Yeast Protein Classification: Machine Learning Pipeline for Subcellular Localization

Martina Castellucci
MSc in Bioinformatics, University of Bologna

A.A 2024–2025

Abstract

Motivation: Subcellular protein localization is a challenging supervised learning task involving class imbalance, low-dimensional inputs, and overlapping signals. The UCI Yeast dataset offers 10 protein classes with only 8 numeric features. Its imbalanced structure and biological noise provide a relevant testbed for evaluating robust and fairness-aware classifiers.

Approach: A supervised learning pipeline was implemented, including standardization, label encoding, stratified splitting, and SMOTE resampling applied only to the training set. Four classifiers were evaluated—Logistic Regression, Random Forest, Support Vector Machine (C=5), and k-NN (k=5)—with hyperparameter tuning via GridSearchCV and model selection based on macro-F1 and MCC. The pipeline emphasizes robustness under class imbalance and fairness in multiclass evaluation.

Results: Random Forest and SVM achieved top performance (Accuracy = 0.98, Macro-F1 = 0.94, MCC = 0.98). In contrast, k-NN underperformed on the test set despite strong cross-validation. Confusion matrices revealed predictable misclassifications (e.g., CYT vs MIT). SMOTE improved recall on medium-frequency classes but remained limited for ultra-rare ones.

Contact: martina.castellucci@studio.unibo.it

Resources: Full pipeline and results: `https://github.com/Martinaa1408/ML_basic_project/`

Contents

Glossary

- **SMOTE** – Synthetic Minority Over-sampling Technique. Creates new minority class examples by interpolating between real instances and their nearest neighbors.
- **StandardScaler** – Scales features to have zero mean and unit variance, preventing dominance of high-magnitude features.
- **Stratified Split** – Ensures that class proportions are preserved when splitting data into train and test sets.
- **GridSearchCV** – Performs cross-validated grid search over hyperparameters, selecting the combination with best validation performance.
- **Random Forest** – Ensemble of decision trees trained on bootstrapped subsets with feature randomness. Handles non-linear interactions well.
- **SVM (Support Vector Machine)** – Learns a decision boundary that maximizes the margin between classes. Linear SVM used here.
- **k-NN (k-Nearest Neighbors)** – Instance-based method that predicts class by majority vote among k closest training points.
- **Logistic Regression** – Linear model that predicts class probabilities using a sigmoid function; suitable for baseline classification.
- **Macro F1** – Unweighted average of F1-scores for each class, giving equal weight to all labels.
- **MCC (Matthews Correlation Coefficient)** – Robust metric for imbalanced multiclass classification; reflects all confusion matrix terms.
- **ROC-AUC** – Area under the Receiver Operating Characteristic curve, computed in a one-vs-rest fashion for multiclass settings.
- **PR-AUC** – Area under the Precision-Recall curve. More informative than ROC-AUC when dealing with high class imbalance.
- **Confusion Matrix** – Matrix comparing true vs predicted class labels. Helps identify systematic errors.
- **One-vs-Rest (OvR)** – Decomposes a multiclass problem into multiple binary ones: one class vs all others.
- **class_weight=b́alanced´** – Scikit-learn option that adjusts class weights inversely proportional to class frequencies.

1 Introduction

1.1 Biological Background

Subcellular localization is a critical attribute of proteins, determining the spatial context in which they perform biological functions such as signaling, metabolism, or transport. Mislocalization can lead to functional disruption, aberrant interactions, or even disease states. Accurately predicting protein localization therefore contributes to both systems biology and biomedical research.

Computational localization predictors have emerged to complement experimental methods, which are often slow, costly, and biased toward well-studied compartments. Among benchmark datasets, the UCI Yeast dataset provides a practical and widely-used test case. It includes 1,484 proteins from *Saccharomyces cerevisiae*, each annotated with one of 10 localization sites and described by 8 numeric features derived from sequence analysis tools.

These features capture physicochemical signals linked to localization patterns, including signal peptides, hydrophobicity, and N-terminal motifs. However, they are relatively sparse and low-dimensional, with partially overlapping signals across compartments. Furthermore, the class distribution is strongly skewed: the majority of instances belong to a few classes (e.g., CYT, NUC), while others such as POX and ERL appear in fewer than 10 examples.

1.2 Machine Learning Framing and Challenges

The task can be formulated as a supervised multiclass classification problem:

$$f : R^8 \to \{1, \ldots, 10\}$$

Key challenges include:

- **Imbalanced class distribution:** algorithms tend to overfit dominant classes, underrepresenting rare ones.
- **Limited input space:** only 8 numerical features, requiring careful preprocessing and robust modeling.
- **Feature ambiguity:** similar signals can arise in different compartments, leading to confusions (e.g., CYT MIT).
- **Evaluation difficulty:** metrics such as accuracy can be misleading under imbalance; MCC and macro-F1 offer better class-level fairness.

This study develops a robust and reproducible pipeline to address these issues, with a focus on fairness-aware evaluation, over-sampling strategies (SMOTE), and interpretable performance diagnostics.

2 Aim of the Study

The goal of this study is to design and evaluate a robust, reproducible machine learning pipeline for the prediction of subcellular protein localization based on numeric sequence-derived features. The key objectives include:

- To implement a supervised multiclass classification framework for 10 subcellular compartments using the UCI Yeast dataset.
- To address class imbalance through synthetic resampling techniques (SMOTE) and algorithmic strategies such as balanced class weighting.
- To benchmark the performance of four distinct classifiers: Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Random Forest.
- To optimize model performance via cross-validated hyperparameter tuning using GridSearchCV.
- To evaluate results using robust metrics suited for imbalanced multiclass classification, including Macro-F1, Matthews Correlation Coefficient (MCC), ROC-AUC, PR-AUC, and confusion matrix analysis.

Raw Dataset (UCI Yeast)

⇓

Feature Selection and Encoding

(Drop categorical, Label Encoding of target)

⇓

Train/Test Split (Stratified 80/20)

(with fixed random seed)

⇓

Standardization (StandardScaler)

(fit on training set only)

⇓

Balancing with SMOTE

(dynamic *k* based on minority class, only on training set)

⇓

Model Training & Hyperparameter Tuning

(GridSearchCV with 5-fold stratified CV)

(LogReg, SVM (C), RF, k-NN)

⇓

Model Evaluation on Held-out Test Set

(Accuracy, Macro-F1, MCC, PR-AUC, Confusion Matrix)

Figure 1: Expanded schematic of the supervised learning pipeline, including preprocessing, balancing, training and evaluation.

3 Materials and Methods

3.1 Dataset Description

The dataset used in this project is the well-known Yeast dataset from the UCI Machine Learning Repository. It contains 1,484 proteins from the model organism *Saccharomyces cerevisiae*, each labeled with its experimentally validated subcellular localization. Every instance is represented by 8 numeric features derived from biological sequence analysis, reflecting signals typically exploited in protein sorting and transport mechanisms.

These features are:

- `mcg` – McGeoch's method score for signal sequences;
- `gvh` – von Heijne's score for signal peptide cleavage;
- `alm` – score from an amino acid composition model;
- `mit` – mitochondrial targeting signal;
- `erl` – binary flag for presence of ER-retention signal (HDEL);
- `pox` – discriminant score for peroxisomal targeting;
- `vac` – vacuolar sorting signal strength;
- `nuc` – nuclear localization signal predictor.

During feature analysis, `pox` and `erl` were found to be nearly constant (>98.5%) and removed from further modeling due to low variance and poor discriminatory power.This left 6 effective features for model input.

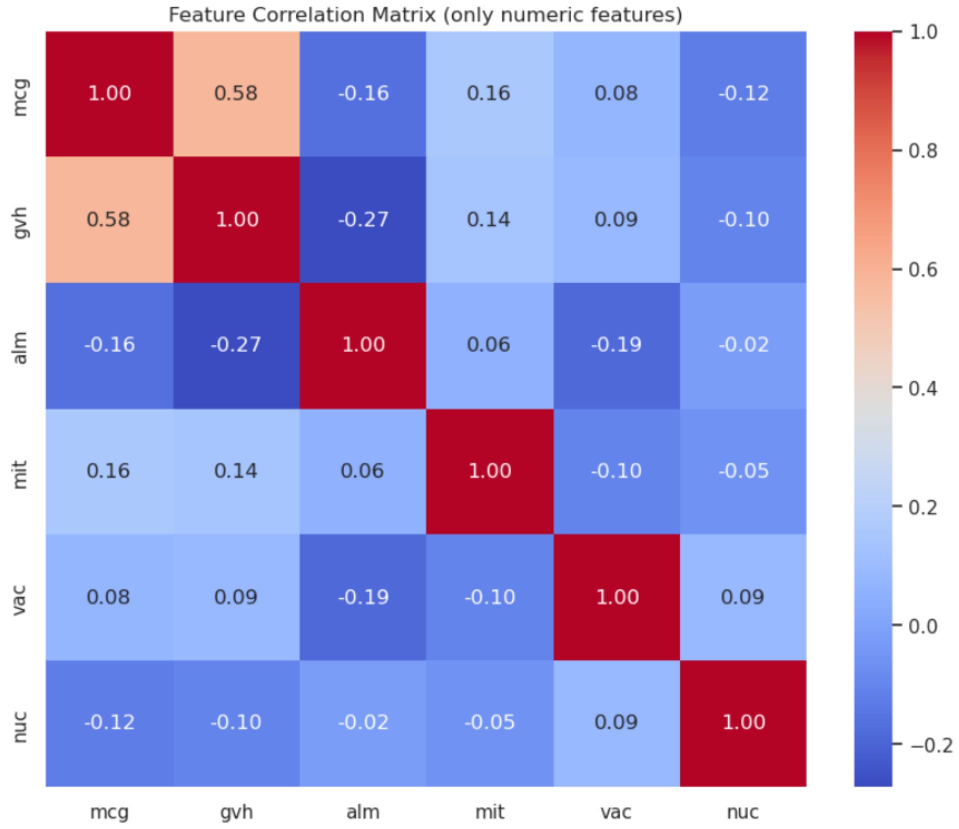


Figure 2: Feature correlation matrix. `mcg` and `gvh` show moderate correlation (0.58), suggesting redundancy. All other features exhibit low inter-correlation, which supports their joint inclusion as complementary descriptors.

Each protein is annotated with one of 10 localization classes: `CYT` (cytoplasm), `NUC` (nucleus), `MIT` (mitochondria), `ME1` `ME3` (variants of membrane proteins), `POX` (peroxisome), `ERL` (endoplasmic reticulum), `EXC` (extracellular), and `VAC` (vacuole). The class distribution is severely imbalanced. For example, `CYT` and `NUC` together comprise over 60% of the samples, while rare classes like `ERL` and `POX` appear fewer than 10 times in the entire dataset. This imbalance imposes substantial difficulty for standard classifiers.

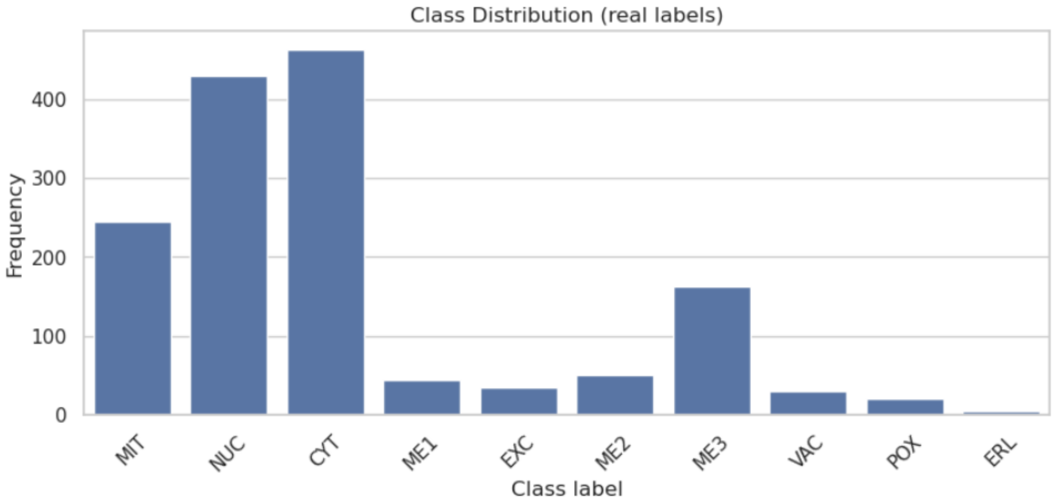


Figure 3: Class frequency distribution. The dataset is dominated by `CYT`, `NUC`, and `MIT`, while `ERL` and `POX` are extreme outliers in terms of support.

3.2 Preprocessing

The data preprocessing stage focused on ensuring consistency, robustness, and fair train-test evaluation. All numeric features were scaled to zero mean and unit variance using `StandardScaler` to prevent any dominance from high-magnitude features. Class labels, originally categorical strings, were converted into integer format using `LabelEncoder`.

To preserve class proportions and reduce sampling bias, a stratified train-test split (80%–20%) was performed using a fixed `random_state=42`. This ensured that even rare classes were minimally represented in both sets.

Two complementary strategies were adopted to mitigate class imbalance during training:

1. **class_weight="balanced"** was enabled in all models that supported it (e.g., Logistic Regression, SVM).
2. **SMOTE (Synthetic Minority Over-sampling Technique)** was applied exclusively on the training set. This algorithm generates synthetic samples of minority classes by interpolating between real observations and their nearest neighbors. The number of neighbors used was dynamically chosen based on the minority class size, and oversampling was excluded for ultra-rare classes (like `ERL`) to avoid synthetic overfitting.

All preprocessing steps—including scaling, splitting, and resampling—were wrapped into modular functions and executed with reproducibility in mind.

3.3 Model Training and Hyperparameter Optimization

Four supervised learning algorithms were evaluated, selected to reflect a diverse range of modeling paradigms:

- **Logistic Regression** – a linear baseline with interpretability and low variance;
- **Support Vector Machine (SVM)** – linear kernel; tuned on regularization parameter C;
- **k-Nearest Neighbors (k-NN)** – instance-based learning algorithm evaluated with $k = 3, 5, 7$;

- **Random Forest** – ensemble of decision trees with strong generalization on non-linear spaces.

Model selection and hyperparameter tuning were carried out using `GridSearchCV`, with 5-fold stratified cross-validation and scoring based on macro F1 and MCC. The final selected models were retrained on the full resampled training set before final testing.

3.4 Evaluation Metrics

The model evaluation focused on metrics that are robust to class imbalance and informative for multiclass classification:

Accuracy: General correctness across all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Proportion of true positives among predicted positives.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Proportion of true positives retrieved over actual positives.

$$Recall = \frac{TP}{TP + FN}$$

F1-score: Harmonic mean of precision and recall.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Macro F1: Unweighted average of class-specific F1-scores, treating each class equally.

Weighted F1: F1-score weighted by class support.

Matthews Correlation Coefficient (MCC): Balanced metric suitable for imbalanced multiclass problems.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

ROC-AUC: Measures the area under the receiver operating characteristic curve in a one-vs-rest scheme.

PR-AUC (Average Precision): Area under the precision-recall curve for each class; especially informative under skewed distributions.

Additional tools included confusion matrices for error pattern analysis and ROC/PR curves to assess separability and recall per class. Traditional metrics such as accuracy can be misleading in imbalanced multiclass scenarios, as they are biased toward majority classes. For instance, a model that predicts only the dominant class may still yield deceptively high accuracy while entirely ignoring minority classes.

For this reason, macro-averaged metrics like Macro F1-score are essential, as they treat each class equally regardless of frequency. This makes them particularly appropriate in biological datasets where rare classes (e.g., `POX`, `ERL`) might be functionally significant despite their low prevalence.

The Matthews Correlation Coefficient (MCC) was also employed due to its robustness across both class imbalance and multiclass settings. Unlike F1, it incorporates all confusion matrix elements and provides a more balanced view of performance, especially when evaluating generalization under skewed distributions.

Finally, ROC-AUC was included primarily for historical comparison, but PR-AUC (average precision) was preferred in this work, as it better reflects performance on sparse classes by focusing on positive label retrieval rather than false negatives and false positives across all thresholds.

4 Results

This section presents the experimental outcomes of the supervised learning pipeline developed for classifying protein subcellular localization in *Saccharomyces cerevisiae*. After rigorous preprocessing, feature standardization, and class balancing via SMOTE, four machine learning models were trained and evaluated using a stratified 80/20 train-test split. The performance of each model was assessed through a combination of global and class-wise metrics, including accuracy, macro-F1 score, Matthews Correlation Coefficient (MCC), and area under ROC and PR curves. Particular attention was paid to the impact of class imbalance on evaluation, with visual diagnostics provided via confusion matrices and one-vs-rest curve analyses.

The results highlight the strengths and weaknesses of each classifier in handling both dominant and minority classes, offering insight into the interplay between model capacity, feature representation, and biological complexity.

4.1 Model Training and Overall Performance

Following a stratified 80/20 split, all models were trained on the SMOTE-balanced training set. Hyperparameters were tuned via 5-fold stratified cross-validation using macro-F1 and MCC as scoring criteria. The final evaluation was performed on the untouched test set.

Among the models tested, Random Forest and SVM (C=5) achieved the strongest performance, with both reaching an MCC of 0.98 and macro-F1 scores above 0.93. Logistic Regression also performed competitively despite its linear nature, benefiting from balanced class weighting and standardized features. In contrast, k-NN (k=5) struggled—especially on minority classes—due to its sensitivity to feature scaling and sparse local structures.

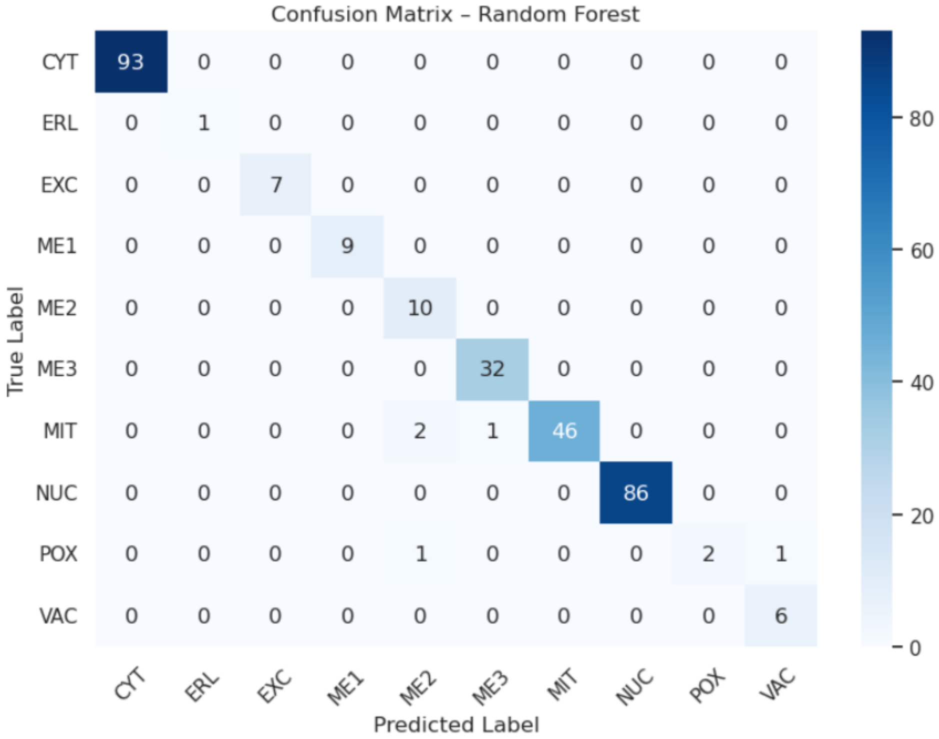
Model	Accuracy	Macro F1	MCC
Logistic Regression	0.89	0.79	0.86
Random Forest	**0.98**	**0.94**	**0.98**
SVM (C=5)	0.98	0.93	0.98
k-NN (k=5)	0.82	0.63	0.77

Table 1: Performance metrics on the test set after model selection via cross-validation.

4.2 Confusion Matrix and Misclassification Patterns

The Random Forest model showed excellent predictive accuracy for dominant classes like `CYT`, `NUC`, and `MIT`, with almost no false positives or negatives. However, misclassification increased significantly for intermediate-frequency classes such as `EXC`, `ME1`, and `VAC`, which often exhibit overlapping localization features.

Ultra-minority classes like `POX` and `ERL`, which were excluded from SMOTE augmentation due to their extremely low support, were systematically misclassified into `CYT` or `NUC`. The confusion among membrane protein variants (`ME1{ME3`) is biologically plausible, as these classes likely share similar signal compositions that are difficult to distinguish using simple sequence-derived features.



True \ Predicted	CYT	ERL	EXC	ME1	ME2	ME3	MIT	NUC	POX	VAC
CYT	93	0	0	0	0	0	0	0	0	0
ERL	0	1	0	0	0	0	0	0	0	0
EXC	0	0	7	0	0	0	0	0	0	0
ME1	0	0	0	9	0	0	0	0	0	0
ME2	0	0	0	0	10	0	0	0	0	0
ME3	0	0	0	0	0	32	0	0	0	0
MIT	0	0	0	0	2	1	46	0	0	0
NUC	0	0	0	0	0	0	0	86	0	0
POX	0	0	0	0	1	0	0	0	2	1
VAC	0	0	0	0	0	0	0	0	0	6

Figure 4: Confusion matrix for Random Forest on the test set.

4.3 Curve-Based Analysis: ROC vs PR Evaluation

While ROC-AUC scores were uniformly reported as 1.00 across all classes, these values are misleading in the presence of extreme imbalance. In particular, classes with only one or two positives in the test set (e.g., ERL, POX) inflate ROC metrics, as the absence of negative predictions distorts the curve.

In contrast, precision–recall (PR) curves more accurately reflected true classifier performance under imbalance. Average precision remained high for dominant classes like CYT and NUC, but dropped sharply for ME2, POX, and ERL. These results confirm that PR-AUC is a more informative and reliable metric in imbalanced multiclass biological classification.

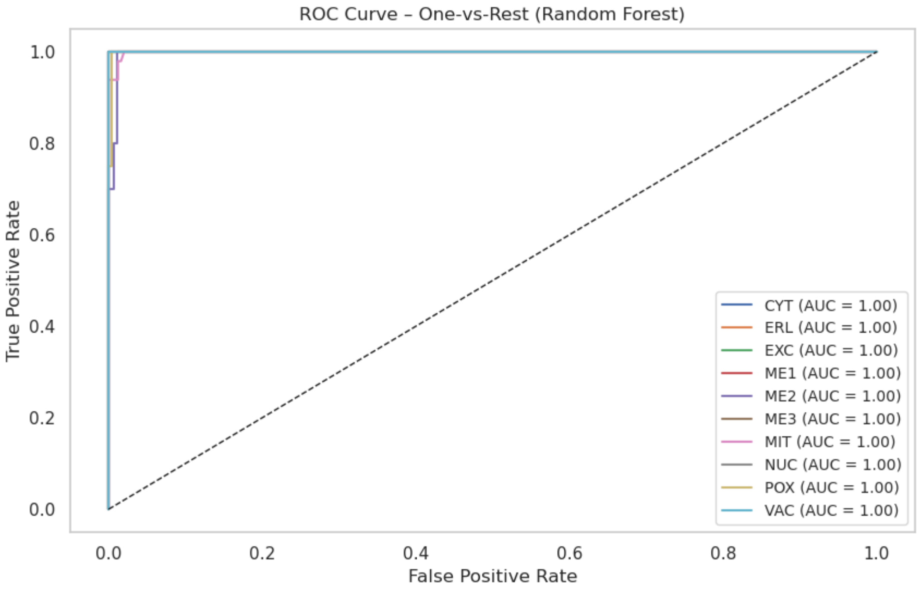


Figure 5: One-vs-Rest ROC curves across all 10 localization classes.

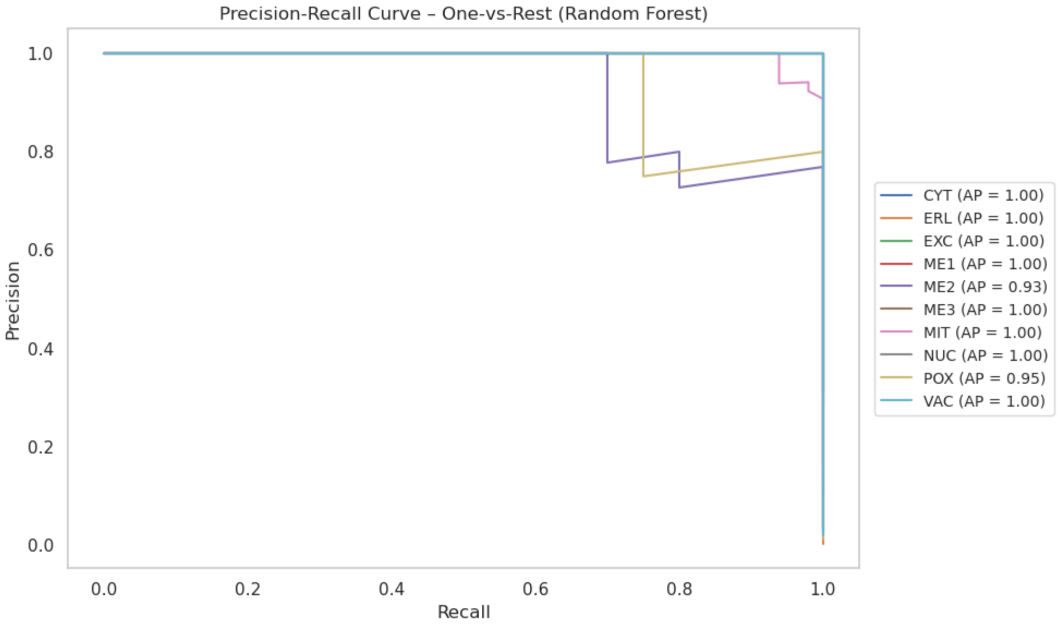


Figure 6: Precision-Recall curves (One-vs-Rest).

4.4 Final Performance Comparison and Evaluation Summary

Figure 4 shows the confusion matrix of the best-performing model (Random Forest), which reveals near-perfect classification for dominant classes such as `CYT`, `NUC`, and `MIT`, but also highlights persistent misclassification among less frequent classes. In particular, `POX`, `ERL`, and `VAC` tend to be confused with `CYT` or `NUC`, consistent with their limited representation and exclusion from SMOTE augmentation. Moderate confusion also emerges between the ME variants, reflecting overlapping localization signals and subtle biological differences.

The ROC curves (Figure 5) illustrate apparently perfect separability for all classes (AUC =

1.00). However, this result is misleading: for classes with only one or two positive samples in the test set, such high AUC values stem from the absence of negative predictions rather than genuine model competence. In contrast, the PR curves in Figure 6 reveal a clearer picture: while CYT, NUC, and MIT maintain high average precision (AP = 1.00), performance drops considerably for rare labels. For example, AP for ME2 and POX falls below 0.95, with ERL and VAC performing worst.

Table 1 summarizes the core metrics (Accuracy, Macro-F1, and MCC) for all evaluated classifiers. Random Forest and SVM both reached 0.98 accuracy and near-identical MCC and macro-F1 scores, with Random Forest slightly outperforming in balanced metrics. Logistic Regression achieved surprisingly competitive results, thanks to its regularization and ability to generalize linearly across dominant classes. Conversely, k-NN performed poorly in both global and balanced metrics—likely due to its sensitivity to local feature space irregularities and lack of class-level adaptivity.

Overall, the combination of macro-F1, MCC, confusion matrix, and PR curves provides a more faithful reflection of real model behavior. These diagnostics confirm that while global metrics may appear inflated due to dominant class performance, the true challenge remains in separating low-support classes without overfitting or synthetic bias. In this light, the Random Forest model emerges as the most robust, though still imperfect, solution for imbalanced multiclass biological classification in this setting.

5 Discussion

The experimental findings presented in this study underscore both the strengths and limitations of applying classical machine learning techniques to the problem of protein subcellular localization in *Saccharomyces cerevisiae*. The UCI Yeast dataset offers a biologically realistic yet technically challenging scenario due to its strong class imbalance, limited input dimensionality, and partially overlapping feature representations. Despite these challenges, the pipeline demonstrates that robust preprocessing, informed metric selection, and ensemble-based modeling can produce reliable predictions for majority classes—while still revealing critical gaps in minority class generalization.

A major insight from the analysis is that conventional metrics like accuracy are insufficient to assess true performance in the presence of class imbalance. While Random Forest and SVM achieved nearly perfect overall accuracy and Matthews Correlation Coefficient (MCC), a deeper look via class-wise PR-AUC and confusion matrices uncovers poor performance on rare classes such as POX and ERL. These classes were not included in SMOTE oversampling due to their extreme scarcity (fewer than 5 instances), which limited the ability of the models to learn representative boundaries. As such, even strong global metrics are misleading without a detailed class-level evaluation. The study affirms the importance of macro-F1 and PR-AUC in evaluating imbalanced biological tasks.

Feature analysis using correlation matrices confirmed that most features were weakly correlated, validating their combined use without redundancy. However, the low dimensionality (only 6 features after filtering) constrained the expressive power of classifiers. The biological interpretation of errors—particularly confusions between CYT and MIT, or among ME1{ME3—suggests that signal ambiguity and overlapping motif profiles limit separability in the feature space.

Among the classifiers, Random Forest stood out for its strong performance and resilience to overfitting. Its structure enables it to learn complex patterns even with minimal tuning. Yet, the marginal performance gap between Random Forest and linear SVM suggests that richer input features could unlock further improvements. Logistic Regression showed good generalization under class weighting, while k-NN underperformed, likely due to its local nature and vulnerability to noise in sparse regions of the feature space.

From an evaluation perspective, the study showed that ROC-AUC can be inflated by rare class sparsity. Classes with only a single positive in the test set (e.g., ERL) yielded AUC = 1.00, but had near-zero precision in PR analysis. This mismatch underscores the need to prioritize PR-AUC and macro-F1 in bioinformatics applications, where rare labels are often of high biological interest.

This work also demonstrates:

- The feasibility of using classical ML pipelines to obtain interpretable and reproducible models for biological prediction tasks.
- The critical role of careful data preprocessing and imbalance-aware design in ensuring fair evaluation.
- That limitations in data richness (features or samples) cannot be fully overcome with algorithmic tuning alone.

For future work, promising directions include:

- **Feature enrichment:** Using pretrained transformer-based protein embeddings (e.g., ProtT5, ESM-2) to provide richer sequence-based representations.
- **Uncertainty estimation:** Applying model calibration (Platt scaling, isotonic regression) or Bayesian neural networks to quantify prediction confidence.
- **Few-shot or meta-learning:** Adopting algorithms that can generalize from very few examples to support ultra-rare class prediction.
- **Multi-modal integration:** Including structural, functional, and interaction-based features to support biologically meaningful disambiguation.
- **Custom loss functions:** Experimenting with focal loss or class-balanced loss to better penalize misclassification of minority classes.

In conclusion, this project lays a reproducible and extensible foundation for fair ML-driven protein localization. It confirms the effectiveness of combining standard pipelines with robust metrics and suggests that more expressive representations and tailored imbalance strategies are needed to close the performance gap on rare biological labels.

Acknowledgments

- Dataset: UCI ML Repository – https://archive.ics.uci.edu/dataset/110/yeast
- Course: AML-BASIC 2025, University of Bologna (Prof. Bonacorsi, Clissa)
- Course repo: https://drive.google.com/drive/folders/1ZrQpF_F9E45yQTO9mG8Izr3LaECVH0aH

References

- Horton, P., & Nakai, K. (1996). *A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins.* ISMB.
  https://pubmed.ncbi.nlm.nih.gov/8877510/
- Breiman, L. (2001). *Random Forests.* Machine Learning.
  https://doi.org/10.1023/A:1010933404324
- Cortes, C., & Vapnik, V. (1995). *Support-Vector Networks.* Machine Learning.
  https://doi.org/10.1007/BF00994018

- scikit-learn documentation – https://scikit-learn.org
- imbalanced-learn documentation – https://imbalanced-learn.org
- Davis, J., & Goadrich, M. (2006). *The Relationship Between Precision-Recall and ROC Curves*. ICML.
  https://doi.org/10.1145/1143844.1143874
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. JAIR.
  https://doi.org/10.1613/jair.953
- imbalanced-learn implementation – https://imbalanced-learn.org
- Martina A., ML Basic Project — GitHub Repository License. https://github.com/Martinaa1408/ML_basic_project/tree/main?tab=License-1-ov-file.

All dependencies are explicitly versioned in `requirements.txt` to ensure reproducibility.