Supplementary Material – Classification Performance and Error Analysis

Yeast Protein Localization via Machine Learning

Martina Castellucci

0. Dataset Overview and Feature Description

The Yeast dataset contains 1,484 proteins annotated with their subcellular localization. Each protein is described by 8 numerical features extracted from the amino acid sequence, reflecting structural, signal-related and motif-based biological properties relevant to intracellular protein sorting. The table below shows an example of 5 proteins and their respective features.

Name	**mcg**	**gvh**	**alm**	**mit**	**erl**	**pox**	**vac**	**nuc**	**Class**
ADT1	0.58	0.61	0.47	0.13	0.50	0.0	0.48	0.22	MIT
ADT2	0.43	0.67	0.48	0.27	0.50	0.0	0.53	0.22	MIT
ADT3	0.64	0.62	0.49	0.15	0.50	0.0	0.53	0.22	MIT
AAR2	0.58	0.44	0.57	0.13	0.50	0.0	0.54	0.22	NUC
AATM	0.42	0.44	0.48	0.54	0.50	0.0	0.48	0.22	MIT

Table 1: Sample of 5 proteins with their numerical features and class labels. Features were derived from biological sequence analysis and computational signal prediction tools.

Feature Description

- **mcg** – McGeoch's method score: quantifies the presence of signal peptides in the N-terminal region using hydrophobicity and segmental amino acid composition.
- **gvh** – von Heijne's method: detects signal peptides based on position-specific scoring matrices and conserved cleavage motifs.
- **alm** – ALOM score for predicting transmembrane helices: useful for distinguishing membrane-bound vs. soluble proteins.
- **mit** – Discriminant score for mitochondrial targeting: reflects amino acid biases specific to mitochondrial import sequences.
- **erl** – Presence of the HDEL motif (binary): this ER-retention signal ensures protein localization in the endoplasmic reticulum lumen.
- **pox** – Peroxisomal targeting signal presence (binary): based on consensus motifs usually found at the C-terminus.

- **vac** – Score from discriminant analysis separating vacuolar from extracellular proteins.
- **nuc** – Nuclear localization score: derived from NLS motif recognition using discriminant functions.
- **Class** – Target label indicating the subcellular localization site. Values include: CYT, NUC, MIT, ME1, ME2, EXC, VAC, POX, ERL.

1. Feature Distribution and Outliers

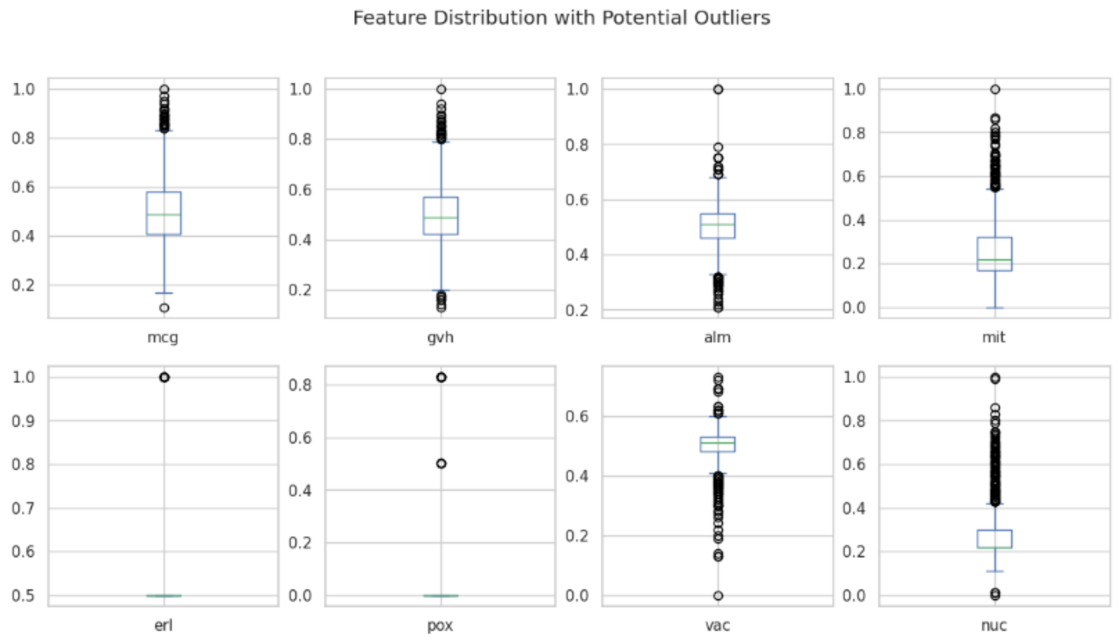
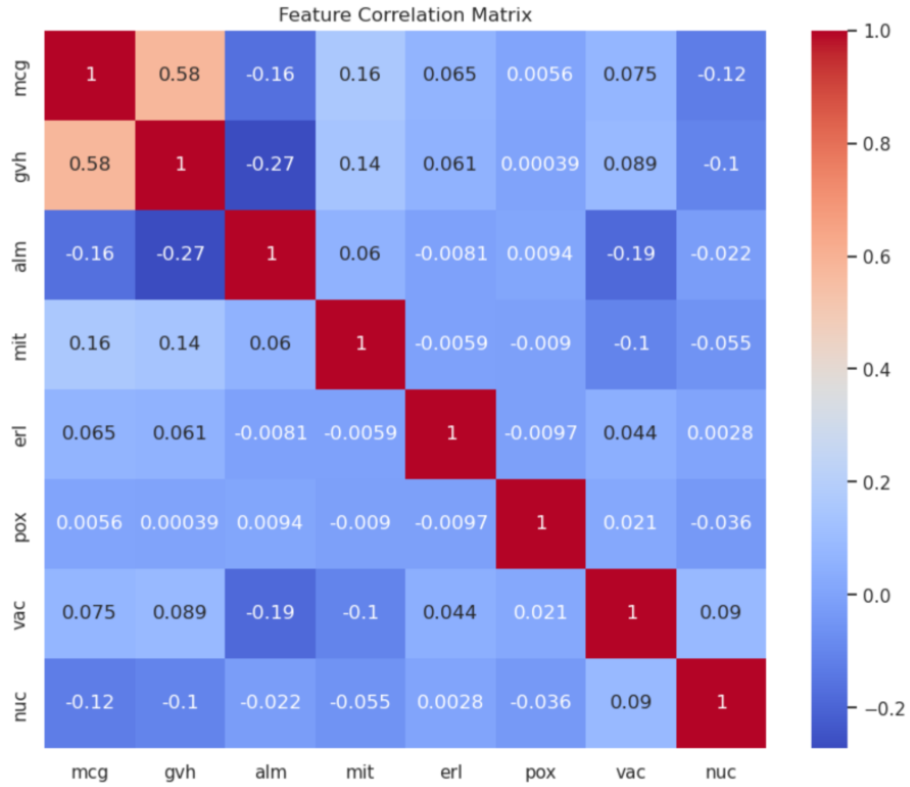


Figure 1: Boxplot of the 8 numerical features in the Yeast dataset. This visualization highlights both dispersion and asymmetry across features. Notably, features such as `mcg`, `gvh`, `alm`, `mit`, and `nuc` exhibit long tails and outliers, indicating potential non-normality. Features like `pox` and `erl` have low variance, suggesting limited discriminatory capacity. These insights guide preprocessing decisions such as normalization, transformation, and feature engineering.

2. Feature Correlation Matrix



	mcg	gvh	alm	mit	erl	pox	vac	nuc
mcg	1	0.58	-0.16	0.16	0.065	0.0056	0.075	-0.12
gvh	0.58	1	-0.27	0.14	0.061	0.00039	0.089	-0.1
alm	-0.16	-0.27	1	0.06	-0.0081	0.0094	-0.19	-0.022
mit	0.16	0.14	0.06	1	-0.0059	-0.009	-0.1	-0.055
erl	0.065	0.061	-0.0081	-0.0059	1	-0.0097	0.044	0.0028
pox	0.0056	0.00039	0.0094	-0.009	-0.0097	1	0.021	-0.036
vac	0.075	0.089	-0.19	-0.1	0.044	0.021	1	0.09
nuc	-0.12	-0.1	-0.022	-0.055	0.0028	-0.036	0.09	1

Figure 2: Pearson correlation heatmap among the 8 numerical features. Only `mcg` and `gvh` display moderate correlation ($r = 0.58$). All other pairwise correlations are weak or negligible ($|r| < 0.3$), indicating low redundancy among predictors. This suggests multicollinearity is not a concern for this dataset, and dimensionality reduction techniques such as PCA are not strictly necessary.

Table 1. Pearson Correlation Matrix

	mcg	gvh	alm	mit	erl	pox	vac	nuc
mcg	1.00	0.58	-0.16	0.16	0.07	0.006	0.08	-0.12
gvh		1.00	-0.27	0.14	0.06	0.000	0.09	-0.10
alm			1.00	0.06	-0.01	0.009	-0.19	-0.02
mit				1.00	-0.01	-0.009	-0.10	-0.06
erl					1.00	-0.010	0.04	0.003
pox						1.00	0.021	-0.036
vac							1.00	0.09
nuc								1.00

3. Class Distribution

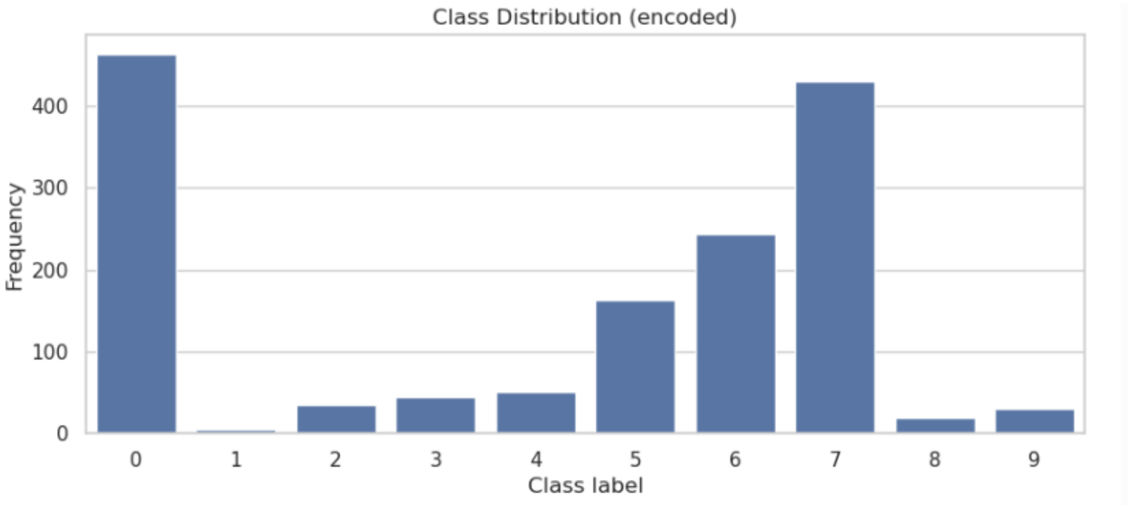


Figure 3: Barplot showing the number of samples per class. The dataset is severely imbalanced: most examples belong to classes CYT and NUC, while rare classes such as ERL and POX have very few samples (fewer than 10). This imbalance can bias the classifier toward the majority classes, decreasing recall for minority classes. To address this, SMOTE (Synthetic Minority Oversampling Technique) and class weighting were applied during model training.

4. ROC Curve – One-vs-Rest

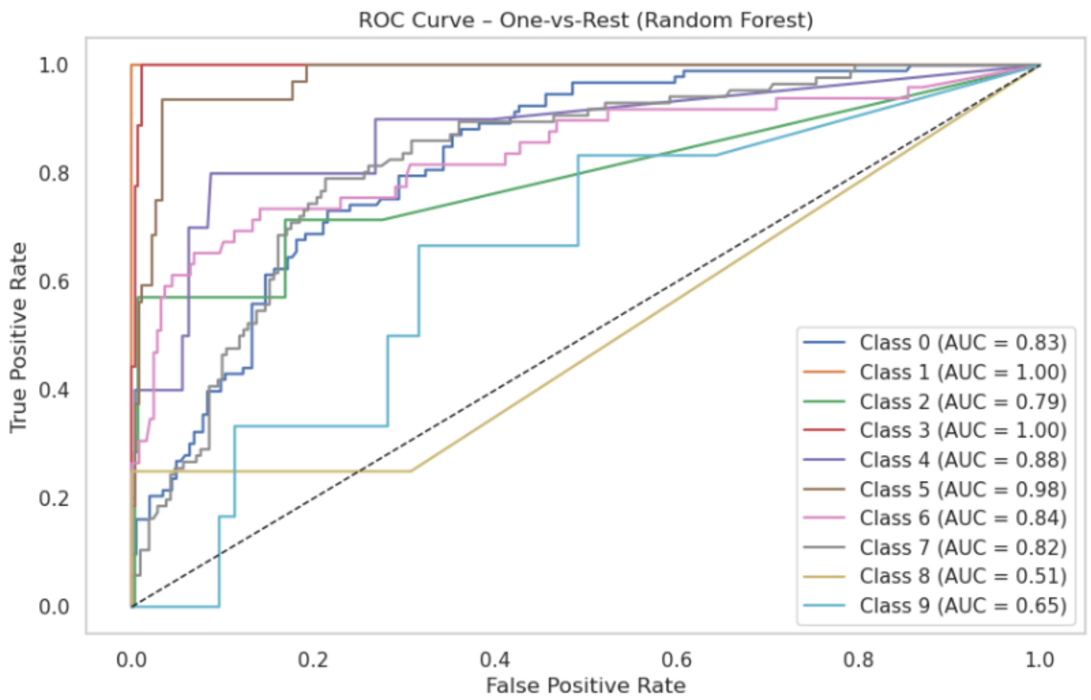


Figure 4: Receiver Operating Characteristic (ROC) curves per class using a One-vs-Rest strategy. The ROC curve evaluates the trade-off between true positive rate (sensitivity) and false positive rate. Perfect AUC values (1.00) in small classes (e.g., ERL, ME2) may reflect overfitting due to very low support. Classes such as POX and PER show poor separation ($AUC < 0.65$), which is typical in highly imbalanced or overlapping distributions.

Table 2. ROC AUC and Class Characteristics

Class	AUC	Support (samples)	Notes
CYT (0)	0.83	463	Well generalized
ERL (1)	1.00	5	Likely overfit
ME1 (2)	0.79	44	Medium sensitivity
ME2 (3)	1.00	20	Risk of overfitting
EXC (4)	0.88	35	Consistent performance
VAC (5)	0.98	30	High separability
MIT (6)	0.84	44	Robust detection
NUC (7)	0.82	429	Strong class balance
POX (8)	0.51	4	Near-random prediction
PER (9)	0.65	30	Low discriminative power

5. Precision–Recall Curve

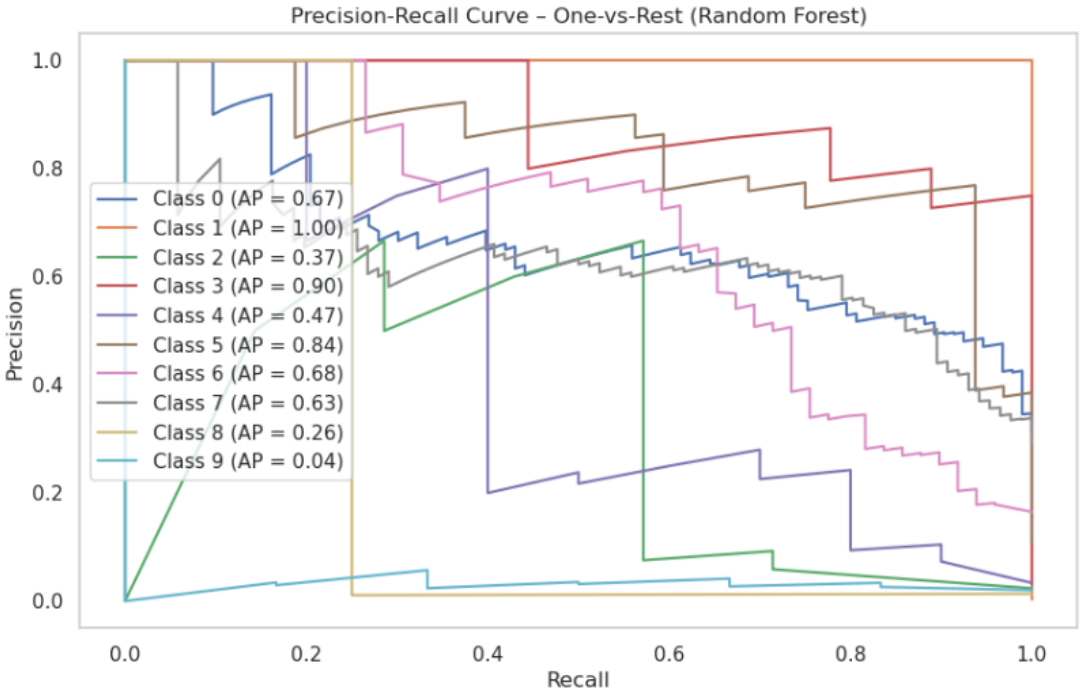


Figure 5: Precision-Recall curves per class. These curves are particularly suited for evaluating classifier performance under class imbalance. While classes such as VAC and MIT display high AP (average precision), minority classes like POX and PER perform poorly, with precision dropping below 0.2. This reflects a high number of false positives and missed predictions.

Table 3. Average Precision and Error Risk

Class	AP	Support (samples)	Risk Analysis
CYT (0)	0.67	463	Reliable predictions with occasional FP
ERL (1)	1.00	5	Overconfident due to low support
ME1 (2)	0.37	44	Many false negatives
ME2 (3)	0.90	20	High precision, few examples
EXC (4)	0.47	35	Balanced precision/recall
VAC (5)	0.84	30	Excellent trade-off
MIT (6)	0.68	44	Robust class
NUC (7)	0.63	429	Stable and well-supported
POX (8)	0.26	4	Class mostly ignored
PER (9)	0.04	30	Severely misclassified

6. Error Analysis Summary

Class	Common Misclassification	Likely Explanation
PER (9)	Predicted as NUC (7)	Similarity in amino acid patterns, weak signals
POX (8)	Predicted as CYT (0)	Very low support, no strong motifs
ME1 (2)	Confused with ME2 (3)	Overlapping hydrophobic profiles
EXC (4)	Confused with VAC (5)	Close functional roles, similar signals
ERL (1)	Misclassified inconsistently	Data sparsity, potential noise

Table 2: Misclassification patterns based on confusion matrix inspection. Most errors involve rare classes or biologically similar categories.

Interpretation: Errors primarily stem from class imbalance and overlapping feature distributions. Suggested improvements include: advanced oversampling techniques, feature engineering focused on motif detection, and ensemble or hierarchical models tailored to protein sorting hierarchies.