ARTICLE (*Pre-print*)

Species Delimitation Using Machine Learning Methods

Philip Lin* and Bing Li

University of Wisconsin - Madison, Madison, US

*Corresponding author: pjlin2001@gmail.com

Abstract

This study examines species delimitation within the Bromeliaceae family using machine learning algorithms to analyze single nucleotide polymorphism (SNP) variants. Traditional species identification methods often rely on morphological traits, which may not capture the full extent of genetic diversity. This project employs two machine learning approaches, XGBoost and Multi-Layer Perceptron, to examine species delimitation. Our analysis includes eight species across five genera, sequenced by low-coverage genome skimming. The results highlight the initial steps towards integrating machine learning models into species delimitation efforts, demonstrating promising potentials of further exploration in taxonomy and conservation biology using machine learning classification methods.

Introduction

This study addresses species delimitation problem by using species-level plant genomic data. Species delimitation plays a central issue that touches on multiple scientific disciplines, including systematics, conservation biology, and evolutionary biology. Recent approaches using genomic data to address species delimitation focuses on developing statistical approaches and incorporating more large datasets. With the recent popularity in machine learning and its application in clustering, it is surprising that machine learning methods have been rarely applied to study taxonomic delimitation, especially in plants. This project aims to leverage the power of single nucleotide polymorphism (SNP) variants to investigate the genetic structure underlying species differentiation across five genera within the Bromeliaceae family, sequenced by low-coverage genome skimming. Bromeliaceae is important not only due to its commercial crops, pineapple, but also as one of the most diverse flowering plant families, contributing greatly to the rich biodiversity of the Neotropics, an understudied global biodiversity hotspot. By focusing on the genomic level, we hope to offer a more accurate method for species delimitation, contributing valuable insights to the fields of taxonomy and conservation.

Motivation

The primary motivation behind this study is to evaluate two machine learning methods, extreme gradient tree boosting approach (XGBoost) ([1] Chen and Guestrin 2016) and Multi-Layer Perceptron (MLP) ([2]Rosenblatt 1962), for identifying species boundaries using plant genomic data derived from single nucleotide polymorphism (SNP) variants. This research is particularly relevant for taxonomists, conservation biologists, and evolutionary biologists, who need reliable methods for species identification to conduct biodiversity assessments, conservation planning, and evolutionary studies.

Traditional species delimitation methods, primarily based on morphological characteristics, often fail to capture the subtle genetic variations that can be crucial for accurate species identification. These methods, while valuable, may not adequately reflect the genetic diversity within and between species, leading to potential misidentification and underestimation of biodiversity ([3]Duminil and Di Michele 2009). The integration of genomic data into species delimitation, facilitated by advancements in sequencing technologies, offers a more detailed resolution of genetic differences. Machine learning

methods, such as XGBoost and MLP, are equipped to handle high-dimensional dataset, making them suitable for analyzing complex genomic datasets to discern species boundaries.

For conservation biologists, accurate species identification is critical for assessing the status of species, especially those that are cryptic or endangered. Misidentification can lead to inappropriate conservation priorities and ineffective resource allocation. For evolutionary biologists, understanding the genetic basis of species delimitation is essential for studying speciation processes, adaptive traits, and phylogenetics. For taxonomists, the integration of machine learning with genomic data represents a transformative approach to taxonomy by providing a more systematic revisions across not only the Bromeliaceae family but potentially various plant and non-plant families. Therefore, this study is driven by the need to address the gaps in traditional species delimitation methods and to harness the potential of machine learning in facilitating biodiversity research.

Background

Species is the fundamental unit in the study of biodiversity and evolution, holding remarkable importance in conservation biology and evolutionary studies. Distinguishing and classifying species, however, presents significant challenges. Historically, the identification of species largely relies on morphological traits (e.g. [4]Pedersen 2010), which may not fully reflect the extent of genetic variation within and between species. Recent breakthroughs in statistical methods of genomic studies and next generation sequencing technologies have facilitated a range of phylogenetic and taxonomic delimitation studies using different types of molecular data, such as whole-genome resequencing (e.g., [5]Liang et al., 2019; [6]Ma et al., 2020), single or low-copy nuclear genes or ultraconserved loci (e.g., [7]Johnson et al., 2019; [8]Yardeni et al., 2022, [9]Heidin et al. 2019), extensive transcriptome sequencing (e.g., [6]Ma et al., 2020), and large-scale SNP data (eg. [10]Adam et al. 2014), etc. Moreover, species delimitation research is increasingly incorporating sophisticated statistical approaches, such as utilizing multivariate algorithms to interpret morphological data ([11]Ezard et al. 2010) and advancing molecular diagnostics with multi-locus sequencing data ([12]Yang and Rannala, 2010).

The major difficulty of addressing species delimitation is to detect the genetic differentiation in population-level and species-level diversity. Speciation is a gradual and continuous evolutionary process with species being represented along a continuum. This results in blurry species boundaries, particularly considering highly differentiated population structure associated with limited gene flow and geographic barriers. For instance, a plant species with a disjointed distribution that experiences limited gene flow between populations tends to exhibit different genetic structure. This can lead to confusion, mistaking population structure for species boundaries. The current most popular statistical model for delimiting species boundary, multispecies coalescent model (MSCM) on Bayesian estimation ([13]Rannala and Yang 2003; [12]Yang and Rannala 2010), accommodates both pre-speciation mutations and incomplete lineage sorting. This approach helps to resolve discrepancies between actual species relationships and those inferred from multi-locus gene trees ([13]Rannala and Yang 2003; [12]Yang and Rannala 2010; [10]Adams et al. 2014). However, tree-based approaches like MSCM often assume speciation to be an instantaneous point event rather than a process, potentially inflating species counts due to the genetic differentiation arising from population structuring ([14]Sukumaran and Knowles, 2017; [15]Yang et al., 2019; [16]Smith and Carstens, 2018).

Machine learning methods are increasingly applied to identify patterns in evolution and biology. Historically, such methods in plant research have focused on detecting phenotypic traits through image processing in model species, like Arabidopsis ([17]Ma et al. 2014; [18]Singh et al. 2016). However, the use of machine learning to interpret plant genomic data, particularly for species delimitation in non-horticultural species, is rare. Techniques like Principal Component Analysis (PCA) have proven successful to clarify species boundaries and population structures using morphological and genomic data ([4]Pedersen 2010; [15]Yang et al. 2019; [19]Cheng et al. 2021). Support vector

machines (SVMs) have been developed to optimize likelihood scores for species-population assignments ([20]Pei et al. 2018). Smith and Carstens ([16]2018) have employed random forests (RF) to assess various speciation models incorporating demographic processes. Additionally, unsupervised machine learning techniques, which do not require predefined species labels, have been adopted to explore genetic structures associated with species delimitation ([21]Derkarabetian et al. 2019).

Species delimitation using machine learning techniques is fundamentally a clustering challenge. It involves identifying and grouping patterns present in a matrix dataset, where the goal is to cluster these patterns based on their similarity across samples. This clustering is not merely about detecting similarities but also about interpreting the biological significance of the resulting groups. Previous studies utilizing machine learning for species delimitation have generally focused on datasets comprising approximately 5 populations or species, each with a moderate sample size (approximately 80) and a moderate number of SNPs (1,000 to 10,000) (e.g. [16]Smith and Carsten 2018; [21]Derkarabetian et al. 2019).

This study presents the first effort to address the issue of species delimitation using two machine learning approach, XGBBoost ([22]Chen and Guestrin 2016) and multi-layer perceptron (MLP, [23]Haykin 1994), to train a SNP dataset with eight different plant species within the pineapple family (Bromeliaceae). To our best knowledge, while XGBoost has been utilized for deciphering tree-based diversification driver in the Cactus family (Cactaceae) ([24]Thompson et al. 2023), its application to SNP datasets for species delimitation remains unexplored. Similarly, our MLP approach echoes neural network strategies like convolutional neural networks, which have been used to discern species or population structure in taxonomically complex Cactaceae ([25]Perez et al. 2022). However, these applications have largely focused on cryptic diversity, showing potential but also highlighting the need for more extensive research to validate their accuracy ([25]Perez et al. 2022). Our study marks the initial attempt to tackle species delimitation using XGBBoost and MLP, aiming to assess the training accuracy of our methods and to explore the limitations and potential enhancements for future studies.

Data

Eight species across five genera within the Bromeliaceae family were sequenced by low-coverage genome skimming by Novogene Inc. We downloaded the reference genome of *Ananas comosus* (pineapple, NC033621, [26]Ming et al. 2015) from GenBank. Raw sequencing reads were trimmed using fastp ([27]Chen et al. 2018), which eliminated adapters and excluded reads with a quality score below 20 or length under 100 base pairs. To ensure our low-coverage genome skimming sufficiently captured the necessary genetic information, we implemented three distinct validation methods. BUSCO ([28]Manni et al. 2021) was used to evaluate the capture of universal single-copy orthologs, while BLAST+ ([29]Camacho et al. 2009) and MiniMap2 independently verified the capture of single-copy orthologs (SCOs) based on criteria from [8]Yardeni et al. (2022), adjusting for mismatch thresholds. These checks confirmed that the genome skimming data was of high quality, with informative genetic representation across each genus. For SNP variant calling, we followed a standard protocol that involved indexing the reference genome with BWA ([30]Li and Durbin 2009), mapping the trimmed reads for each sample using Bowtie2 ([31]Langmead and Salzberg 2012), and applying BCFtools ([32]Li et al. 2011) for the variant calling process.

The core of our dataset is the Variant Call Format (VCF) ([33]Daneck et al. 2011), a specialized format designed to store gene sequence variations with high precision. This study specifically zooms in on the SNP variants, represented in the GT:PL format. Here, GT denotes the genotype, revealing the genetic constitution at a specific loci, while PL indicates the normalized Phred-scaled likelihoods, offering a quantitative measure of the genotype's accuracy by estimating the probability of each genotype possibility. This dual-layered data structure allows for exploring genetic variations, serving as a critical tool for understanding the genetic diversity and evolutionary relationships

among species.

Prior to the analytical phase, the datasets undergo a rigorous preprocessing stage. This essential step ensures that all data adheres to strict quality and consistency standards, paving the way for accurate and reliable species delimitation analysis. The final processed dataset is encapsulated in a dataframe with dimensions of shape (210859, 33), signifying the integration of 210859 SNP records across 33 distinct features. This structured approach not only facilitates efficient data manipulation and analysis but also enhances the clarity and interpretability of the results.

Methods

Our approach to addressing the problem of species delimitation through genetic data involves treating it as a classification problem using machine learning algorithms. Once the data is transformed into a matrix form, we will employ techniques such as XGBoost ([1] Chen and Guestrin 2016) and multi-layer perceptrons ([2]Rosenblatt 1962) to potentially enhance the predictive accuracy beyond that achievable with traditional clustering models. Our baseline for comparison will be a random forest model ([34]Breiman 2001), configured with 800 estimators and trained on the same dataset.

The three machine learning techniques, Random Forest, Multi-Layer Perceptron (MLP), and XGBoost algorithms, were used to understand species delimitation within the Bromeliaceae family was undertaken using. These methods were chosen based on their theoretical capability to manage high-dimensional data, like the SNP variants used in this study. The Random Forest model was used as a baseline model to evaluate the efficiency of XGBoost and MLP models.

The effectiveness of our models will be rigorously evaluated using the F1 score ([35]Chinchor 1992). F1 score will be used to assess the balance between precision (the proportion of positive identifications that are correct) and recall (the proportion of actual positives that were correctly identified) for each species. This is particularly important in the context of imbalanced classes, which is common in species data where some species may be more frequent than others.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \qquad \text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- TP = True Positives: the number of correct predictions that an instance is positive,
- TN = True Negatives: the number of correct predictions that an instance is negative,
- FP = False Positives: the number of incorrect predictions that an instance is positive,
- FN = False Negatives: the number of incorrect predictions that an instance is negative.

Model validation will employ the hold-out method, reserving 20% of the data for testing the models' ability to accurately delimit species not encountered during the training phase. The assignment of labels for training will be meticulously conducted based on manually verified genetic data. To ensure accuracy and consistency, we use the GT:PL format (Genotype: Probability List) which integrates expert knowledge into the genetic data interpretation process. Each species label is derived from annotations made by seasoned biologists or from previously validated datasets known for their reliability.

These expert annotations are critical as they serve as the gold standard for our analyses. The robustness of this gold standard is essential because it directly impacts the reliability of our model's performance evaluation. By grounding our label assignment in well-established genetic evidence and expert validation, we ensure that our model validation reflects the models' effectiveness in a realistic and operational context where precise species identification is paramount. This comprehensive

approach guarantees that the machine learning models are not only theoretically sound but also practically effective in real-world biodiversity research settings.

Results

Class Prediction Error

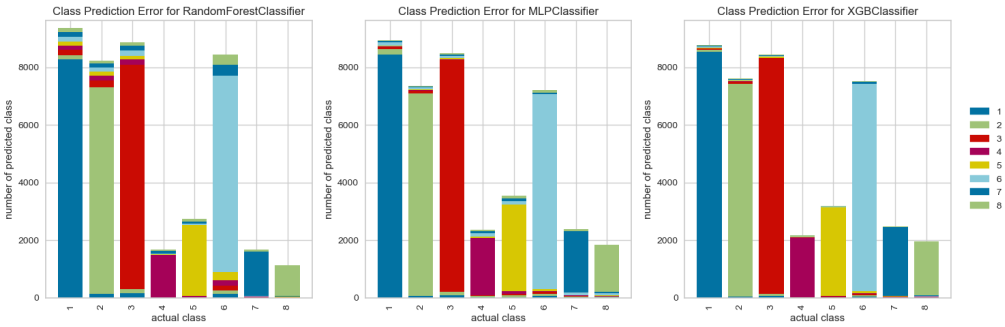

Figure 1. Class prediction errors for three classifier (Random Forest, MLP, XGBoost). The histograms display the count of misclassified instances across the eight plant species. The x-axis of each graph shows the actual class labels, and the y-axis indicates the number of predicted instances for each class. The bars are segmented by color, with each color representing a different predicted class. To interpret this figure, for example, if a bar representing the actual class 1 has significant segments of colors other than the one designated for class 1, it indicates that instances of class 1 are frequently misclassified as other classes. The random forest model has higher class prediction error, and this inference is drawn from the variety of colors across the bars corresponding to each actual class, indicating that predictions were more frequently assigned to incorrect classes, which suggests a less consistent prediction for several classes compared to the other models.

Figure 1 visualizes which classes (8 classes/species) are being mistakenly predicted by the each of the three classifiers (Random Forest, MLP, XGBoost). The XGBoost model demonstrates the best performance in terms of class prediction accuracy (Fig.1), while Random Forest exhibits the highest level of misclassification among the three models (Fig.1). The bars in the XGBoost plot predominantly match the color that represents the correct class, with minimal presence of other colors (Fig.1). This uniformity suggests that the XGBoost model has the most consistent and accurate class predictions, making it the superior model among the three in handling this particular dataset. MLP shows a moderate degree of misclassification, with fewer instances of incorrect class predictions than the Random Forest (Fig.1). This suggests that while the MLP may not be as precise as the XGBoost model, it still maintains a reasonable level of accuracy across most classes.

Confusion Matrix

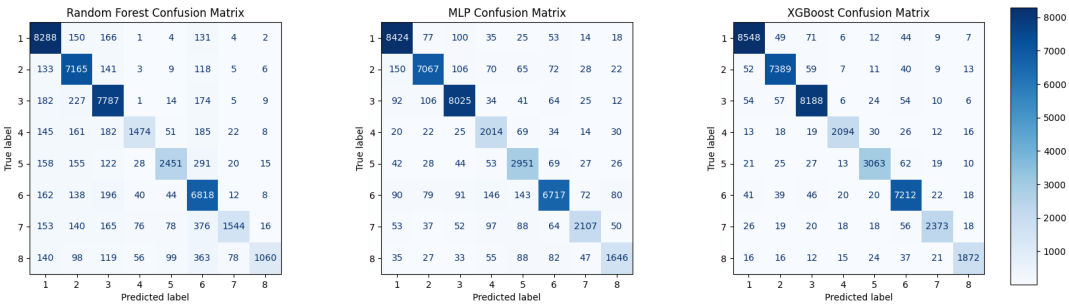

Figure 2. Confusion matrix for three models (Random Forest, MLP, XGBoost). Each matrix has the predicted labels on the x-axis and the true labels on the y-axis. The cells within the matrix contain values that represent the number of instances for each combination of predicted and true labels.

To illustrate the misclassification across classes, the confusion matrices shows which specific classes are being confused and the extent of the confusion (Fig.2). A cell with a notably high value off the diagonal denotes a substantial count of misclassifications for the respective pairing of predicted and actual labels. Conversely, high values along the diagonal signify accurate classifications.

The XGBoost model demonstrates a more distinct diagonal in its confusion matrix, which represents a higher rate of correct predictions, while the Random Forest and MLP show more spread across non-diagonal cells, indicating more frequent misclassifications (Fig.2). In all three models, species 1 (*Brocchinia accuminata*), species 2 (*B. paniculata*), and species 2 (*B. reducta*) are more commonly confused with one another than with other species (Fig.2).

Predicted Probability Distribution

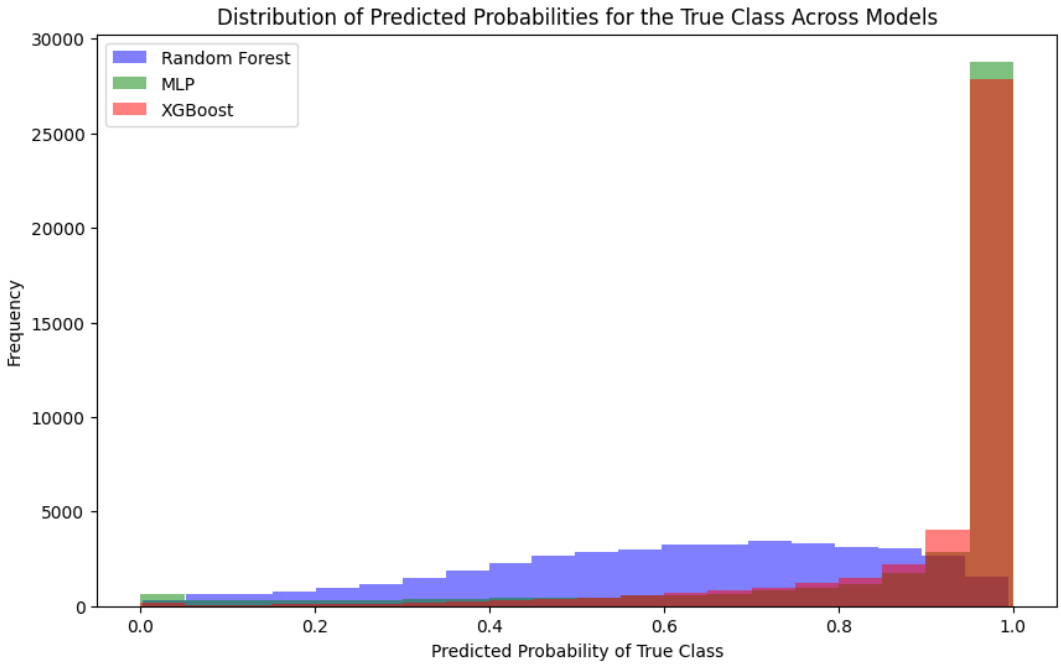

Figure 3. The histogram compares the confidence of predictions made by three different machine learning models: Random Forest, MLP, and XGBoost. The x-axis of the histogram represents the predicted probability of the true class, which ranges from 0.0 to 1.0. A predicted probability close to 1.0 indicates a high confidence in the prediction, while a probability close to 0.0 indicates low confidence. The y-axis shows the frequency or number of instances that were assigned each probability value by the models. The histogram bars are color-coded, with Random Forest in blue, MLP in green, and XGBoost in red. The concentration of bars towards the higher end of the probability range indicate a higher confidence in the predictions.

Based on the distribution of predicted probability, the Random Forest model shows a more uniform distribution of probabilities, while the MLP and XGBoost models have more instances with high confidence predictions, as indicated by the concentration of bars towards the 1.0 end of the x-axis (Fig.3). The distribution of predicted probabilities for the true class illustrates the confidence of each model in its predictions. The XGBoost model again shows a strong performance, with a significant majority of predictions having high confidence (probability close to 1) (Fig.3). This shows that XGBoost not only makes accurate predictions but also makes these predictions with a high degree of certainty. Meanwhile, Random Forest and MLP exhibit broader distributions of probabilities (Fig.3), indicating less consistent confidence across their predictions.

Model Accuracy

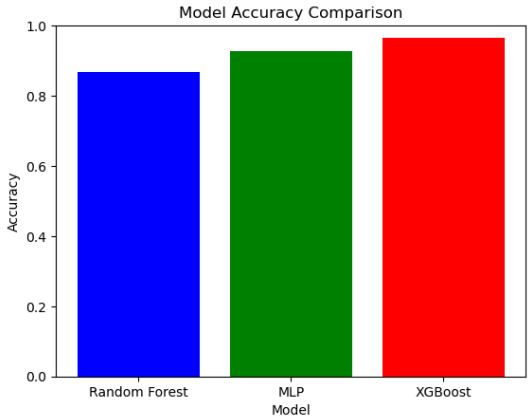

Figure 4. Model accuracy comparison. The x-axis lists the models, and the y-axis represents the accuracy scores, which range from 0 to 1, with 1 being perfect accuracy. Each model's accuracy is represented by a colored bar: blue for Random Forest, green for MLP, and red for XGBoost. The height of each bar corresponds to the model's accuracy score.

Regarding model accuracies for all classifiers, the Random Forest model's accuracy is around 0.87, the MLP model's accuracy is higher at around 0.92, and the XGBoost model has the highest accuracy, approximately 0.97 (Fig.4). The XGBoost model outperforms both the Random Forest and MLP models, with a near-perfect accuracy score.

Discussion

This study represents a pioneering effort in plant biology by introducing XGBoost and MLP methods to explore species delimitation using large-scale genomic data. The results of this study provide strong evidence for potential application of machine learning algorithms in species delimitation by using SNP data. The high accuracy and confidence of the XGBoost model demonstrates its potential as a powerful tool for better identification of species boundaries that may not be discernible through traditional methods.

The superior performance of XGBoost could be attributed to its ensemble learning properties that effectively captures complex patterns in the data by combining multiple weak learners into a strong one ([1]Chen and Guestrin 2016). Our results suggest that the gradient boosting technique employed by XGBoost is highly effective for this dataset (Chen and Guestrin 2016), possibly due to its capacity for handling the complexities and interactions within the SNP data. The presence of highly confident predictions aligns with the theoretical efficiency of the algorithm in handling various types of data distributions and interactions among features.

XGBoost has been effectively employed to study evolutionary biology questions. For example, XGBoost has been used to understand drivers for diversification in the Cactus family by predicting the effects of biotic and abiotic variables on diversification rates along phylogenetic lineages ([24] Thompson et al. 2023). Despite the application of XGBoost in addressing other biological questions in plants, our study marks the first effort to showcase the promising potential of employing XGBoost to understand species delimitation using whole genome SNP data in plant biology.

The MLP model showed varying degrees of confidence in its predictions, possibly due to its sensitivity to the scale of input data and the choice of architecture. Although MLP has been widely used in various biological fields, such as agricultural and ecological imaging training (e.g. [36] Sumsion et al.

2019), tumor classification (e.g. [37] Potghan et al. 2018), etc, MLP has not been explored in training SNP data to understand species delimitation. The observed errors suggest that further optimization of network hyperparameters or the inclusion of regularization techniques might enhance its predictive capacity to fully explore its capacity in classifying species with genomic data.

The Random Forest model, while generally robust and less prone to overfitting, employs multiple decision trees through classification tree and tree bagging techniques. Each tree makes its own classification, and the final decision is made by combining the outputs of all trees through a voting process ([38] Breiman 1996; [34] Breiman 2001). The poor performance of our Random Forest model may be attributed to the high dimensionality and complexity of SNP data. Its misclassifications across species suggest a need for parameter tuning or potentially incorporating feature selection methods to reduce noise and focus on more informative genetic markers.

Despite its limited model accuracy in species classification shown by this study, Random Forest has been proven to be useful in addressing evolutionary biology questions using genomic data. For instance, comparative studies between phylogenetic methods and classification models employing DNA barcode data and Random Forest have demonstrated their utility in classifying taxonomic groups ([39] Austerlitz et al. 2009). Previous research also shows that machine learning classification methods, such as RF, could be easily misled by mutations shared by different species, while traditional phylogenetic methods are more likely to treat different individuals from the same species as different species ([39] Austerlitz et al. 2009). Furthermore, previous attempts using unsupervised Random Forest for species delimitation have highlighted its potential in accurately identifying species boundaries, yet its effectiveness in discerning various evolutionary scenarios such as adaptive radiation and cryptic species remains largely unexplored (Derkarabetian et al., 2019).

The variation in model performance also highlights the complexity of species delimitation. Highly related species (species 1:(*Brocchinia accuminata*), species 2 (*B. paniculata*), and species 3 (*B. reducta*) are more easily to be misclassificed with each other due to their similar genetic profiles. This observation aligns with previous phylogenetic studies indicating a potential need for taxonomic refinement within the genus *Brocchinia* (unpublished). This result suggests that identifying species with similar genetic profile requires more refined data or advanced modeling techniques to differentiate accurately. The inconsistencies between models for highly related species suggest that integrating multiple modeling approaches or refining feature selection could improve performance, especially for species with high misclassification rates.

Furthermore, the variability in error rates across species implies that while machine learning can significantly facilitate species delimitation, there remains a need for domain expertise to interpret the biological significance of the predictions. Previous research assessing the performance of multispecies coalescent models in species delimitation underscores the importance of cautiously interpreting genomic data and statistical models in this context, which advocates for an integrative approach that combines morphological and genomic data to comprehensively understand species identification ([14] Sukumaran and Knowles 2017).

These findings highlight the importance of comprehensive model evaluation, not only in terms of overall accuracy but also in considering the reliability of predictions for each class, applications under different evolutionary scenarios, and ability to deal with different taxonomic groups with more complicated between-population level genetic structure. This study also points to the significance of probability scores in understanding model confidence and making informed decisions in species conservation efforts.

Conclusion and Future Work

Our study demonstrates the use of machine learning techniques, particularly XGBoost, in the field of species delimitation based on SNP data. The XGBoost model, in particular, has shown remarkable accuracy and prediction confidence, underscoring its potential as a robust tool for taxonomic classification. Looking ahead, the potential for enhancing species delimitation using machine learning is vast. Exploring additional machine learning models could provide complementary strengths to those observed with XGBoost and MLP. The integration of deep learning ([40] Hinton et al. 2006) and ensemble methods might offer new insights and potentially superior performance in managing complex genetic data.

Despite its success, the research also unveiled several challenges. First, there is a high misclassification rate for all models when dealing with closely-related species, including species 1 (*Brocchinia accuminata*), species 2 (*B. paniculata*), and species 3 (*B. reducta*). These findings highlight the necessity for more refined data and sophisticated modeling techniques to accurately delineate species.

Additionally, our study was constrained by a limited dataset comprising only eight samples from five genera. Each sample represented a distinct species, leading us to treat all labeled classes as distinct entities. Consequently, we were unable to assess the model's performance in classifying data with shared labels. A comprehensive training dataset for species delimitation typically require multiple individuals from the same species but from different populations. The absence of population-level genomic data in our dataset limited our ability to further investigate the model's performance in deciphering population-level genetic structure. To overcome these limitations, it is important to augment the training dataset by increasing sample sizes and incorporating population-level diversity. Such augmentation would not only enhance the training process but also improve the predictive capabilities of the models. Moreover, expanding the dataset to encompass broader genomic regions and integrating multi-omics data could significantly enrich our understanding of species differentiation.

Third, advanced feature selection methods represent another promising area for future research. These methods could help pinpoint the key genetic markers that are crucial for distinguishing species, thereby improving model accuracy and facilitating a deeper comprehension of the specific SNPs responsible for the observed species delimitation. Additionally, developing a hybrid approach that combines the computational power of machine learning with the detailed biological insights offered by traditional phylogenetic and coalescent-based methods could lead to more accurate and robust species delimitation tools.

In conclusion, despite the recent popularity of studying machine learning methods, their application in evolutionary and conservation biology in plants has been limited to model organisms or a few taxonomic groups (e.g. [21] Derkarabetian et al. 2019). Applying these methods to other plant families and even non-plant, non-model organisms could help validate the utility and effectiveness of machine learning approaches in species delimitation across various biological systems. This could pave the way for their application in conservation biology, where they could assist in the accurate identification of endangered species and the planning of effective conservation strategies based on detailed understandings of genetic diversity and species boundaries. Through these efforts, we can enhance our ability to understand and protect biological diversity using the latest advances in computational techniques.

References

[1] T Chen **and** C Guestrin. XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754. 2016. URL: https://arxiv.org/abs/1603.02754.

[2] F Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books, 1962.

[3] J. Duminil **and** M. Di Michele. Plant Species Delimitation: A Comparison of Morphological and Molecular Markers. **in***Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*: 143.3 (2009), **pages** 528–542.

[4] H Pedersen. Species delimitation and recognition in the Brachycorythis helferi complex (Orchidaceae) resolved by multivariate morphometric analysis. **in***Botanical Journal of the Linnean Society*: 162.1 (2010), **pages** 64–76. DOI: 10.1111/j.1095-8339.2009.01015.x.

[5] Z Liang, S Duan, J Sheng **andothers**. Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. **in***Nature Communications*: 10 (2019), **page** 1190. DOI: 10.1038/s41467-019-09135-8.

[6] ZY Ma, J Wen, JP Tian, LL Gui **and** XQ Liu. Testing morphological trait evolution and assessing species delimitations in the grape genus using a phylogenomic framework. **in***Molecular Phylogenetics and Evolution*: 148 (2020), **page** 106809. DOI: 10.1016/j.ympev.2020.106809.

[7] MG Johnson, L Pokorny, S Dodsworth **andothers**. A Universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. **in***Systematic Biology*: 68.4 (2019), **pages** 594–606. DOI: 10.1093/sysbio/syy086.

[8] G Yardeni **andothers**. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. **in***Molecular Ecology Resources*: 22 (2022), **pages** 927–945.

[9] M Hedin, S Derkarabetian, A Alfaro, MJ Ramírez **and** JE Bond. Phylogenomic analysis and revised classification of atypoid mygalomorph spiders (Araneae, Mygalomorphae), with notes on arachnid ultraconserved element loci. **in***PeerJ*: 7 (2019), e6864. DOI: 10.7717/peerj.6864.

[10] KF Adam DL and Matthew, NM Vladimir **and** RB Remco. Species Delimitation using Genome-Wide SNP Data. **in***Systematic Biology*: 63.4 (2014), **pages** 534–542.

[11] THG Ezard, PN Pearson **and** A Purvis. Algorithmic approaches to aid species' delimitation in multidimensional morphospace. **in***BMC Evolutionary Biology*: 10 (2010), **page** 175.

[12] Z Yang **and** B Rannala. Bayesian species delimitation using multilocus sequence data. **in***Proceedings of the National Academy of Sciences of the United States of America*: 107.20 (2010), **pages** 9264–9269. DOI: 10.1073/pnas.0913022107.

[13] B Rannala **and** Z Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. **in***Genetics*: 164.4 (2003), **pages** 1645–1656.

[14] J. Sukumaran **and** L. L. Knowles. Multispecies coalescent delimits structure, not species. **in***Proceedings of the National Academy of Sciences of the United States of America*: 114.7 (2017), **pages** 1607–1612.

[15] L Yang, H Kong, JP Huang **and** M Kang. Different species or genetically divergent populations? Integrative species delimitation of the Primulina hochiensis complex from isolated karst habitats. **in***Molecular Phylogenetics and Evolution*: 132 (2019), **pages** 219–231. DOI: 10.1016/j.ympev.2018.12.011.

[16] M. L. Smith **and** B. C. Carstens. Demographic model selection using random forests and the site frequency spectrum. **in***Systematic Biology*: 67.4 (2018), **pages** 605–618. DOI: 10.1093/sysbio/syx082. URL: https://academic.oup.com/sysbio/article/67/4/605/4569848.

[17] C Ma, M Xin, KA Feldmann **and** X Wang. Machine learning–based differential network analysis: A Study of stress-responsive transcriptomes in Arabidopsis. **in***The Plant Cell*: 26.2 (2014), **pages** 520–537.

[18] D. Singh, D. D. Pant **andothers**. Influence of High and Low Levels of Plant-Beneficial Heavy Metal Ions on Plant Growth and Development. **in***Frontiers in Environmental Science*: 4 (2016).

DOI: 10.3389/fenvs.2016.00069. URL: https://www.frontiersin.org/articles/10.3389/fenvs.2016.00069/full.

[19] S Cheng, W Zeng, J Wang, L Liu, H Liang, Y Kou, H Wang, D Fan **and** Z Zhang. Species Delimitation of Asteropyrum (Ranunculaceae) Based on Morphological, Molecular, and Ecological Variation. **in***Frontiers in Plant Science*: 12 (2021), **pages** 681–864. DOI: 10.3389/fpls.2021.681864.

[20] J Pei, C Chu, X Li, B Lu **and** Y Wu. CLADES: A classification-based machine learning method for species delimitation from population genetic data. **in***Molecular Ecology Resources*: 18 (2018), **pages** 1144–1156. DOI: 10.1111/1755-0998.12887.

[21] S Derkarabetian, S Castillo, PK Koo, S Ovchinnikov **and** M Hedin. A demonstration of unsupervised machine learning in species delimitation. **in***Molecular Phylogenetics and Evolution*: 139 (2019), **page** 106562. DOI: 10.1016/j.ympev.2019.106562.

[22] T Chen **and** C Guestrin. XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754. 2016. URL: https://arxiv.org/abs/1603.02754.

[23] S Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[24] JB Thompson, T Hernández-Hernández, G Keeling **and** NK Priest. Identifying the multiple drivers of Cactus diversification. **in***bioRxiv*: (2023). DOI: 10.1101/2023.04.24.538150.

[25] MF Perez, IAS Bonatelli, M Romeiro-Brito, FF Franco, NP Taylor, DC Zappi **and** EM Moraes. Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. **in***Molecular Ecology Resources*: 22 (2022), **pages** 1016–1028. DOI: 10.1111/1755-0998.13534.

[26] R Ming, R VanBuren, C Wai **andothers**. The pineapple genome and the evolution of CAM photosynthesis. **in***Nature Genetics*: 47 (2015), **pages** 1435–1442.

[27] S Chen, Y Zhou, Y Chen **and** J Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. **in***Bioinformatics*: 34.17 (2018), **pages** i884–i890.

[28] M Manni, MR Berkeley, M Seppey **and** EM Zdobnov. BUSCO: Assessing genomic data quality and beyond. **in***Current Protocols*: 1 (2021), e323. DOI: 10.1002/cpz1.323.

[29] C Camacho, G Coulouris, V Avagyan **andothers**. BLAST+: architecture and applications. **in***BMC Bioinformatics*: 10 (2009), **page** 421.

[30] H Li **and** R Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. **in***Bioinformatics*: 25.14 (2009), **pages** 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[31] B Langmead **and** SL Salzberg. Fast gapped-read alignment with Bowtie 2. **in***Nature Methods*: 9.4 (2012), **pages** 357–359. DOI: 10.1038/nmeth.1923.

[32] H Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **in***Bioinformatics*: 27.21 (2011), **pages** 2987–2993.

[33] P. Danecek **andothers**. The variant call format and VCFtools. **in***Bioinformatics*: 27.15 (2011), **pages** 2156–2158.

[34] L Breiman. Random Forests. **in***Machine Learning*: 45 (2001), **pages** 5–32. DOI: 10.1023/A:101093340432.

[35] Nancy Chinchor. MUC-4 evaluation metrics. **in**MUC4 '92: (1992), **pages** 22–29. DOI: 10.3115/1072064.1072067. URL: https://doi.org/10.3115/1072064.1072067.

[36] G. R. Sumsion, M. S. Bradshaw, K. T. Hill, L. D. G. Pinto **and** S. R. Piccolo. Remote sensing tree classification with a multilayer perceptron. **in***PeerJ*: (2019). DOI: 10.7717/peerj.6101.

[37] Sneha Potghan, R. Rajamenakshi **and** Archana Bhise. Multi-Layer Perceptron Based Lung Tumor Classification. 2018. DOI: 10.1109/ICECA.2018.8474864.

[38] L. Breiman. Bagging predictors. **in***Machine Learning*: 24 (1996), **pages** 123–140.

[39] F. Austerlitz, O. David, B. Schaeffer **andothers**. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. **in***BMC Bioinformatics*: 10.Suppl 14 (2009), S10. DOI: 10.1186/1471-2105-10-S14-S10. URL: https://doi.org/10.1186/1471-2105-10-S14-S10.

[40] G. E. Hinton, S. Osindero **and** Y. W. Teh. A fast learning algorithm for deep belief nets. **in***Neural Computation*: 18.7 (2006), **pages** 1527–1554.