Machine-Learning Classification of three *Aspergillus* Species Using ITS Barcode Data

Student: Lishita Rowjee (1093319)
Course: BINF*6210
Data Sources: BOLD, NCBI

GitHub link: https://github.com/LishitaR/BINF-6210-Assignment-4

INTRODUCTION: Why identify *Aspergillus* species?

Aspergillus species belong to fungi kingdom

- Ubiquitous in the environment
- Morphological identification is unreliable
- Accurate identification needed for research, medicine and biodiversity studies
- Internal transcribed spacer (ITS) widely used for DNA fungal barcoding (Schoch et al., 2012)

Limitations of ITS:

- May not distinguish closely related species (Schoch et al., 2012)
- Public databases (BOLD, NCBI) contain mislabeled/low-quality sequences (Nilsson et al., 2019)

Gap: Will machine learning be able to detect subtle ITS sequence patterns → relevant for species discrimination?

Research question

- Is ITS DNA barcoding from combined BOLD and NCBI datasets sufficient for machine learning algorithms (Random Forest, SVM, XGBoost) to reliably classify closely related *Aspergillus* species, considering database sequence quality and bias?

- Species selected for this project: *Aspergillus fumigatus (A. fumigatus)*, *Aspergillus niger (A. niger)*, *Aspergillus flavus (A. flavus)*

- Reason: These species can exhibit similar ITS sequence patterns, making them challenging to separate using traditional barcoding approaches (Sugita et al., 2004)

Hypotheses

01

H1: ITS 3-mer profiles contain enough variation to allow accurate classification of *A. fumigatus*, *A. niger*, and *A. flavus* using machine-learning models.

02

H2: ITS k-mer profiles of *A. fumigatus*, *A. niger*, and *A. flavus* may overlap, leading to misclassification in PCA and machine-learning models.

03

H3: Sequencing errors, low-quality sequences, and ambiguous bases are present in public ITS datasets, and these quality issues can be detected through exploratory sequence metrics and removed through preprocessing

04

H4: Unequal numbers of sequences per species exist in the dataset, which may influence downstream machine learning analyses.

Machine Learning Approaches

Model	Key feature	Interpretation
Random Forest	Nonlinear ensemble, robust to noise	High accuracy = good feature signal
SVM (Radial)	Boundary-based classifier	Sensitive to overlap between species
XGBoost	Gradient Boosting	Tests model generalization strength

- If Accuracy >80% across all models: H1 supported (ITS discriminates species).
- If Accuracy ~50–70% and confusion matrix shows overlap: H2 supported.

Datasets Used

BOLD

- Downloaded through portal search “Aspergillus [tax]” (on November 16th, 2025)
- Contains sequences + metadata (species, country, marker_code)

NCBI

- Query through Rentrez
- 14,117 hits for ITS-containing *A. fumigatus / niger / flavus* (on November 16th, 2025)
- FASTA downloaded

Key Variables

- Species
- ITS nucleotide sequence
- Sequence length
- GC content
- Ambiguous nucleotides (N-content)

Project Workflow: Methods

- **Download sequences** (BOLD + NCBI)
- **Combine & clean data**

Remove short sequences < 400 bp

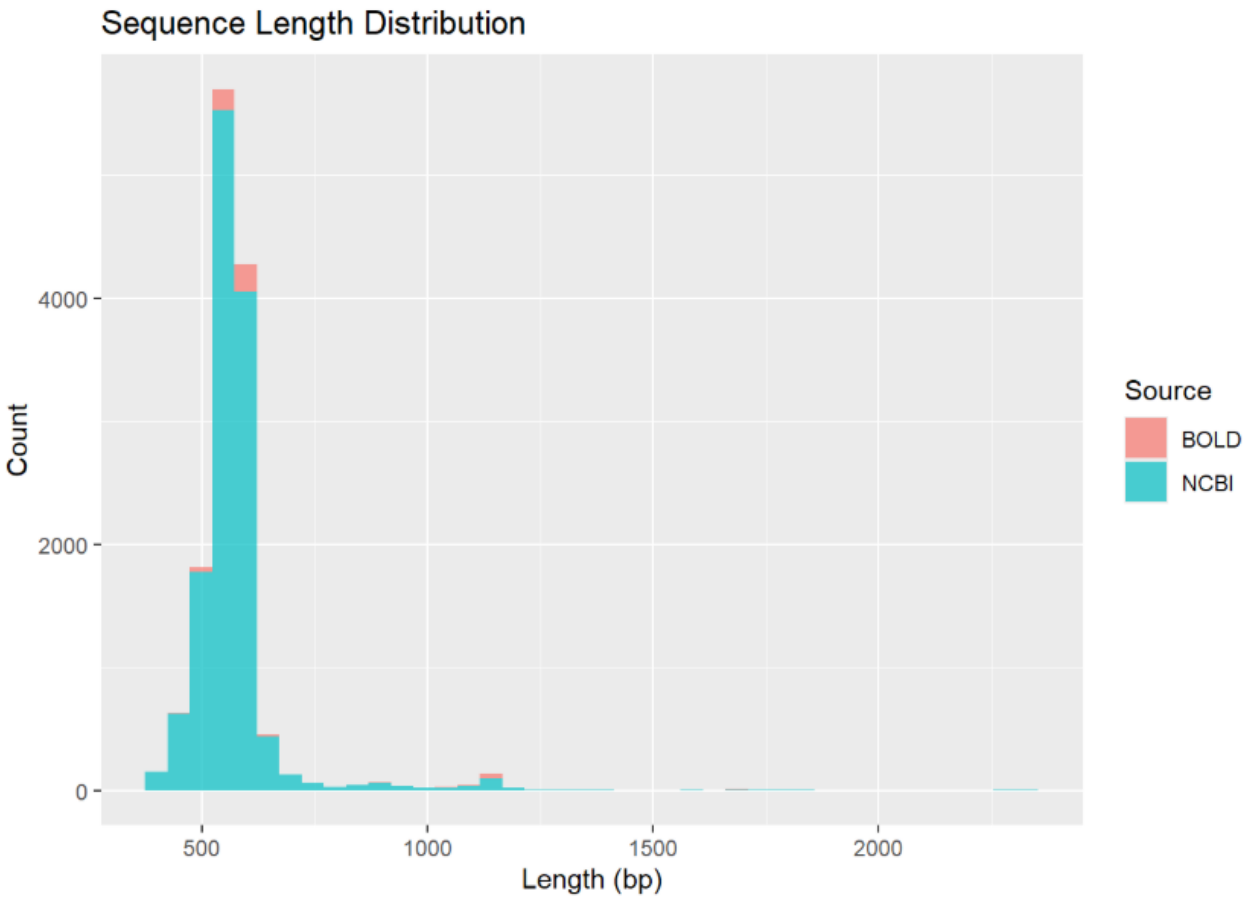
Remove >3% N content

Keep species with ≥200 sequences

- **Compute metrics:** GC, N%, length
- **Encode ITS using k-mers (k=3)**
- **PCA**
- **Train ML models:**
  - Random Forest
  - SVM (RBF)
  - XGBoost
- **Evaluate performance**
- **Interpret results**

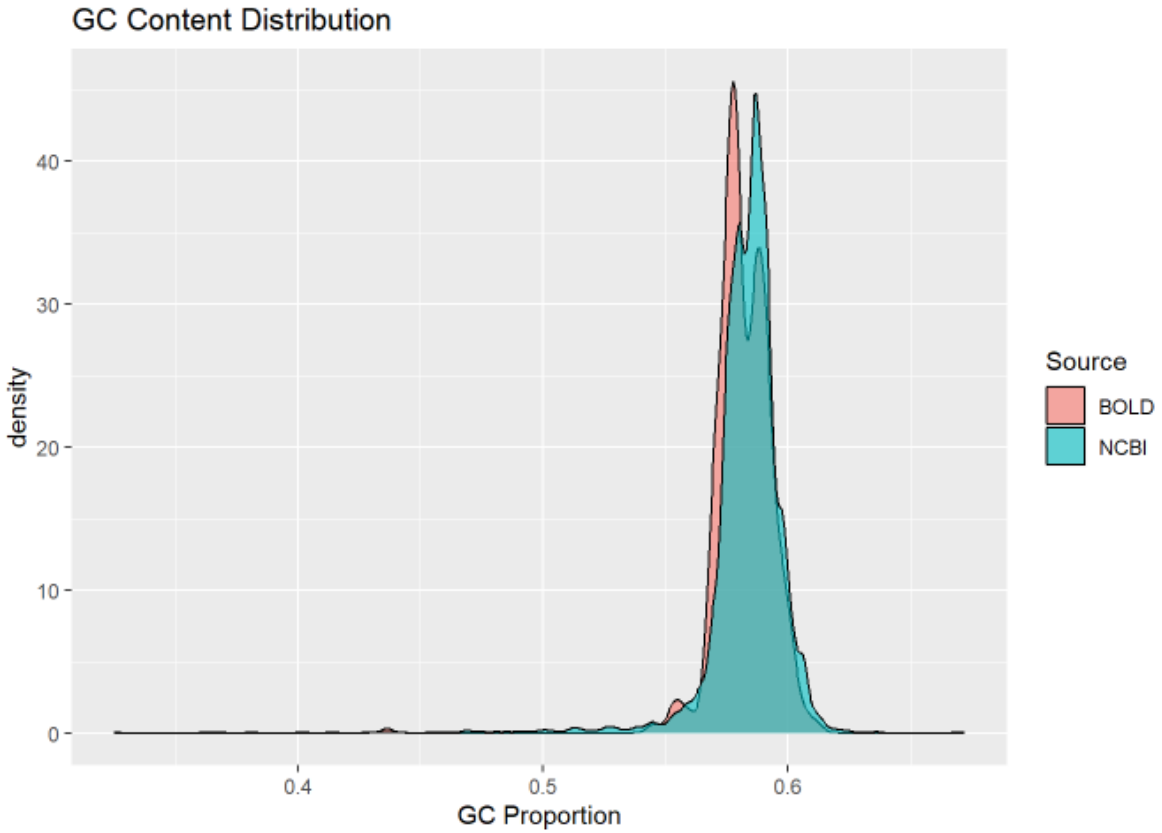
```
## ├── Research question
##     ├── Hypotheses
## ├── 1. Setup: Install and Load packages
## ├── List of required packages
## ├── Function to install (if needed) and load a single package
## ├── Apply the function to all packages using lapply
## ├── 2. Data acquisition and loading
##     ├── 2.1 BOLD database:
##     ├── 2.2 NCBI
## ├── 3. BOLD and NCBI Data wrangling
##     ├── 3.1 Bold data:
##     ├── 3.2 NCBI Fasta Sequences
##     ├── 3.3 Combining BOLD and NCBI dataframe
## ├── Check combined size
## ├── 4. Data Cleaning for H3
## ├── 5. Inspect cleaned data set (H4)
## ├── 6. Exploration figures for H3
##     ├── 6.1 Sequence Length Histogram
##     ├── 6.2 GC content
## ├── 7. k-mer Encoding (for PCA and Machine Learning)
## ├── 8. PCA (for H1 and H2)
## ├── 9. Machine Learning Models (Test H1-H3)
##     ├── 9.1 Splitting data
##     ├── 9.2 Random Forest
##         ├── Training
##         ├── Prediction
##         ├── Confusion matrix
##         ├── Importance
##     ├── 9.3 SVM model
##         ├── Training
##         ├── Prediction
##         ├── Confusion matrix
##     ├── 9.4 XG Boost Model
##         ├── Training
##             ├── Convert labels to integers
##             ├── Train multi-class XGBoost
##         ├── Prediction
##         ├── Map back to original species names
##         ├── Confusion Matrix
## ├── 10. Summary Table (randomForest, SVM and XGBoost)
## ├── Ensure all three confusion matrices exist
## ├── XGBoost confusion matrix
## ├── summary table
## ├── 11. Generate Pipeline
```

Figure 1: Cleaned ITS sequences show broad but comparable length distributions



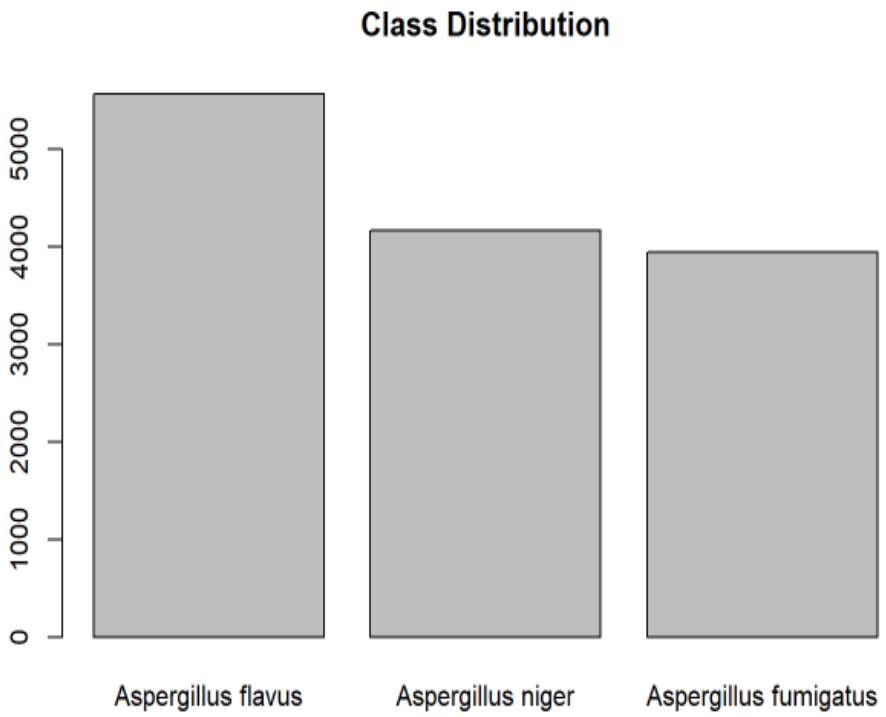
- Histogram displays the lengths of all ITS sequences after the cleaning step (Length ≥ 400 bp) and compares two data sources: BOLD (orange) and NCBI (teal)
- Histogram shows that almost all ITS sequences fall within the expected 450–650 bp range, confirming that the cleaned dataset contains biologically realistic ITS fragments.
- NCBI contributes most sequences, while BOLD provides a very small proportion.
- Both sources share nearly identical length distributions, indicating no length-based bias between databases.
- This supports **H3**: data cleaning improves consistency.
- Only a small number of long sequences (>800 bp) are present, likely reflecting ITS regions with flanking rRNA segments.

Figure 2: Variation of GC content between databases could indicate technical biases



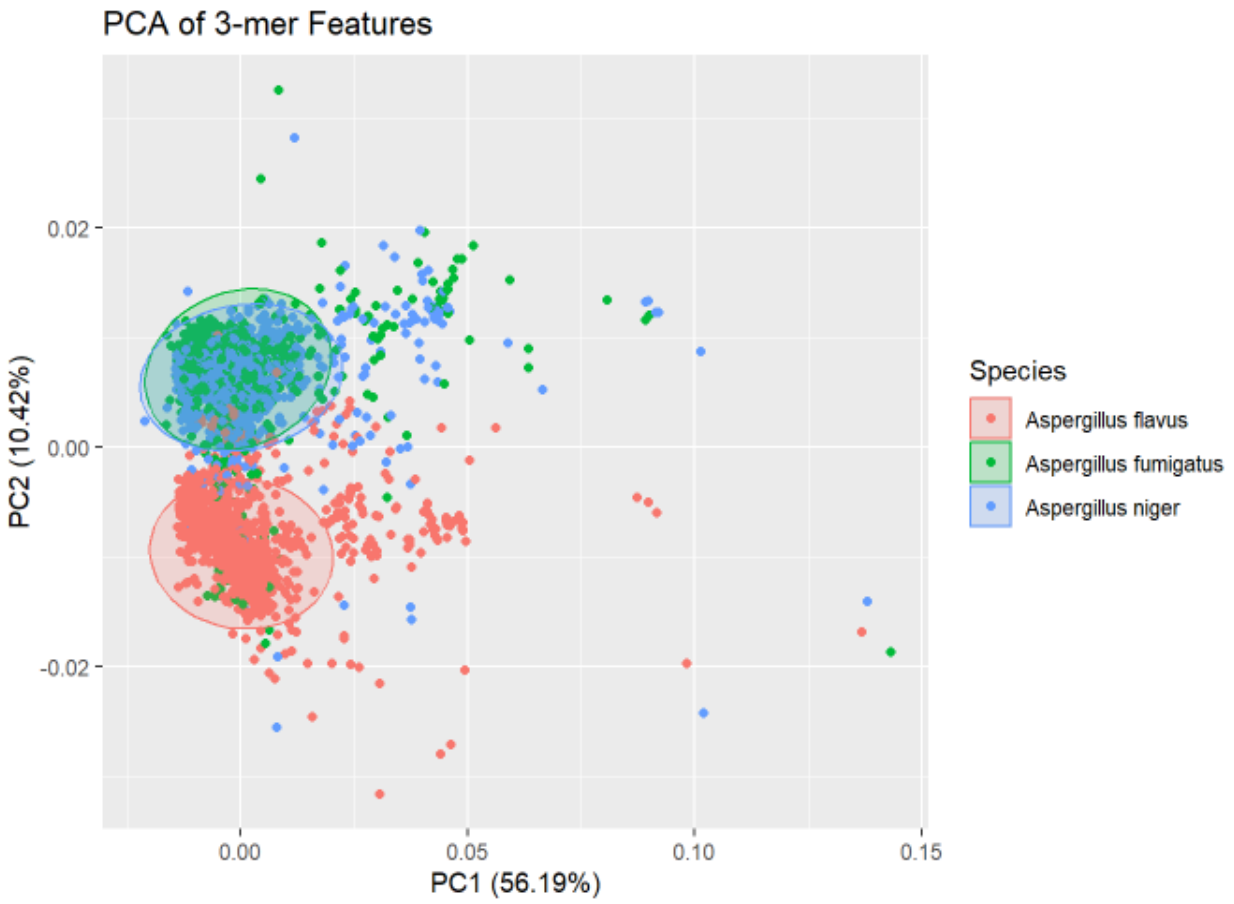
- Interpretation for H3: If one source (e.g., BOLD) has a wider GC spread or more outliers, that could degrade ML performance.
- Small shoulder or low-GC peak might suggest possible contaminants or misidentified sequences.
- Both sources peak around 0.58–0.6 GC, showing consistent biology.

Figure 3: Strong class imbalance in cleaned dataset (H4 confirmed)



- Species counts differ - may lead to Machine learning bias toward majority class.
- Needs to be careful at interpreting accuracy.
- Consider looking at kappa statistic.

Figure 4: PCA – 3 mers shows partial clustering but overlapping profiles among species (H1 + H2).



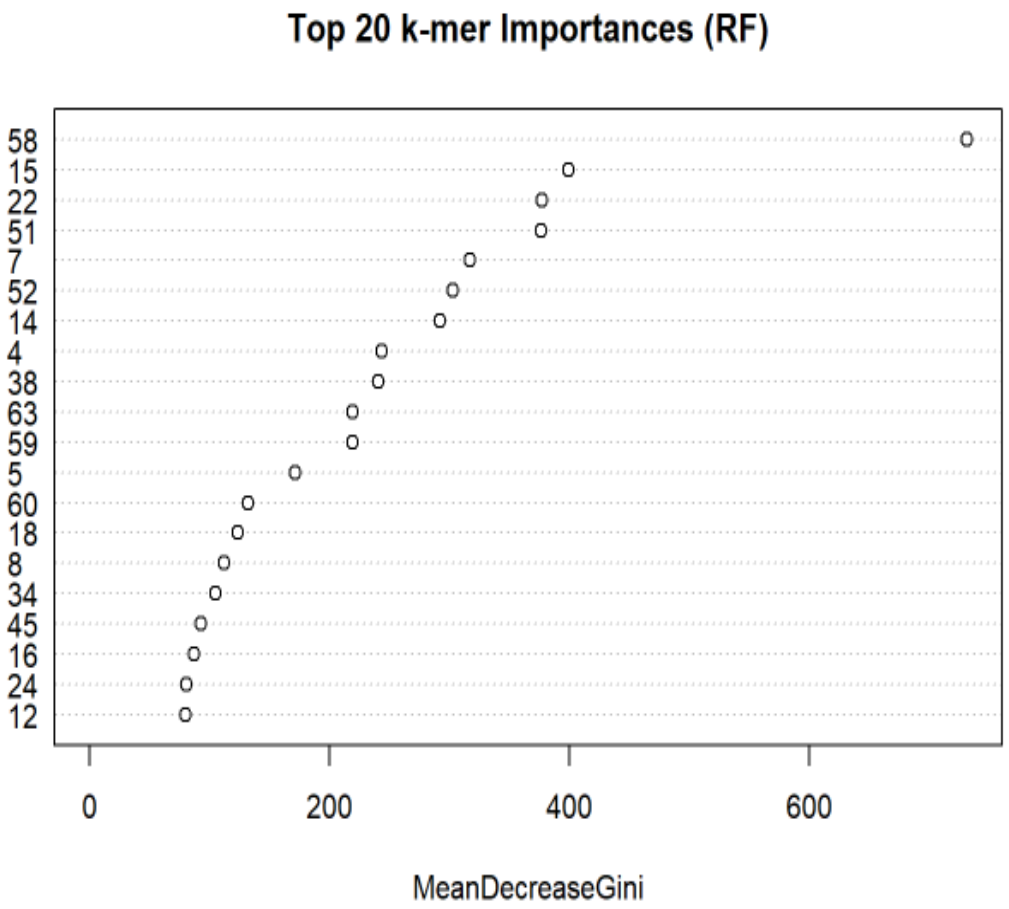
Separation of Species

- *A. flavus* (red) forms a somewhat distinct cluster on the left side, showing some separation along PC1.
- *A. fumigatus* (green) and *A. niger* (blue) overlap substantially, especially in the left central area.
- Overlap between *A. fumigatus* and *A. niger* indicates that these two species share similar 3-mer profiles, at least along the first two principal components.

Cluster Tightness

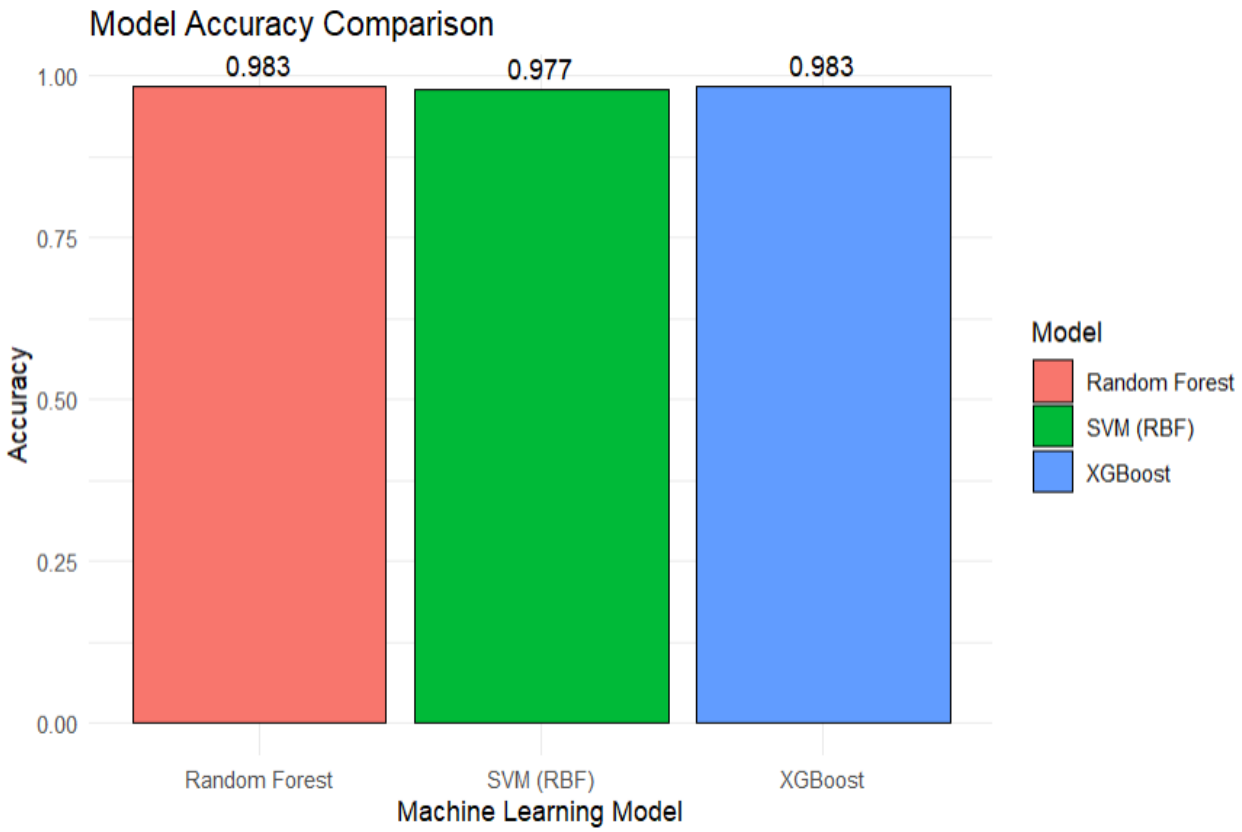
- The *A. flavus* cluster looks tighter and more compact, indicating more consistent k-mer profiles across sequences.
- *A. fumigatus* and *A. niger* clusters are more dispersed and overlap, indicating more variation within these groups and less clear distinction between them.

Figure 5: Top 20 Most Important 3-mers for Aspergillus Species Classification (Random Forest Mean Decrease Gini)



- Rightmost dot (highest MeanDecreaseGini) corresponds to most important 3-mer in classifying Aspergillus species.
- Rest of the points descend leftwards in importance.
- Suggests that certain 3-mers are much more informative than others in separating your species
- Since the k-mers are derived from ITS sequences, the top k-mers likely capture key sequence variations that help differentiate between *A. fumigatus*, *A. niger*, and *A. flavus*.

Figure 6: Random Forest + SVM + XGBoost Performance Comparison



- SVM and Random Forest outperform SVM on overlapping ITS profiles.

	Model	Accuracy	Kappa
1	Random Forest	0.9831625	0.9744050
2	SVM (RBF)	0.9773060	0.9655231
3	XGBoost	0.9827965	0.9738424

Results Summary

Models performed extremely well, showing ITS 3-mer profiles contain enough discriminatory signal:

Random Forest

Very balanced sensitivity/specificity across species

Few misclassifications, mostly *A. flavus* ↔ *A. niger*

SVM (RBF kernel)

Slightly lower accuracy than RF but still excellent

Misclassification patterns like RF

XGBoost

Performance nearly identical to Random Forest

Most errors were *A. flavus* predicted as *A. niger*

Interpretation

All ML models consistently achieve 97–98% accuracy, strongly supporting H1 (ITS k-mers are sufficient for species-level classification)

Slight overlap and misclassification patterns align with H2

Discussion: Interpreting Findings Using Literature

- ITS 3-mers can classify these *Aspergillus* species accurately → supports H1
- PCA plot and confusion matrices show some overlap are likely to happen between *A. niger* & *A. flavus* → supports H2
- NCBI + BOLD would show biases, errors, ambiguous bases if not cleaned → supports H3
- Species counts are uneven, and it could potentially affect model stability → supports H4

Why does ITS sometimes fail to distinguish Aspergillus?

- ITS barcode recognized globally (Schoch et al., 2012), but not ideal for all clades.
- Nilsson et al. (2019) noted high rates of mislabeled public fungal ITS sequences.
- PCA and Machine learning results reflect some of these limitations.
- GC content variation suggests platform-level biases.

Study Caveats

- Database mislabeled sequences may remain even after filtering (e.g. there were some sequences that were >800 bp)
- Only ITS used, additional loci such as β-tubulin, CaM) could improve accuracy (Balajee et al., 2007)
- Class imbalance affects model performance
- K=3 k-mers may have insufficient resolution (He et al., 2025)

Next Steps / Future Directions

- Run the models again with uncleaned data and compare the accuracy results with the models ran on cleaned data to confirm whether sequence biases affect machine learning
- To reduce computation time, I used 3-mer frequencies and ntree=300, but we can run 4-mer frequencies + ntree>300 for Random Forest
- Conduct phylogenetic validation of each Machine learning predictions
- Use additional loci to reduce risk of overfitting to single loci
- Multi-locus k-mer profiles capture more of the genome's evolutionary history (Ametrano et al., 2022).

Reflection

- Through the BINF*6210 classes and assignments , I was able to learn the full bioinformatics pipeline from data acquisition to Quality Control to Machine learning.
- Additionally, I was able to improve my debugging skills, handling missing data, and reproducibility throughout the semester.
- I definitely gained confidence in R workflows, classifiers, and interpreting outputs now compared to back in September when I was a beginner at RStudio .
- I believe that these skills will be directly applicable to me when it comes to the analysis of my project for my thesis.
- The skills learnt will also help me in my career after graduate school since machine learning is evolving at a high rate.

Acknowledgements

- I would like to acknowledge Dr Karl Cottenie for the R script provided on Random Forest classifier as it helped me a lot for the Random Forest part of my code.

- Building on the knowledge provided by my group members from Assignment 2, I was able to learn how to load my packages more effectively.

- I would also like to acknowledge fellow classmate Hannah Glowacki for looking over my storyboard and offering me editing advice on my slides.

- Lastly, I would like to acknowledge our TA Saira Asif for directing me to resources for XGBoost and SVM models. She also helped me iron out my research question so that it does not become too overwhelming.

References

Scientific papers

- Ametrano, C. G., Jensen, J., Lumbsch, H. T., & Grewe, F. (2025). UnFATE: A Comprehensive Probe Set and Bioinformatics Pipeline for Phylogeny Reconstruction and Multilocus Barcoding of Filamentous Ascomycetes (Ascomycota, Pezizomycotina). *Systematic Biology*, syaf011. https://doi.org/10.1093/sysbio/syaf011
- Balajee, S. A., Houbraken, J., Verweij, P. E., Hong, S.-B., Yaghuchi, T., Varga, J., & Samson, R. A. (2007). Aspergillus species identification in the clinical setting. *Studies in Mycology*, *59*, 39–46. https://doi.org/10.3114/sim.2007.59.05
- He, L., Huang, M., Yiming, G., Zhu, Y., Liu, R., Chen, J., & Yau, S. S. T. (2025). A new alignment-free method: K-mer Subsequence Natural Vector (K-mer SNV) for classification of fungi. *BMC Bioinformatics*, *26*(1), 170. https://doi.org/10.1186/s12859-025-06152-x
- Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2019). Mycobiome diversity: High-throughput sequencing and identification of fungi. *Nature Reviews Microbiology*, *17*(2), 95–109. https://doi.org/10.1038/s41579-018-0116-y
- Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding Consortium, Fungal Barcoding Consortium Author List, Bolchacova, E., Voigt, K., Crous, P. W., Miller, A. N., Wingfield, M. J., Aime, M. C., An, K.-D., Bai, F.-Y., Barreto, R. W., Begerow, D., … Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*. *Proceedings of the National Academy of Sciences*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109
- Sugita, C., Makimura, K., Uchida, K., Yamaguchi, H., & Nagai, A. (2004). PCR identification system for the genus Aspergillus and three major pathogenic species: Aspergillus fumigatus, Aspergillus flavus and Aspergillus niger. *Medical Mycology*, *42*(5), 433–437. https://doi.org/10.1080/13693780310001656786

References

Websites

- https://www.datacamp.com/tutorial/support-vector-machines-r
- https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/SVM
- https://www.rdocumentation.org/packages/e1071/versions/1.7-16/topics/svm
- https://stackoverflow.com/questions/37190135/how-can-i-speed-up-the-training-of-my-random-forest
- https://www.numberanalytics.com/blog/optimizing-random-forests-ml-performance
- https://developer.ibm.com/tutorials/awb-implement-xgboost-in-r/
- https://medium.com/@heyamit10/xgboost-in-r-a-practical-guide-f14b722866c1
- https://rpubs.com/nicklepore/XGBoost
- https://search.r-project.org/CRAN/refmans/directotree/html/description.html

References

YouTube videos

- https://www.youtube.com/watch?v=OJcFCs7Toe4 (XGBoost tutorial)
- https://www.youtube.com/watch?v=gKyUucJwD8U (XGBoost tutorial)
- https://www.youtube.com/watch?v=QkAmOb1AMrY (SVM tutorial)
- https://www.youtube.com/watch?v=ueKqDlMxueE (SVM tutorial)
- https://www.youtube.com/watch?v=6EXPYzbfLCE (Random Forest tutorial)
- https://www.youtube.com/watch?v=zk017BiA-ZE (Confusion matrices tutorial)
- https://www.youtube.com/watch?v=5vgP05YpKdE (PCA plot analysis)
- https://www.youtube.com/watch?v=NLrb41ls4qo (PCA plots in R)
- https://www.youtube.com/watch?v=u4IxOk2ijSs (Random Forest Mean Decrease Gini analysis tutorial)