Multi-Label Enzyme Class Prediction Using Machine Learning Models

BIOL 3340 – Bioinformatics

Instructor: Dr. Jeffery Demuth

Zainab Siddiqui, Alain Siddiqui

Day 3 of Data Science Bootcamp!

Questions? Feedback? Ideas? Reach out to us!

Zainab: zxs2546@mavs.uta.edu

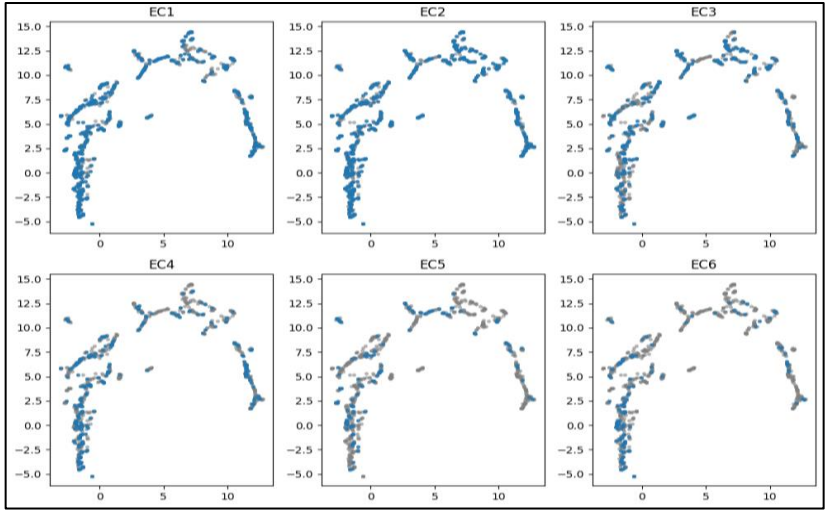
Alain: axs6901@mavs.uta.edu

Background & Research Question

- Biological or computational problem at hand:
  - Enzyme promiscuity and overlapping functions make EC classification difficult.
  - Traditional sequence-based tools struggle to accurately predict multi-label enzyme class
- Why does this matter?
  - Accurate functional annotation is essential for drug discovery and biotechnology
  - Misclassification can affect drug safety
  - Improved computational models reduce reliance on costly experiments
- Research Question: Can machine learning models using molecular fingerprints and biochemical features improve multi-label enzyme class prediction?
  - Hypothesis: Combining molecular-level descriptors with ML models will outperform traditional sequence-only approaches in predicting EC classes

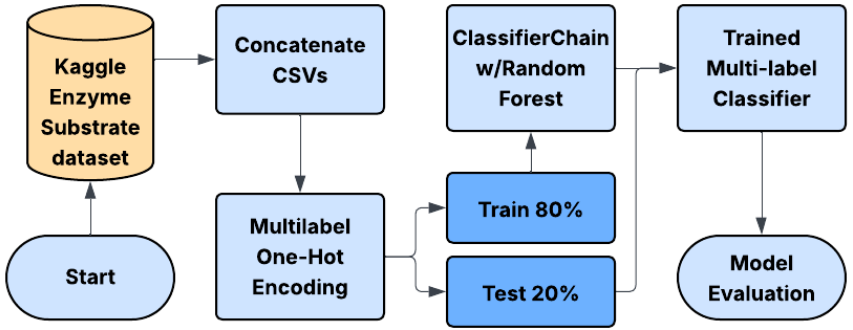
Data & Methods

- Source: Kaggle's [Multi-label Classification of Enzyme Substrates dataset]
- Type(s): tabular CSVs of molecular/chemical fingerprint data
- Size / structure:
  - 3.2 MB
  - 1039 instances
  - 1229 features split across 3 CSV files



UMAP plots of data, color-coded per class.

- Key tools: Python, Colab, SciKit-Learn
  - Models: Random Forest & Classifier Chain
  - PCA/kPCA, MLSMOTE, RFECV
- Evaluation metrics:
  - Jaccard Score (naturally handles multilabel classification)
    - Train and test Jaccard scores were recorded to investigate overfitting
  - Recall (controls for false negatives)
  - Precision (controls for false positives)



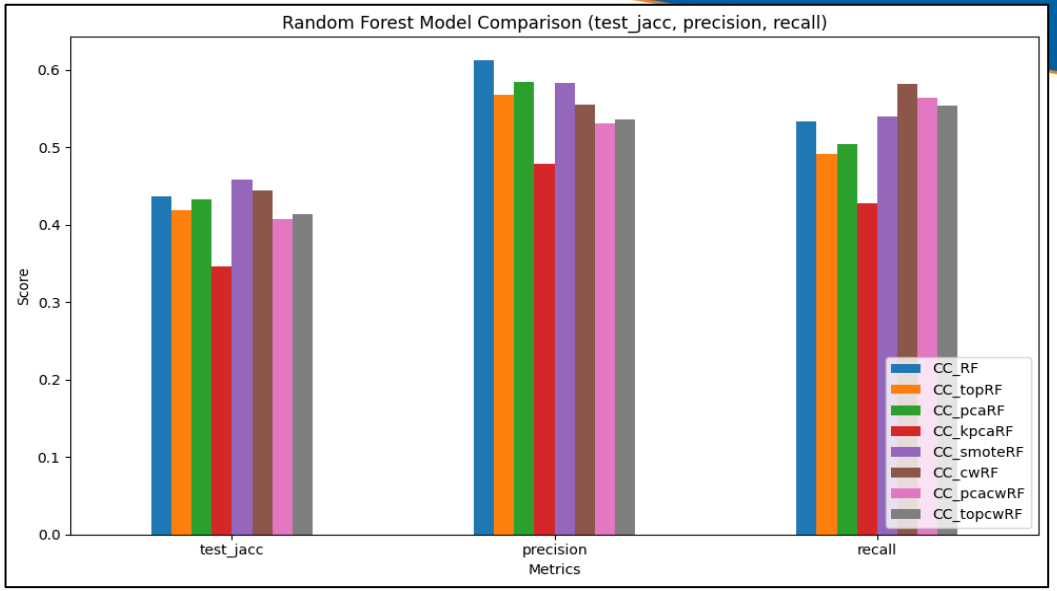
Flowchart of modelling pipeline.

Results

CC_cwRF (Brown)	precision	recall	f1-score	support
0	0.71	0.72	0.71	121
1	0.61	0.71	0.66	112
2	0.49	0.48	0.49	64
3	0.32	0.34	0.33	44
4	0.32	0.30	0.31	30
5	0.44	0.33	0.38	24
micro avg	0.56	0.58	0.57	395
macro avg	0.48	0.48	0.48	395
weighted avg	0.56	0.58	0.57	395
samples avg	0.55	0.57	0.52	395

CC_RF (Blue)	precision	recall	f1-score	support
0	0.73	0.70	0.71	128
1	0.58	0.77	0.66	115
2	0.45	0.31	0.37	49
3	0.56	0.21	0.31	47
4	0.56	0.16	0.25	31
5	0.67	0.08	0.14	26
micro avg	0.62	0.53	0.57	396
macro avg	0.59	0.37	0.41	396
weighted avg	0.61	0.53	0.53	396
samples avg	0.61	0.53	0.52	396

Classification report comparison between top RF models (according to bar graph). The class-weighted model is observably more consistent despite lower test Jaccard Score and precision.

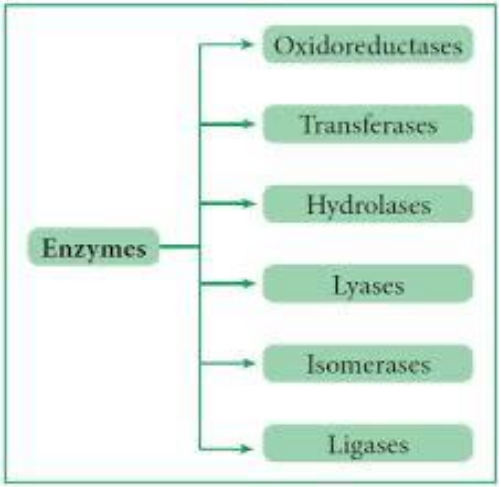


Metrics comparison of baseline (i.e., one vs. rest) RF models. Values are inflated due to the simplicity of binary classification, with minority classes suffering from underfitting.

Class	Majority Class %	Test Acc.	Minority Precision	Minority Recall	Weighted F1
EC1	0.54	0.62	0.57	0.56	0.61
EC2	0.58	0.64	0.52	0.44	0.64
EC3	0.71	0.74	0.61	0.35	0.71
EC4	0.77	0.74	0.44	0.13	0.68
EC5	0.87	0.86	0.2	0.08	0.83
EC6	0.87	0.88	0.38	0.13	0.85

Interpretation & Conclusion

- Overlapping enzyme functions and rare classes make dimensionality reduction and standard ML methods less effective
- Class-weighted Random Forest improved prediction of minority enzymes, reflecting the importance of capturing rare but biologically meaningful functions
- Our machine learning models partially outperform sequence-only approaches
  - Perform well on common classes but struggle with overlap
- Conclusions:
  - Tree-based models with class weighting provide robust predictions
- Limitations
  - Class imbalance
  - High-dimensional features
  - Small, “wide” dataset
- Future Works
  - Add domain-guided feature engineering
  - Custom stacked models
  - Dimensionality reduction
  - Robust class imbalance techniques



Reproducibility & Acknowledgements

- Reproducibility:
- GitHub repo.
- AI Use: Visualizations and simplifying code-logic
- Acknowledgments
  - Kaggle
- Data Source: https://www.kaggle.com/datasets/gopalns/ec-mixed-class?select=mixed_ecfp.csv
- Key Reference: Visani, G. M., Hughes, M. C., & Hassoun, S. (2021). *Enzyme promiscuity prediction using hierarchy-informed multi-label classification*. arXiv. https://doi.org/10.48550/arXiv.2106.00870