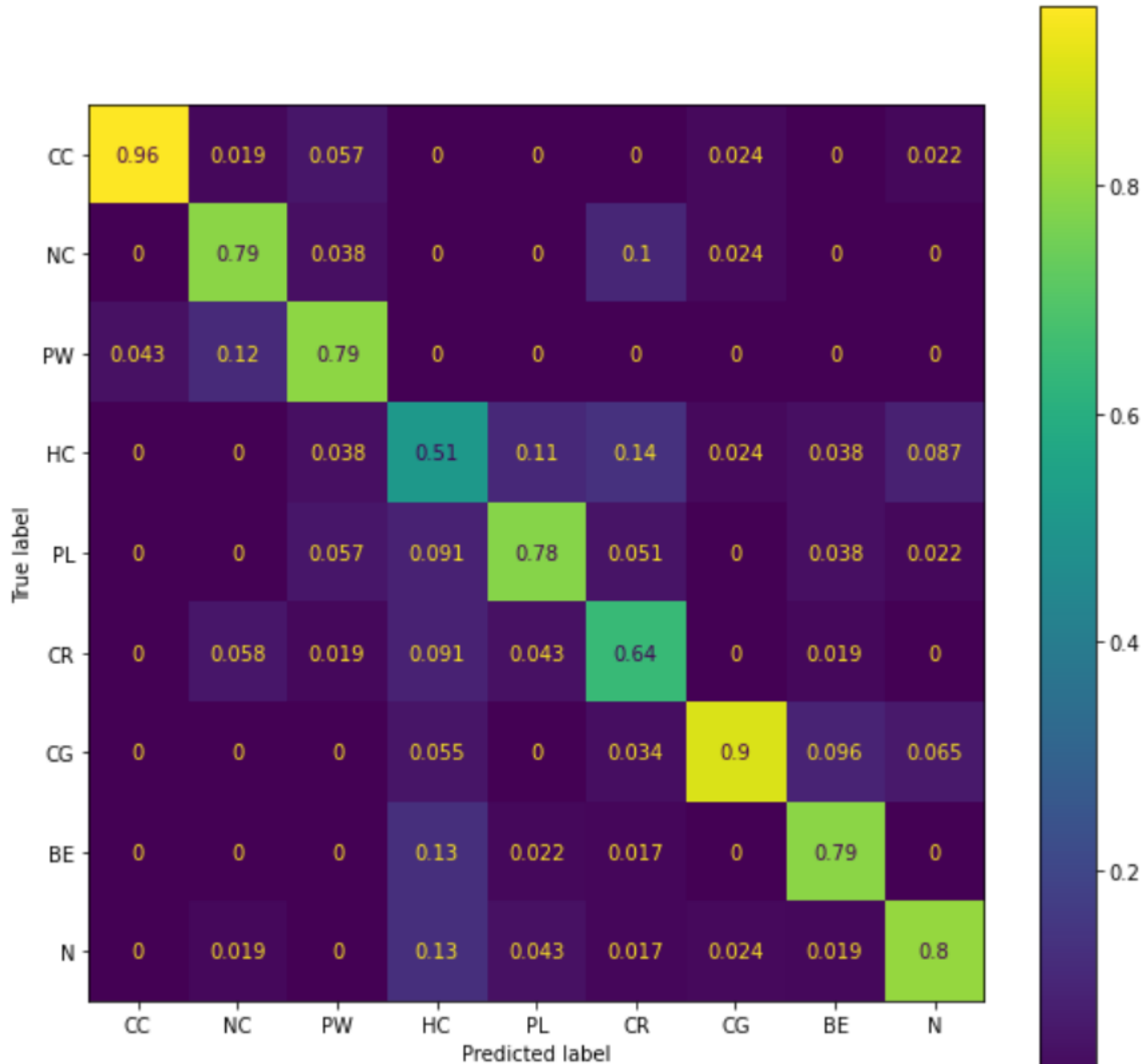


LSTM

Test Accuracy: 0.7644

Test Macro-F1 Score: 0.7672

	precision	recall	f1-score	support
CC	0.96	0.88	0.92	50
NC	0.79	0.82	0.80	50
PW	0.79	0.84	0.82	50
HC	0.51	0.56	0.53	50
PL	0.78	0.72	0.75	50
CR	0.64	0.76	0.70	50
CG	0.90	0.74	0.81	50
BE	0.79	0.82	0.80	50
N	0.80	0.74	0.77	50
accuracy			0.76	450
macro avg	0.77	0.76	0.77	450
weighted avg	0.77	0.76	0.77	450



GRU

Test Accuracy: 0.7489

Test Macro-F1 Score: 0.7543

	precision	recall	f1-score	support
CC	0.93	0.82	0.87	50
NC	0.88	0.74	0.80	50
PW	0.91	0.80	0.85	50
HC	0.49	0.56	0.52	50
PL	0.59	0.68	0.63	50
CR	0.76	0.76	0.76	50
CG	0.89	0.80	0.84	50
BE	0.68	0.78	0.73	50
N	0.75	0.80	0.78	50
accuracy			0.75	450
macro avg	0.77	0.75	0.75	450
weighted avg	0.77	0.75	0.75	450

