Species Delimitation Using Machine Learning

Philip Lin and Bing Li

University of Wisconsin Madison
Department of Biostatistics and Medical Informatics

April 25, 2024

Why It's Interesting

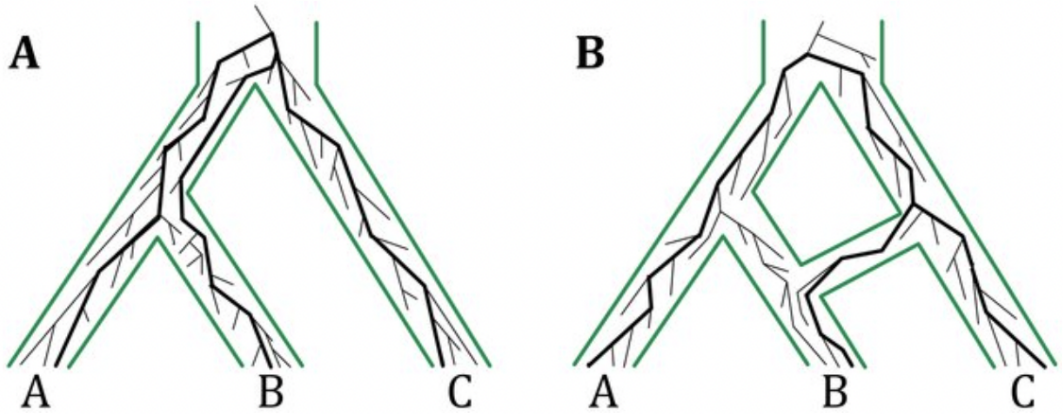
- **The Problem:** Traditional species delimitation relies heavily on morphological traits.
- **Importance for Science and Conservation**

Introduction to Species Delimitation Challenges

- **Research Question:** Can machine learning algorithms such as XGBoost and MLP improve the accuracy and reliability of species delimitation within the Bromeliaceae family?
- **Significance of Machine Learning:** Machine learning offers innovative tools to address these limitations, potentially revolutionizing species identification and classification.

Existing Methods for Species Delimitation

- **Traditional Morphology-Based Methods:** Historically, species were identified and delimited based primarily on physical traits. However, these methods can be subjective and overlook genetic diversity.
- **Molecular Genomics Approaches:** Recent years have seen the development of models such as the multispecies coalescent model using genomics data.



Limitations of Current Methods

- **Computational Demands:** Advanced statistical models, while powerful, often demand substantial computational resources and expert knowledge, limiting their accessibility and scalability.
- **Need for Refined Data:** Many methods require high-quality, comprehensive genetic data, which is not always available, especially for non-model organisms.

Why Machine Learning?

- **Overcoming Limitations:** Machine learning algorithms, such as XGBoost and MLP, can handle large datasets with complex patterns, potentially providing more accurate species delimitation with less human bias.
- **Innovative Potential:** These methods have rarely been applied to non-model organism species delimitation.
- **Expected Benefits:** Improved accuracy and reliability in species identification, efficient processing of large genomic datasets, and enhanced capability to uncover hidden genetic relationships.

Data Overview

- **Organisms Studied:** Bromeliaceae family, including eight species across five genera, crucial for their ecological and commercial value (e.g., pineapple).
- **Data Source:** Genomic data obtained via low-coverage genome skimming provided by Novogene Inc.
- **Data Structure:** Genetic data stored in Variant Call Format (VCF), which includes SNP variants detailed in GT:PL format.
- **GT:PL Format Explained:**
  - **GT (Genotype)**: Indicates the genetic constitution at specific loci.
  - **PL (Phred-scaled Likelihoods)**: Provides a quantitative measure of the accuracy of the genotype call.

Sample Sizes and Data Processing

- **Sample Sizes:** Analysis includes 210,859 SNP records across the eight species, providing a comprehensive genetic overview.
- **Data Processing:**
  - **Preprocessing:** Raw sequencing reads were trimmed using tools like fastp for quality control, removing adapters and low-quality reads.
  - **Alignment and Variant Calling:** Reads were aligned to the reference genome of *Ananas comosus* using Bowtie2, with variants called using BCFtools.
  - **Quality Checks:** Data quality validated through BUSCO for capturing universal single-copy orthologs and BLAST+ for further verification of single-copy orthologs.
- **Dataframe Structure:** The final processed dataset is structured into a dataframe with dimensions (210,859 rows $\times$ 33 features), facilitating data manipulation and analysis.

Machine Learning Models for Species Delimitation

- **Choice of Algorithms:** Focused on XGBoost and Multi-Layer Perceptron (MLP) for their robustness in handling complex datasets.
- **XGBoost:** An ensemble learning method that uses gradient boosting frameworks, known for its high performance on structured data.
- **MLP:** A type of neural network suitable for capturing nonlinear relationships in data through multiple layers of neurons.

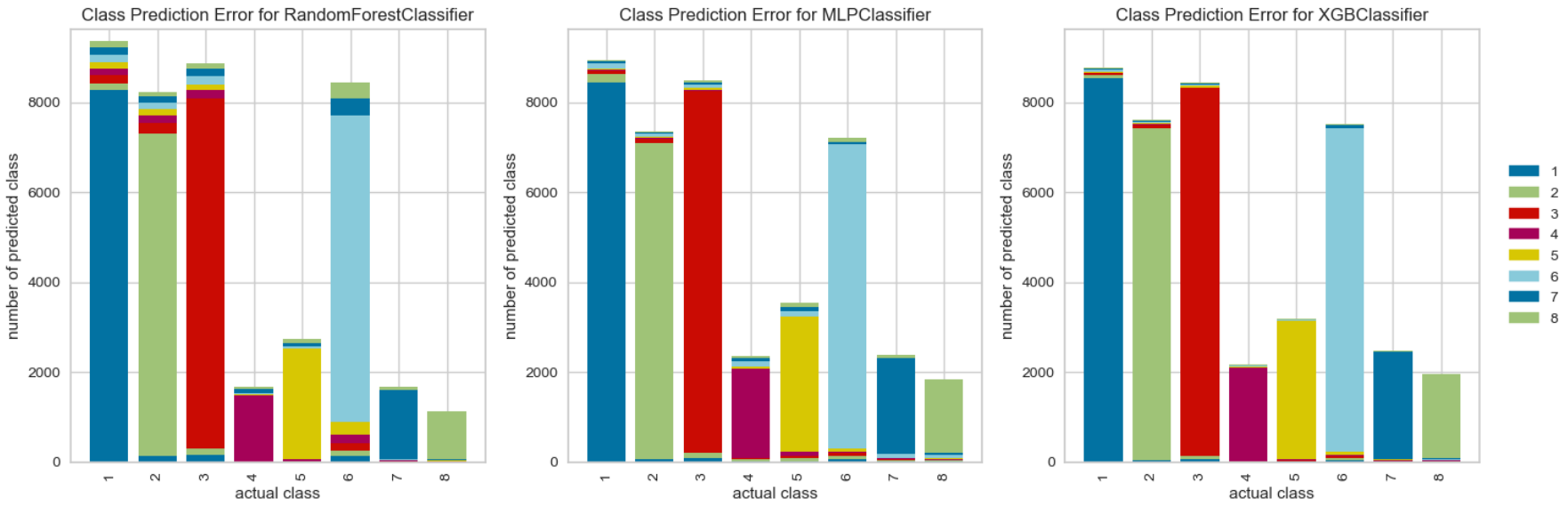
Data Preparation and Model Configuration

- **Data Transformation:** SNP data was transformed into a matrix format suitable for analysis with both XGBoost and MLP.
- **Feature Selection:** Important genetic markers were identified using statistical techniques to enhance model performance and relevance.
- **Model Parameters:**
  - **XGBoost:** Configured with parameters optimized for depth, learning rate, and number of trees to manage overfitting and ensure robust generalization.
  - **MLP:** The architecture was carefully tuned, including the number of hidden layers and neurons, to balance between complexity and computational efficiency.
- **Validation Technique:** Employed the hold-out method, reserving 20% of the data for testing to assess the model's performance on unseen data.

Evaluation Metrics and Model Validation

- **Evaluation Metrics:** Focused on precision, recall, and the F1 score to measure the balance between model accuracy and its ability to generalize.
- **Model Testing:**
  - **Precision:** Proportion of true positives among the labels classified as positive.
  - **Recall:** Ability of the model to identify all relevant instances (true positives).
  - **F1 Score:** Harmonic mean of precision and recall, providing a single score that balances both the precision and the recall.
- **Cross-Validation:** Used to ensure that the model is not tailored to a specific subset of the data, thereby enhancing its robustness and reliability across different sets of data.

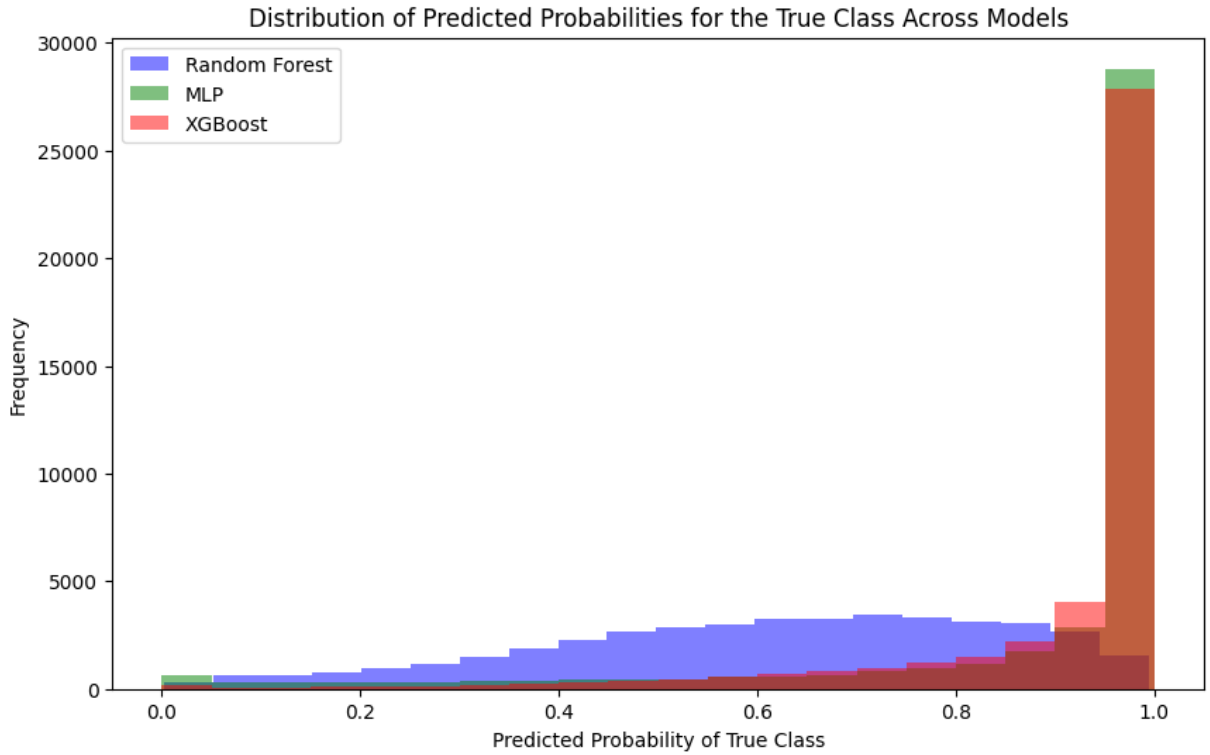
Class Prediction Error



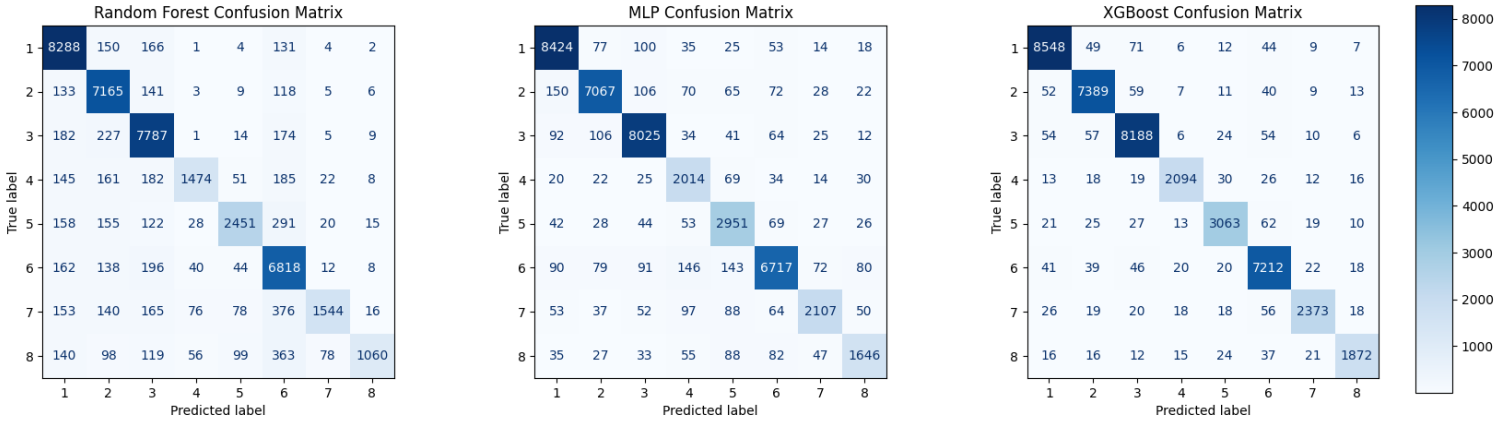
- Each bar represents the count of misclassified instances for an actual class.
- Colors denote the predicted class mistaken for the actual class.
- Inference: The RandomForestClassifier has the most diverse color distribution among the actual classes, suggesting higher misclassification compared to the MLPClassifier. The XGBoost model displays the least color variation.

Predicted Probability Distribution

- Horizontal axis represents the predicted probability of the true class.
- Vertical axis indicates the frequency of instances in the dataset.
- Inference: XGBoost shows a higher concentration of predictions with high confidence (probabilities close to 1), suggesting it is the most confident model.



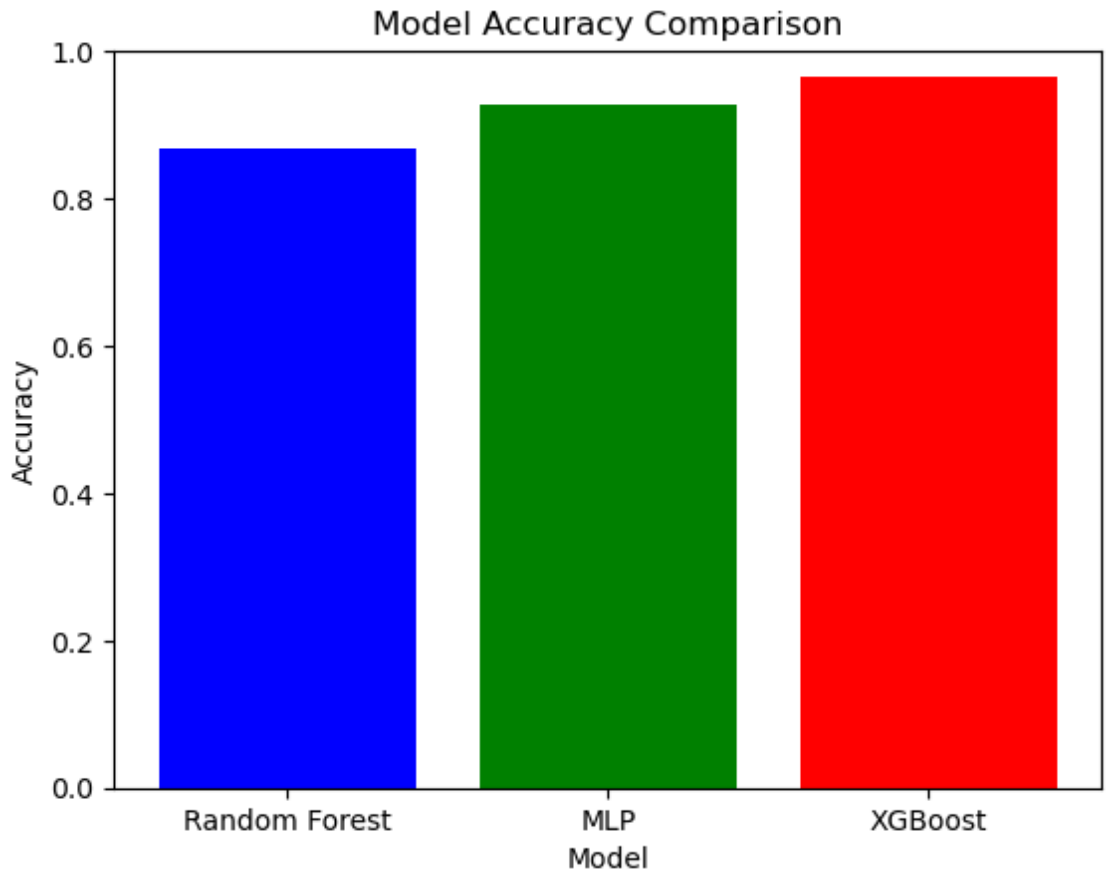
Confusion Matrix for Classifiers



- Matrices show the count of predictions for each true label vs. predicted label.
- Ideal results would show high values along the diagonal (correct classifications).
- Inference: The XGBoost classifier demonstrates a stronger diagonal, suggesting a higher overall classification accuracy.

Model Accuracy Comparison

- Each bar represents the accuracy of a model.
- Inference: The XGBoost model outperforms Random Forest and MLP in terms of accuracy, aligning with expectations from its higher confidence in predictions.



Interpretation of Results

- **Insights Gained:** The application of machine learning models has demonstrated a clear advantage in distinguishing between species within the Bromeliaceae family.
- **Model Efficacy:** XGBoost's performance indicates its capability to handle complex genetic patterns and offers a promising tool for species delimitation in taxonomy.
- **Challenges Addressed:** The use of machine learning algorithms has helped overcome the ambiguity often encountered in species delimitation due to overlapping genetic signatures.
- **Implications for Taxonomy:** These findings suggest a shift towards integrating computational methods in taxonomic practices could greatly enhance accuracy and efficiency.

Broader Implications and Future Work

- **Beyond Bromeliaceae:** The methodologies could be applied to other families and groups, potentially reshaping species classification across a wider biological spectrum.
- **Conservation Efforts:** Accurate species identification is crucial for conservation planning, and these methods could improve strategies for preserving biodiversity.
- **Next Steps:** Future work will focus on refining the models, incorporating larger datasets, and exploring additional machine learning techniques to further improve the accuracy and reliability of species delimitation.
- **Integration with Traditional Methods:** Combining machine learning with traditional taxonomic methods could yield a powerful hybrid approach, leveraging the strengths of both disciplines.

Citations

[1] Wen, Dingqiao & Yu, Yun & Nakhleh, Luay. (2016). Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. PLOS Genetics. 12. e1006006. 10.1371/journal.pgen.1006006.
[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754
[3] Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, DC: Spartan Books.
[4] Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, 1, e323. doi: 10.1002/cpz1.323
[5] Camacho, C., Coulouris, G., Avagyan, V. *et al.* (2009) BLAST+: architecture and applications. **BMC Bioinformatics 10**, 421 (2009).
[6] Langmead B and Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.