SPECIES DELIMITATION BY MACHINE LEARNING METHODS

by

JIAN WU

(Under the Direction of Liang Liu)

ABSTRACT

Species, in biology, are determined by a classification of related organisms that share common characteristics and are capable of interbreeding. People are interested in how species can be delimited, or whether their common ancestor can be found based on genetic sequences. In general, there are several methods used in species delimitation, such as the Automated Barcode Gap Discovery method, the General Mixed Yule Coalescent method, and the Poisson Tree Process method, etc. However, these methods have several disadvantages, including being time-consuming, hard to solve the big dataset problem, etc. In our design, we explore using supervised machine learning methods (Catboost, XGboost, Classification Tree, Support Vector Machine, K-nearest Neighbors), an unsupervised machine learning method (K-means Clustering), and a deep learning method (Neural Network) in species delimitation. Five species trees are determined to be our treatments. The results show that supervised machine learning models have the highest accuracy compared to the unsupervised machine learning model and deep learning model.

INDEX WORDS: Species, Species Delimitation, Machine Learning, Deep Learning.

SPECIES DELIMITATION BY MACHINE LEARNING METHODS

by

JIAN WU

B.S. Xiangtan University, 2012

M.S. Kunming University of Science and Technology, 2014

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment

of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2024

SPECIES DELIMITATION BY MACHINE LEARNING METHODS

by

JIAN WU

Major Professor:	Liang Liu
Committee:	Shuyang Bai
	Ting Zhang

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2024

ACKNOWLEDGEMENTS

First, I will thank my advisor Prof. Liang Liu. He gave me a lot of support and guidance. I don’t have much statistical background, but he was very patient and gave me a lot of advice which is vital for my project.

Second, I will thank my committee members (Prof. Ting Zhang and Prof. Shuyang Bai) for their generous help and prompt suggestions.

Finally, I will thank my wife and my family. Pursuing two master’s degrees at the same time is very difficult. They gave me warmth and support when I was helpless. They are my biggest motivation to persist in completing my degree.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Species and Species Delimitation

Species are considered as the currency of biodiversity, have fundamental importance to ecological and evolutionary studies, and are key to formulating conservation efforts (Martin et al., 2021). Different species that are related to the dynamics of the speciation continuum are considered to be elements of diverse properties (Perez et al., 2021). Currently, the rate of extinction of species is difficult to estimate because the species boundaries are unclear and about 80-90% of species are undiscovered or undescribed. The most likely is that numerous contemporary species are not documented by scientists even though they are already distinct (Rannala and Yang, 2020). Defining species or the importance of process by new species arise continue to be contentious topics.

Species delimitation research is continuously advancing with new data types, innovative scientific approaches, and theoretical frameworks (Derkarabetian et al., 2019). However, there are several significant confusions on species delimitation (Pei et al., 2018). The most significant issue is that the widely acceptable definition of the concept of species is lacking. For example, The evolutionary species concept (ESC) defines a species as a lineage that maintains an identity distinct from other lineages, has unique evolutionary trends, and has a specific historical destiny. This

concept emphasizes the importance of considering the current characteristics and evolutionary history of lineages when defining species (Pei, 2017). It is hard to delimit species. Furthermore, the challenging task of selecting appropriate methods to describe species becomes particularly evident in species complexes, as highlighted by Pinheiro et al. (2018). Effectively identifying gaps between early divergence stages, which are often prevalent across species complexes, requires considerable effort and a multidisciplinary approach to evaluate various forms of evidence substantiating constraints in different species (Perez et al, 2021).

1.2 History of Species Delimitation

1.2.1 Numerical Taxonomy and Molecular Taxonomy

In the 1960s, the concept of using computer algorithms to classify species based on morphological characteristics was proposed. Numerical taxonomy, which was first used by bacterial taxonomists, to use a large number of trait measurements can compensate for the lack of unique traits among bacterial strains. However, numerical taxonomy did not solve the problems related to convergent evolution and caused unstable classification at the same time. Although it is possible to cluster individual operational taxonomic units (OTUs) into groups with the same characteristics, it cannot provide a theoretical basis for distinguishing species or establishing species boundaries, nor can it trace the origin of species (Rannala and Yang, 2020).

In the 1970s, molecular sequence data were widely recognized for distinguishing groups such as bacteria, but there was no method for widespread use among groups of organisms in determining species boundaries and classifying species. The need for large sample sizes makes it difficult to

find convincing evidence, even in pilot trials targeting morphological shapes (Rannala and Yang, 2020).

1.2.2 Gene Trees

The “genealogical species concept” is proposed by Avise and Ball Jr (1990) and accelerates the proposal and research on the definition of the common ancestor model of genetic tree species. The criterion of this concept is “reciprocal monophyly”, that all sequences of a species in the corresponding gene tree form a monophyletic group. However, considering gene trees in this method belong to observations, this method cannot determine the errors when inferring gene trees and will reduce the efficiency of reciprocal monophyly. The size of the population has a significant impact on the use of this method. When the population size is very large, this method cannot accurately identify species because the isolation time required to achieve exclusivity at the site is longer. When the population size is well known, the isolation period of the population is too short and it is difficult to distinguish them into different species (Rannala and Yang, 2020).

1.2.3 DNA Barcoding

In the 2000s, the use of molecular data for species assignment and delimitation was a very important issue, the most representative of which was the proposal for the "DNA Barcoding" initiative. The rationale for this act is to use a single locus to provide a rationale for species delimitation. However, there are many critics of the use of DNA barcoding for species delineation. Because interspecific distances and intraspecific distances will be very similar in large populations, the threshold cannot be defined, and there is no fixed threshold. At the same time, using a single locus method for species definition is also a very inefficient method (Rannala and Yang, 2020).

1.2.4 Multispecies Coalescent

In the 2000s, researchers proposed the multispecies coalescence (MSC) method to distinguish species. This method can provide probability distributions and statistical models for the possible gene trees given a specific species tree, provided there is no gene flow between species. This statistical model can map individuals to species based on single-site or multi-site sequence data to achieve species delimitation and species boundary determination (Rannala and Yang, 2020). Hebert et al. (2003) provide an alternative approach to delimitation species based on the percent sequence divergence between species which is sensitive to the levels of polymorphism within populations. The method of multispecies coalescent does not require mutual monophyly to define species and can also reduce the statistical uncertainty of gene trees (Rannala and Yang, 2020).

1.3 Species Delimitation Methods

In the species delimitation methods, they can be divided into two categories (heuristic methods and parametric methods). Heuristic methods use summary statistics or algorithms to delimit species and do not rely on a statistical model. This method is effective to use in a large dataset, but the results may not be easy to interpret due to the bad statistical properties. The most common methods include the Automated Barcode Gap Discovery (ABGD) method, the General Mixed Yule Coalescent (GMYC) method, and the Poisson Tree Process (PTP) method. Compared to the heuristic methods, parametric methods are based on a probabilistic model which is a model of the biological process generating gene trees and DNA sequences. Most of the parametric methods belong to Multispecies coalescent (MSC) models, especially the Bayesian phylogenetics and

phylogeography (BPP) method. The following part will have a brief description of these two methods (Rannala and Yang, 2020).

1.3.1 Automated Barcode Gap Discovery

The concept of barcode gaps was proposed after early studies of serial differences in trip pairs between and within species. Differences caused by intra-group differences and species differences can be distinguished through this gap. The ABGD program, proposed by Puillandre et al. (2012), aims to determine the threshold of barcode gaps in an automated process. ABGD may perform best in cases where the gene tree is monophyletic in nature, that is, where there is no incomplete lineage ordering of the gene tree. Like other barcoding methods, ABGD only utilizes information from a single locus. Its computational cost is relatively low and suitable for processing large sequence samples. However, ABGD also has some shortcomings. It relies on simple pairwise distance calculations and clustering operations and does not fully exploit all the information in the sequence data. The distribution of intraspecific distances may be polymodal due to factors such as population growth or selection, which may lead to spurious barcode gaps (Rannala and Yang, 2020).

1.3.2 General Mixed Yule Coalescent

The GMYC method is a technique for delimiting species using a likelihood approach, which involves fitting intra- and inter-species branching models to reconstructed gene trees (Fujisawa and Barraclough, 2013). Based on the coalescence events or branch lengths in a gene tree, the waiting times can be divided into two classes: the coalescent process determines the rate within species and a generalization of the Yule process model of species divergences determines between

species (Pons et al., 2006). It is widely used in inferring gene trees, but it has several disadvantages. It will ignore errors in the gene trees, and this method only be used in a single locus, even though some researchers try to use it in multiple loci. Furthermore, the GMYC method has good results for data with small population sizes and long speciation intervals, because this method assumes that the gene trees are mutually monophyletic and does not consider the merger relationship within the species population (Rannala and Yang, 2020).

1.3.3 Poisson Tree Process

The PTP methods through the distribution of branch lengths in the gene tree to determine the species status (Zhang et al., 2013). PTP is very similar to the GMYC method. Both of them bridge the gap between species of trees and restrict to a single lotus (Kapli et al., 2017). Based on the maximum likelihood, the tree and branch length can be inferred without error. Even if there are multiple loci, they can be connected to infer the gene tree with branch lengths.

PTP models the branching process in terms of the expected number of substitutions accumulated between subsequent speciation events. This model aims to determine the transition point from interspecific to intraspecific processes by comparing two parametric models, one for speciation and the other for the aggregation process, to optimally fit the data. Unlike other methods, PTP directly adopts the alternative method without the need for a hypermetric input tree, avoiding the time-consuming and errors that may occur during the inference process (Kapli et al., 2017). Therefore, compared to GMYC, PTP can usually provide a more accurate delimitation. Some weaknesses of this approach can be easily identified. This approach is essentially a single-locus

approach because concatenated sequences across genetic loci or genome segments cannot account for random fluctuations in the merging process between loci (Rannala and Yang, 2020).

1.3.4 Multispecies Coalescent Model

MSC models are novel species delimitation methods in which individual gene trees are estimated simultaneously or separately with species trees as a means of estimating phylogenetic relationships. This model is widely used in maximum likelihood, Bayesian, and nonparametric approaches (Edwards et al., 2016). The marginal likelihoods and the posterior probabilities of different species delimitation models are calculated through multi-locus sequence data. In contrast to traditional phylogenetic analysis, which assumes that all gene loci lie under the same tree, MSC accounts for the merging of modern and ancestral species and the resulting conflict of gene trees with species trees. Therefore, reliable estimates of species phylogenies can be made even if the information at each site is weak, resulting in a high degree of uncertainty in the gene tree (Heled and Drummond, 2010). Moreover, BPP is the most popular method using the MSC model. BPP is a Bayesian Markov chain Monte Carlo program for analyzing DNA sequence alignments (Yang, 2015). Although this approach is attractive in theory, its utility in actual data analysis still needs to be rigorously tested. In particular, this analysis may be sensitive to priors on ancestral population size, species divergence time, and interspecies gene flow (Zhang et al., 2011).

CHAPTER 2

METHODS

2.1 Machine Learning Methods

Machine learning is a dynamic field of computational algorithms that aims to replicate human intelligence by adapting to and learning from the environment (Naqa and Murphy, 2015). There are several advantages of machine learning, including computational efficiency and predictive accuracy in analyzing large, complicated, and high-dimensional datasets (Salles and Domingos, 2023). In the field of artificial intelligence, machine learning has become the preferred method for creating functional software for computer vision, speech recognition, natural language processing, robot control, and various other applications (Jordan and Mitchell, 2015). The effect of machine learning is broad across from computer science (Mitchell, 2006) to multiple other fields including industries with data-intensive issues (Chen and Zhang, 2014), biology (Tarca et al., 2007), cosmology (Ntampaka et al., 2019), social science (Grimmer et al., 2021), etc. In particular, machine learning methods can be divided into two categories: supervised machine learning (SML) and unsupervised machine learning (UML). Moreover, deep learning (DL) has been widely used recently which belongs to the subfield of machine learning. The following three parts will introduce SML, UML, and DL.

2.1.1 Supervised Machine Learning Methods

The SML method is widely used in solving classification and regression problems. It predicts data points using a training dataset which includes labeled data (features) input and output (response variable). Supervised learning is the most common technique for training neural networks and decision trees. Figure 1 is the SML model. As shown in Fig.1, the algorithm identifies observed data X, which typically constitutes the training data provided to the model during the training phase, usually in the form of structured data. In this process, the algorithm builds a predictive model. After training, the model predicts the most likely label for a new set of samples X in the test dataset (Nasteski, 2017).

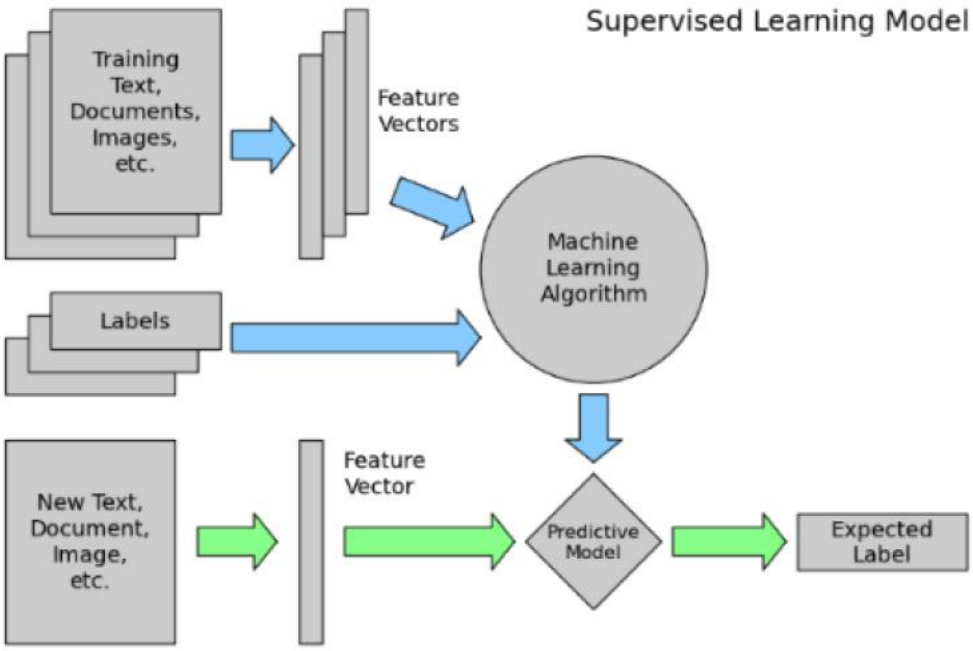


Figure 1: Supervised machine learning model (Inspired by Nasteski, 2017)

Moreover, the SML method is becoming used in genetic data, and in the analysis process, there are six steps about the SML method used in species delimitation, including (Salles and Domingos, 2023):

(i) Designing Evolutionary model

(ii) Simulating data

(iii) Choosing how to represent the biological data

(iv) Applying the appropriate SML algorithm in the training dataset

(v) Evaluating performance and optimizing parameters

(vi) Applying the appropriate SML algorithm in the testing dataset

2.1.2 Unsupervised Machine Learning Methods

UML in artificial intelligence is a type of machine learning that learns from data without human supervision. Unlike SML, UML models are given unlabeled data and are allowed to discover patterns and insights without any explicit guidance or instructions. UML only depends on the inherent data structure to group samples. UML methods can be found in diverse integrative taxonomic studies, including genetic, morphometric, etc. (Salles and Domingos, 2023). In general, UML methods include three categories, such as clustering, dimensionality reduction, and association rules.

Clustering is a valuable technique in data science that aims to identify cohesive structures in data sets by emphasizing similarities within clusters and differences between different clusters. Hierarchical clustering was one of the pioneering methods originally employed by biologists and social scientists, while cluster analysis developed into a specialized field in statistical multivariate

analysis (Sinag and Yang, 2020). There are generally four types of clustering, including exclusive clustering, overlapping clustering, hierarchical clustering, and probabilistic clustering. Figure 2 is the k-means clustering plot which belongs to exclusive clustering. k-means clustering is a vector quantization method that aims to partition n observations into k clusters, where each observation belongs to the cluster with the nearest mean (cluster center or cluster centroid), as the cluster.

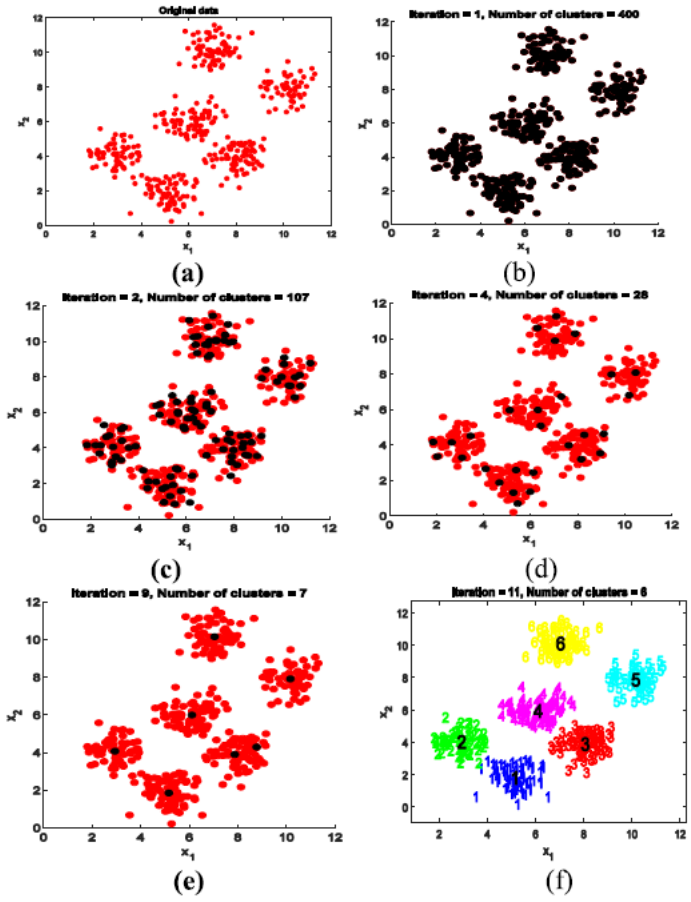


Figure 2: K-means clustering plot (a) Original data set; (b)-(e) Processes of the U-k-means after 1, 2, 4, and 9; (f) Convergent results. (Inspired by Sinag and Yang, 2020)

Dimensionality reduction is an important and efficient method to solve large dataset problems (Dash et al., 1997). The key to this method is reducing the number of features or dimensions in the

dataset. The principal component analysis (PCA) and singular value decomposition (SVD) are the most commonly used methods in dimensionality reduction methods. Figure 3 is an example of PCA used in the species delimitation.

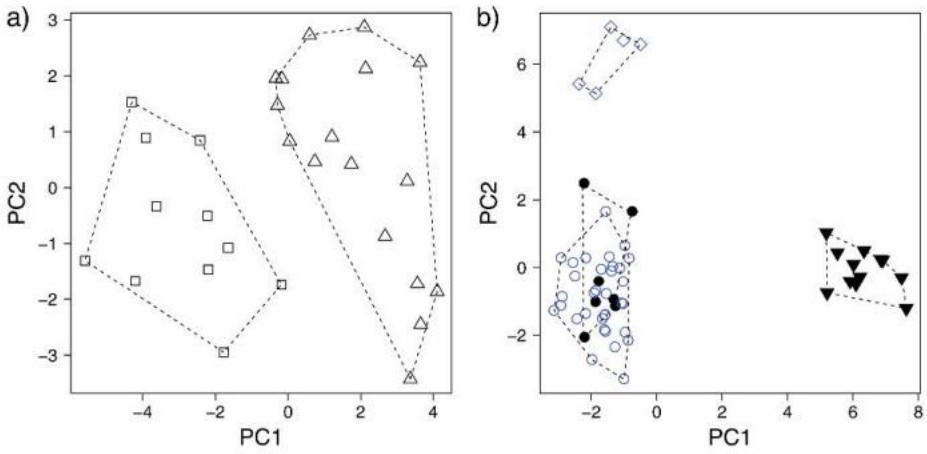


Figure 3: PCA analysis among hypothesized Escallonia species. a and b represent two different examples (Inspired by Zapata and Jiménez, 2012).

Association rule learning in the UML method is used to reveal associations between variables in massive data sets. Unlike certain machine learning methods, association rule learning can accept non-numeric data points. Briefly speaking, association rule learning focuses on how specific variables are linked (Naeem et al., 2023).

2.1.3 Deep Learning Methods

Deep learning is indeed a subfield of machine learning that focuses on learning representations of data through layered architectures called neural networks. It is widely used in traditional artificial intelligence domains, such as transfer learning, natural language processing, computer version, etc

(Guo et al., 2016). There are three reasons why deep learning is widely used in our daily lives: the low cost of computing hardware, the development of machine learning algorithms, and the dramatically increased chip processing abilities (Deng., 2014). In general, deep learning methods can be divided into four categories: Convolutional Neural Networks, Restricted Boltzmann Machines, Autoencoder, and Sparse Coding (Guo et al., 2016).

2.2 Current Machine Learning Applications on Species Delimitation

The machine learning method has been used for species delimitation recently. In Table 1, we introduce the current machine learning application on species delimitation, including method, algorithm, input, etc. Also, we will give details about how these researchers use SML or UML to delimit species. Figure 4 is the comparison plot between UML and SML algorithms on species delimitation. In particular, the key to UML is related to clustering and dimensional reduction, and the key to SML is focused on using simulated datasets to train a classifier.

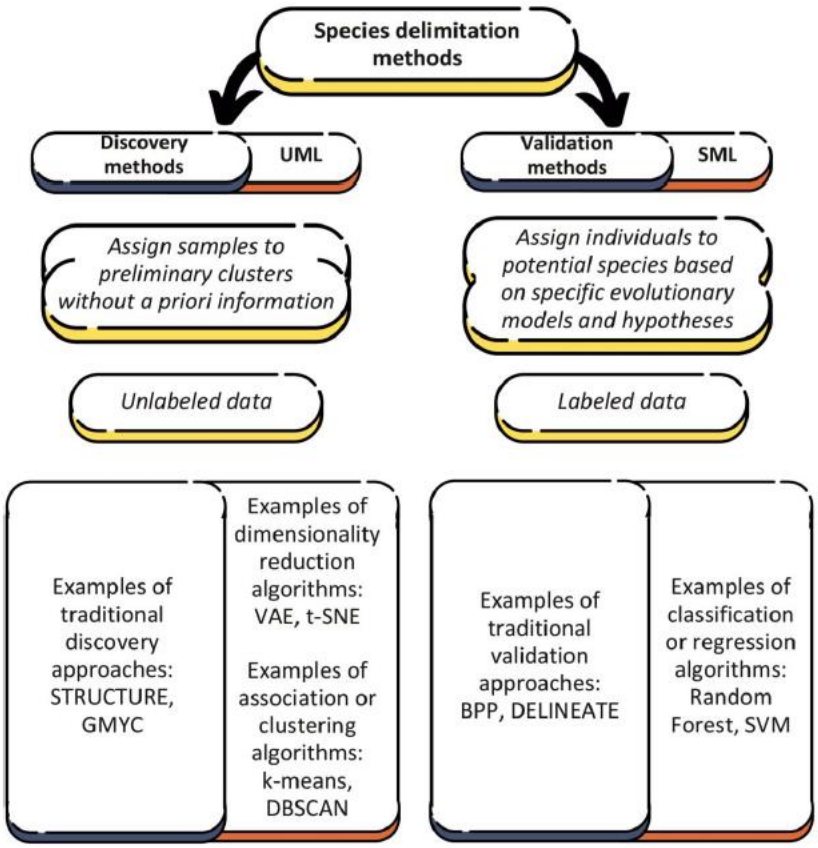


Figure 4: Comparing UML and SML algorithms on species delimitation. (Inspired by Salles and Domingos, 2023)

2.2.1 Supervised Machine Learning Model

Smith and Carstens. (2018) represented data using the site frequency spectrum (SFS), compared the model using random forest (RF) by applying the R-package delimitR, and identified species boundaries in the reticulate taildropper slug (Prophysaon andersoni). DelimitR enables users to analyze data by employing a default set of models along with user-defined priors. These models exhibit variation not only with the inclusion or exclusion of population-level processes such as migration and population size changes but also with variation in the number of species or populations considered. There are three steps about species delimitation in delimitR, including

generating the default model and adding models to this model set; simulating data in fsc26 using the default model or a user-specified model set, and constructing the RF classifier and calculating error rates using out-of-bag (oob) error rate. The results show that even though the detection of migration becomes more challenging due to recent divergence between lineages, the RF method which is used by delimitedR has a low error rate and can successfully delimit species.

Smith and Carstens. (2020) designed a simulation study and considered four scenarios in a three-population system, including no population divergence, the divergence between two populations, divergence among three populations, and divergence among all three populations which has secondary contact between two populations using delimitR. In the analysis process, the moderate divergence time (50000-100000 generations) and recent divergence time (5000-10000 generations) are considered with 500-10000 unlinked Single Nucleotide Polymorphisms (SNPs). The results show that when the SNPs are equal to 1500 under recent divergence time has almost zero error rate. Considering a more complicated model which includes divergence, gene flows, population size variation, etc., the error rate will be moderate (~0.15 with 10,000 SNPs).

Pei et al. (2018) explored the CLAssification-based DElimitation of Species (CLADES) method that was utilized to perform species delimitation to evaluate the multilocus genomic data. There are two outcomes to using a trained classifier to classify a test sample: “same species” or “different species”. The CLADES is using maximizing the likelihood of species assignment for multiple populations to delimit species. The main advantage of CLADES is that it can accurately determine two outcomes than the Bayesian method, especially to the recent genetic isolation or to the gene flow. In this study, support vector machine (SVM) algorithms are introduced. The key of SVM is

to separate all training samples into two clusters with the least miss-classification score. The SVM regression is calculated using the training sample statistics (Salles and Domingos, 2023). The results show that when the migration parameter $M < 1$, the accuracy of the model can be more than 80%.

2.2.2 Unsupervised Machine Learning Model

Derkarabetian et al. (2019) utilized three UML methods (Random Forests (RF), Variational Autoencoders (VAE), and t-distributed Stochastic Neighbor Embedding (t-SNE)) for species delimitation. The data related to an arachnid taxon belongs to a high population genetic structure (Opiliones, Laniatores, Metanonychus). First, through combining phylogenetic analysis of mitochondrial DNA sequences and examination of morphology, the priori species hypothesis is generated. Second, through SNPs to be the database, they successfully use the above three UML methods to cluster a priori species. The results show that the UML method can successfully cluster samples based on the species-level divergences. Moreover, UML methods can offer data visualization under two-dimensional space with multiple data types. Also, Martin et al. (2021) used the same method to delimit the North American box turtles (Terrapene spp.). The results also show that these UML methods can significantly determine the species boundary.

Saryan et al. (2019) used principal component analysis (PCA), non-metric multidimensional scaling (nMDS), and spectral clustering which is t-distributed Stochastic Neighbor Embedding (t-SNE) to delimit the 16 morphological characters within the genus Hedychium to group 93 individuals from 10 taxa. The t-SNE method is a non-linear projection method, which has a better visualization result compared to the other dimensional reduction method. The results show that

there are 5, 9, and 12 clusters of taxa in the species complexes to be examined and at least 4 characters are selected to define these clusters. Cooperated with the character analysis method, the t-SNE method can significantly delimit the species boundary, which means that combining with a character selection analysis can promote the morphometric analysis and species delimitation.

Pyron et al. (2023) utilized a UML method which is called Self-Organizing (or “Kohonen”) Maps (SOM) for species delimitation. The data is related to Seal Salamanders (Desmognathus monticola). These two geographic genetic clusters have limited phenotypic divergence, demographic and spatial models of ecology and gene flow provide strong support for species delimitation. SOM is a single-layer artificial neural network that includes a random large, multidimensional feature class input vector. It can maximize the similarity between the distance matrix between the input and output and build a two-dimensional configuration of multi-dimensional data. They calculate the allele frequencies from the 2201 SNP matrix which the column has the 2-4 possible states at each SNP locus and the rows are specimens. The results show that the two genomic clusters are successfully determined. Moreover, they found a new species (D. cheaha) which is supported by multiple genetic data and analysis. Also, Pyron. (2023) provides another UML method called SuperSOM. Compared to the SOM, the input of the matrix in the SuperSOM method can be divided into multiple distinct layers and each one belongs to the output grid. The results show that the SuperSOM method includes the capacity to independently integrate multiple data types and determine a unified species delimitation model.

2.2.3 Deep Learning Model

Perez et al. (2022) used a deep learning algorithm based on Convolutional Neural Networks (CNNs) to delimit the species in the multispecies coalescent model. In order to test the accuracy of CNNs, they test the species boundary based on the precious taxonomic delimitations as well as genetic data (41 loci) in *Pilosocereus aurisetus*, which is a cactus species complex with a sky-island distribution and taxonomic uncertainty. Also, they fit another dataset related to the genus data (*Drosophila*) to validate this method. The results show that After 250 epochs, the CNN showed accuracies of 96.81% and 92.49% for the training and validation sets, respectively to the *Pilosocereus aurisetus* dataset which means that CNN has a high capacity to distinguish among the species. To the Drosophila dataset, the accuracy of CNN also can reach 80%.

Table 1: Machine Learning Applications on Species Delimitation

References	Category	Algorithms	Language	Input
Smith and Carstens, 2020	SML	Random Forests	Python	SNPs
Pei et al., 2018	SML	Support Vector Machine	R	SNPs
Smith and Carstens, 2018	SML	Random forests	Python	SNPs
Derkarabetian et al., 2019	UML	Random Forests, Variational Autoencoders, t-distributed Stochastic Neighbor Embedding	R/Python	SNPs
Pyron, 2023	UML	Self-Organizing Maps	R	SNPs
Pyron et al, 2023	UML	Super Self-Organizing Maps	R	MNPs
Martin et al., 2021	UML	Random Forests, t-distributed Stochastic Neighbor Embedding, Variational Autoencoders	Python	SNPs
Saryan et al., 2019	UML	t-distributed Stochastic Neighbor Embedding	Python	SNPs
Perez et al., 2022	DL	Convolutional Neural Networks	Python	SNPs

CHAPTER 3

DATASETS PREPARATION AND SIMULATION

3.1 Dataset Preparation

In the dataset, we determine the use of the species trees which include four animal species: chimpanzee, human, gorilla, and orangutan, and derive gene trees using the R package (Phybase) (Liu and Yu, 2010) under R studio software (Boston, Massachusetts, USA). In the simulation process, we can consider five simulation species trees (Treatment 1 to 5), the difference between treatments 1 to 5 is based on the adjusting ratio of population size (θ) and branch length of each species (B) which represents the evolutionary time, the details are in the section 3.2. In general, if the ratio is smaller than 0.1, the four species are easily delimited, and if the ratio is larger than 10, the results are opposite.

Considering we hope to explore the complicated dataset and delimit the species boundaries, we use multiple gene trees to simulate multi-locus data using Seq-gen v1.3.4 (Rambaut and Grassly, 1997) under Ubuntu 22.04.3 LTS software (London, UK). Under each treatment, we generate 10 gene trees, and each gene tree generates 100 DNA sequences (Each species contains 25 DNA sequences). Each DNA sequence is 250-based pairs in length. By combining multiple DNA sequences from gene trees, we get the multi-locus DNA sequence.

In our design, we consider generating 1-lotus, 2-lotus, 5-lotus, and 10-lotus DNA sequences under each treatment, and using different SML and UML methods to delimit species through comparing the accuracy of each machine learning method.

We separate the dataset into a training dataset and test dataset based on the ratio of 7:3. If we consider using the machine learning method to analyze the DNA sequence dataset, we need our input to be numerical vectors x and y. The acronym ACGT represents the four bases present in a DNA molecule: adenine (A), cytosine (C), guanine (G), and thymine (T). DNA is composed of two strands entwined around each other, and the bonds between the bases hold the strands together. Adenine forms pairs with thymine, while cytosine pairs with guanine. We need to do some transformations for the dataset.

One hot encoding is a technique that we use to represent categorical variables as numerical values in a machine learning model. The advantage of one hot coding is improving model performance by providing more information to the model about the categorical variable. We assume A, C, G, and T to be (1,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1).

3.2 Simulation Datasets

3.2.1 Treatment 1

In treatment 1, we build the following species tree and the code is as follows: "(((H:0.00402#0.01,C:0.00402#0.01):0.00304#0.01,G:0.00707#0.01):0.00929#0.01,O:0.01635#

0.01)#0.01;". H, C, G, and O represent four species. The number before the '#' sign represents the branch length of a species tree. In our design, the B is not changed due to the constant species which means that the key is changing θ which follows the '#' sign. When the θ is larger, the genetic divergence between two species is smaller which makes it harder to determine the boundary of species. The visualization of treatment 1 is in Figure 5.

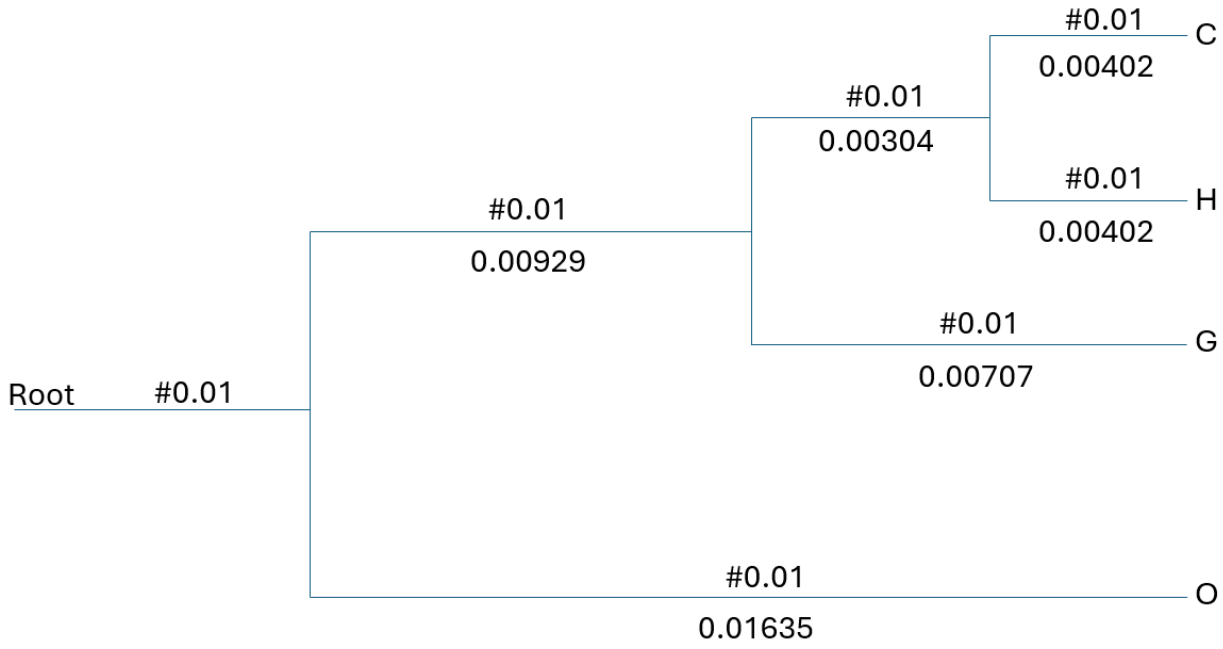


Figure 5: Visualization of species trees in treatment 1

3.2.2 Treatment 2

In treatment 2, we build the following species tree and the code is as follows: "(((H:0.00402#0.001,C:0.00402#0.001):0.00304#0.01,G:0.00707#0.01):0.00929#0.01,O:0.01635#0.01)#0.01;". Compared to treatment 1, we change the value θ of species C and species H from 0.01 to 0.001 which means that it will be easier to distinguish species C and species H compared to treatment 1. (I use red to mark the difference between treatment 1 and 2). The visualization of treatment 2 is in Figure 6.

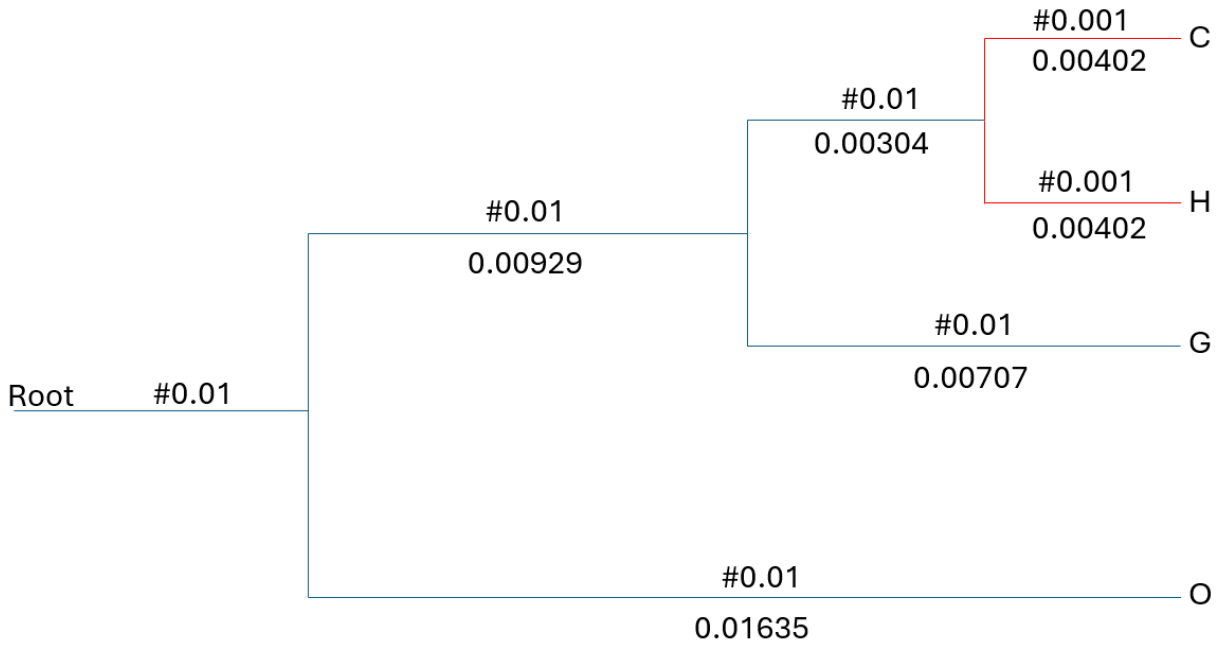


Figure 6: Visualization of species trees in treatment 2

3.2.3 Treatment 3

In treatment 3, we build the following species tree and the code is as follows: "(((H:0.00402#0.1,C:0.00402#0.1):0.00304#0.01,G:0.00707#0.01):0.00929#0.01,O:0.01635#0.01)#0.01;". Compared to treatment 1, we change the value θ of species C and species H from 0.01 to 0.1 which means that it will be harder to distinguish species C and species H compared to treatment 1. (I use red to mark the difference between treatment 1 and 2). The visualization of treatment 3 is in Figure 7.

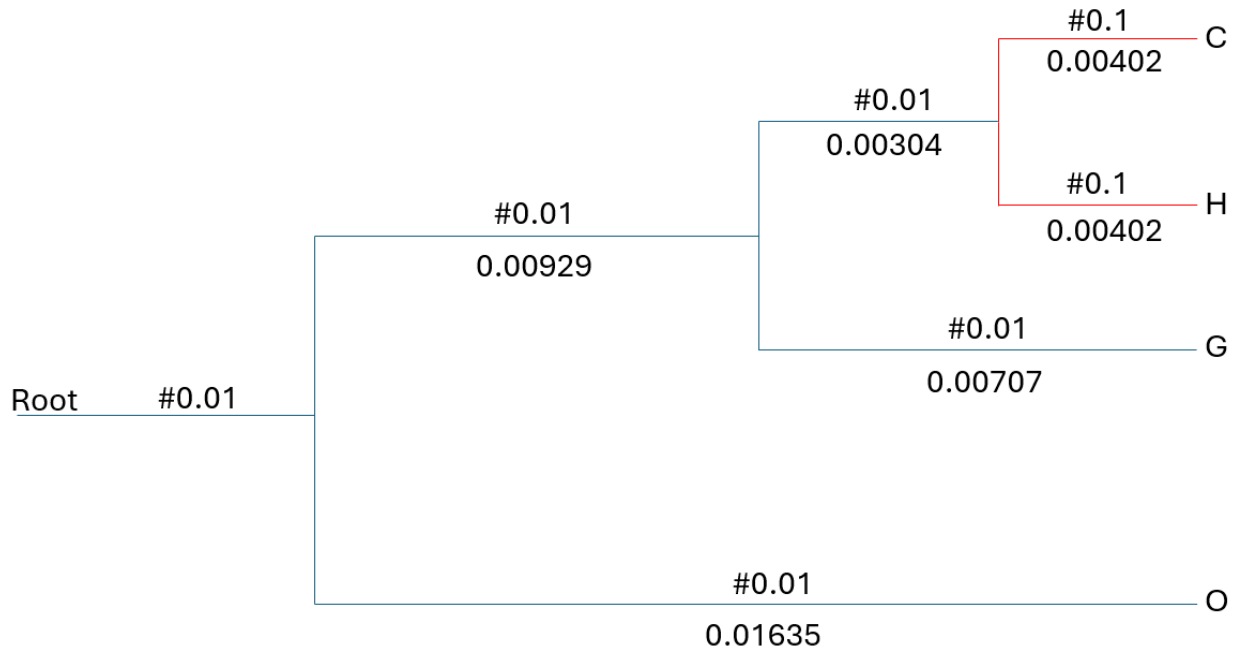


Figure 7: Visualization of species trees in treatment 3

3.2.4 Treatment 4

In treatment 4, we build the following species tree and the code is as follows: "(((H:0.00402#0.01,C:0.00402#0.01):0.00304#0.001,G:0.00707#0.001):0.00929#0.001,O:0.01635#0.001)#0.001;". Compared to treatment 1, we keep value θ of species C and species H are same, but we change the value θ to 0.001 except for the red branches. Because the value θ is smaller than treatment 1, it will be easier to distinguish species G and O from species C and H. The visualization of treatment 4 is in Figure 8.

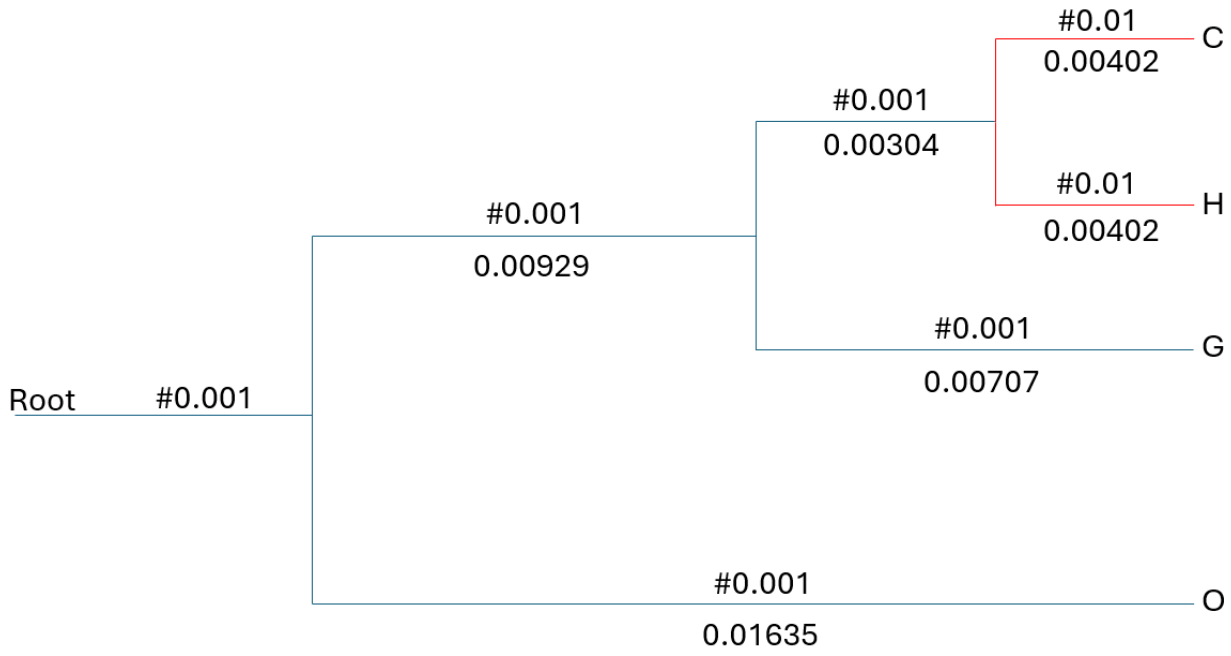


Figure 8: Visualization of species trees in treatment 4

3.2.5 Treatment 5

In treatment 5, we build the following species tree and the code is as follows: "(((H:0.00402#0.01,C:0.00402#0.01):0.00304#0.1,G:0.00707#0.1):0.00929#0.1,O:0.01635#0.1) #0.1;". Compared to treatment 1, we keep the value θ of species C and species H are same, but we change the value θ to 0.1 except for the red branches. Because the value θ is larger than treatment 1, it will be harder to distinguish species G and O from species C and H. The visualization of treatment 5 is in Figure 9.

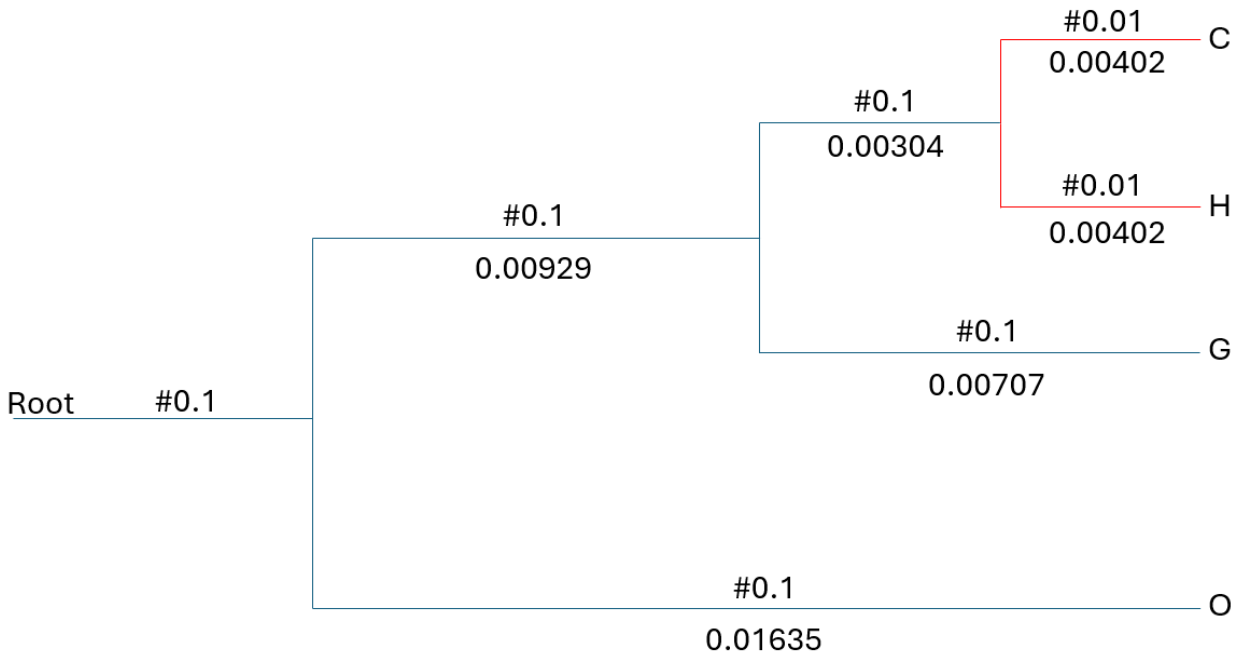


Figure 9: Visualization of species trees in treatment 5

3.3 Empirical Datasets

In addition to the simulation datasets, we randomly choose one of the empirical datasets and fit our machine learning models. The DNA sequence data is from Bank et al. (2013). *Oxystele* is a genus in the diverse marine gastropod superfamily *Trochoidea* found only in southern Africa. These mollusks are the most abundant mollusks along the southern African coastline and play a vital role in rocky intertidal habitats, making a significant contribution to coastal biodiversity. The species delimitation is controversial, especially between *Oxystele impervia,* and *Oxystele variegata*. The author provides 56 specimens and tries to delimit five molecular operational taxonomic units (*Oxystele sinensis*, *Oxystele tabularis*, *Oxystele tigrine*, *Oxystele variegate,* and *Oxystele impervia*). The sequence lengths are different. We also use one-hot encoding to make

AGCT transfer to numbers. To the unknown nucleotides in the sequence (-), we code this word as (0.5, 0.5, 0.5, 0.5) (Pang, 2021). Species trees are in Figure 10.

In the dataset, we find that each DNA sequence is not the same. We assume the DNA sequences from the literature to be our raw dataset, and we use MAFFT software to pairwise data. After that, we get the new DNA sequence and we set this dataset to be our updated dataset.

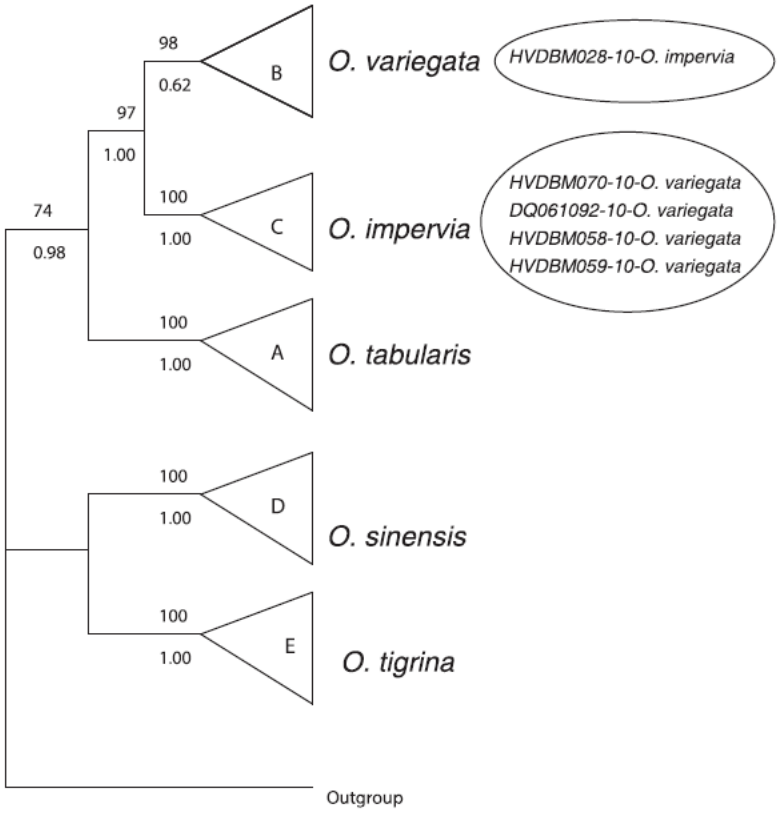


Figure 10: Visualization of the Bayesian tree and the parsimonious tree in empirical datasets. Values above branches indicate bootstrap supports; values under branches indicate posterior probability. (Inspired by Bank et al., 2013)

3.4 Data Analysis Methods

In our design, we consider using SML methods (K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Classification Tree, and Gradient Boosting (Catboost and XGboost)), UML methods (K-means Clustering), and DL (Neural Network) to fit the different DNA sequence dataset and delimit the species boundary. The machine learning analysis is utilized by Python under Anaconda Navigator (Austin, TX, USA) and R software under R studio (Boston, MA, USA).

3.4.1 K-Nearest Neighbors

The KNN is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point (Pandey and Jain, 2017). In KNN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. We used KNN classification methods to complete the classification task. In other words, KNN methods help me to group the samples into four clusters: chimpanzee, human, gorilla, and orangutan.

3.4.2 Support Vector Machine

A SVM is a type of deep learning algorithm that performs supervised learning for the classification or regression of data groups. In AI and machine learning, supervised learning systems provide both input and desired output data, which are labeled for classification. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces (Sen et al.,

2020). We utilized the training data set to train an SVM model separately and evaluate them on the test data sets by referring to the classification report and confusion matrices.

3.4.3 Classification tree

Decision trees are arranged in a hierarchical tree-like structure and are simple to understand and interpret. The decision tree method can be divided into regression and classification trees based on whether the response we need to predict is quantitative or qualitative (Malehi and Jahangiri, 2019). Here, the response we endeavor to predict is four species (chimpanzee, human, gorilla, and orangutan), so the classification tree is appropriate for this problem.

3.4.4 Gradient Boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees (Zhang and Haghani, 2015).

Catboost is an open-source software library developed by Yandex. It provides a gradient boosting framework that among other features attempts to solve for Categorical features using a permutation-driven alternative compared to the classical algorithm (Siddamsetty et al., 2021). Catboost is an implementation of gradient boosting, which uses binary decision trees as the base predictor.

XGBoost stands out as a finely tuned distributed gradient boosting library, renowned for its exceptional efficiency, adaptability, and portability (Kiangala and Wang, 2021). Within the

Gradient Boosting framework, it seamlessly implements advanced machine learning algorithms. The core of XGBoost lies in its parallel tree boosting, commonly referred to as Gradient Boosted Decision Trees or Gradient Boosting Machines. This approach efficiently addresses a myriad of data science challenges, ensuring swift and precise solutions.

3.4.5 Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters (Friedman and Meulman, 2004). The similarity between different objects can be measured by their distances. We selected the K-means clustering to be our clustering method and we employed clustering methods to complete the classification task. Based on Figure A.1, we choose the number of clusters as 4 and fit the model using simulation datasets under this condition. Based on Figure A.3, we choose the number of clusters as 5 and fit the model using the empirical dataset under this condition.

3.4.6 Neural Network

A neural network, that belongs to a machine learning program or model, uses processes by mimicking the way biological neurons work together to recognize patterns, assess alternatives, and make informed decisions (Prieto et al., 2016). The neural network consists of layers of nodes or artificial neurons, including input layers, multiple hidden layers, and output layers. The data or the input travels in one direction. The data passes through the input nodes and exits on the output nodes (Islam et al., 2019). Based on Figure A.2, we choose the number of epochs as 20 and fit the model using simulation datasets and the empirical dataset under this condition.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Dataset Introduction

In our dataset, we choose four gene lengths (250, 500, 1250, and 2500 base pairs) and 7 machine learning methods (SVM, KNN, Neural Network, classification tree, Catboost, XGboost, and K-means clustering) to be our explanatory variables. The response variable is accuracy which means that that measures how often a machine learning model correctly predicts the outcome. However, if we only generate once under each species tree, the accuracy may not be accurate because the sample size is too small. In our design, we generate 10 times under each gene length scenario and get the mean of accuracy to be our results of accuracy under each machine learning method.

4.2 Simulation Datasets

4.2.1 Treatment 1

In treatment1, based on fitting different machine learning models, the results of average accuracy utilizing different machine learning methods are shown in Table 2. Moreover, the boxplot of accuracy for different machine learning methods is shown in Figure 11. From Figure 11, we can find that except for using K-means Clustering, the range of other machine learning methods is small. Based on Table 2, gradient boosting (Catboost and XGboost) has the highest accuracy which

is larger than 0.95 at most gene length treatments. KNN and CT also have good fitting results (>0.9). To most data related to the accuracy of fitting different machine learning models, when the gene lengths increased, the accuracy also increased except for the SVM. Compared to the other machine learning models, the KC doesn't have a good fitting.

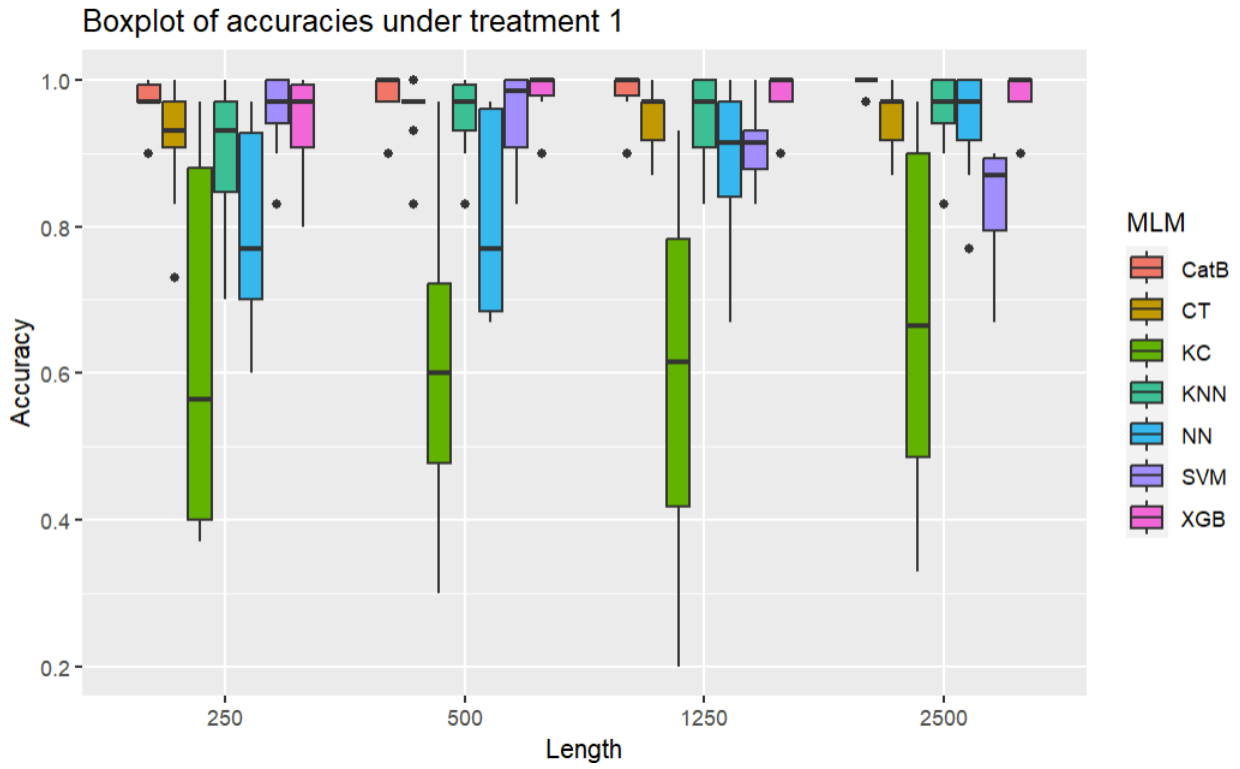


Figure 11: Boxplot of accuracies of treatment 1

Table 2: Average accuracy for treatment 1 (CatB is Catboost, KC is K-means Clustering, KNN is K-Nearest Neighbors, NN is Neural Network, CT is Classification Tree, SVM is Support Vector Machine, and XGB is XGboost)

Length	CatB	KC	KNN	NN	CT	SVM	XGB
250	0.965	0.630	0.900	0.795	0.916	0.957	0.941
500	0.981	0.594	0.950	0.812	0.958	0.953	0.984
1250	0.984	0.596	0.950	0.888	0.949	0.909	0.981
2500	0.994	0.683	0.957	0.945	0.949	0.835	0.981

4.2.2 Treatment 2

In treatment 2, Figure 12 represents the boxplot of accuracy of different machine learning models under different gene sequences and Table 3 provides the average accuracy under different treatments. From Figure 12 and Table 3, we can find that except for SVM, NN, and KC, the accuracy of the other model is close to 1 which means that it is easier to delimit the species and find the species boundary under treatment 2. When we designed treatment 2, we just changed the value θ of species C and species H from 0.01 to 0.001, which will make it easier to distinguish the species in theory. The data of accuracy matches the theory.

Table 3: Average accuracy of treatment 2

Length	CatB	KC	KNN	NN	CT	SVM	XGB
250	0.980	0.800	0.970	0.819	0.980	0.970	0.980
500	1.000	0.867	0.983	0.926	1.000	0.983	1.000
1250	1.000	0.826	0.980	0.926	1.000	0.974	1.000
2500	1.000	0.870	0.980	0.944	1.000	0.871	1.000

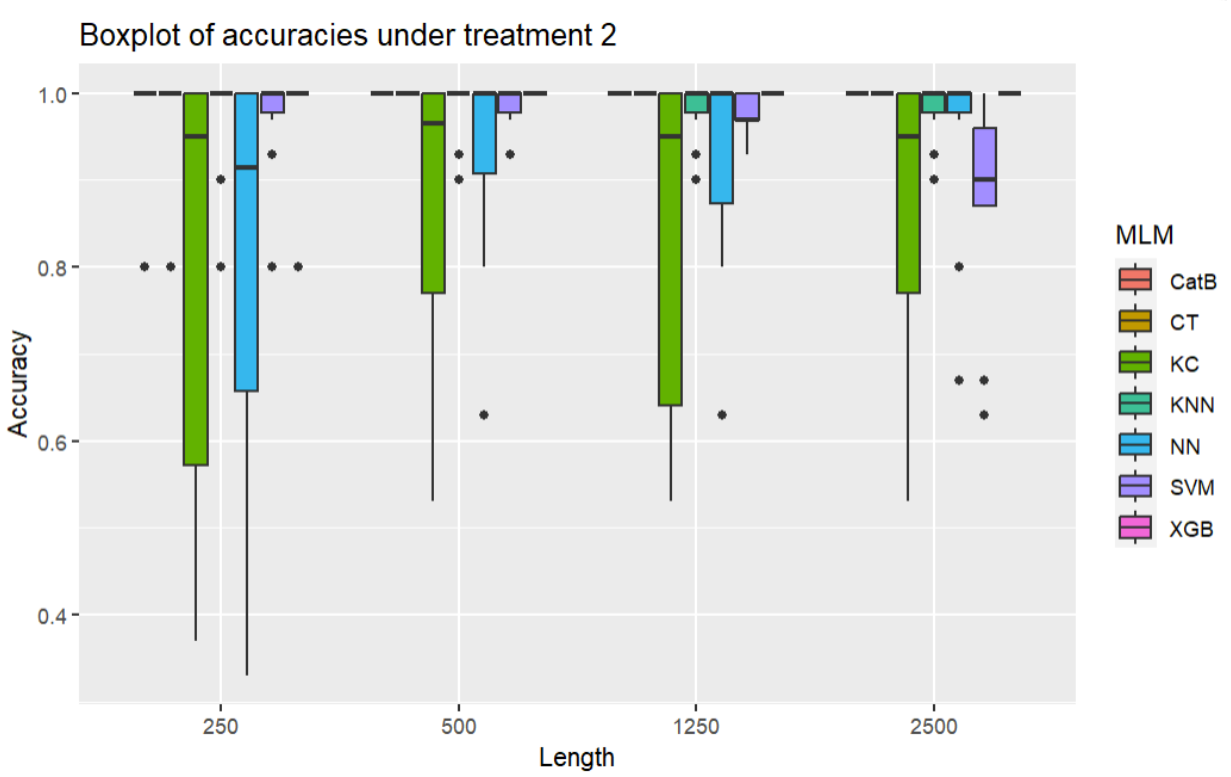


Figure 12: Boxplot of accuracies of treatment 2

4.2.3 Treatment 3

Figure 13 and Table 4 represent the boxplot of accuracy and average accuracy of different machine learning methods under different length gene sequences, separately. From Figure 13 and Table 4, we can easily find that compared to treatments 1 and 2, the accuracy is decreased. As we change the value θ of species C and species H from 0.01 to 0.1, it is very common to get this result. From Table 4, we can find that gradient boosting has the highest accuracy even though the accuracy is below 0.9. The KC's accuracy is close to 0.5 which means that the data doesn't have a good fit under this method. Except for SVM, when the gene sequence length is increased, the accuracy is increased.

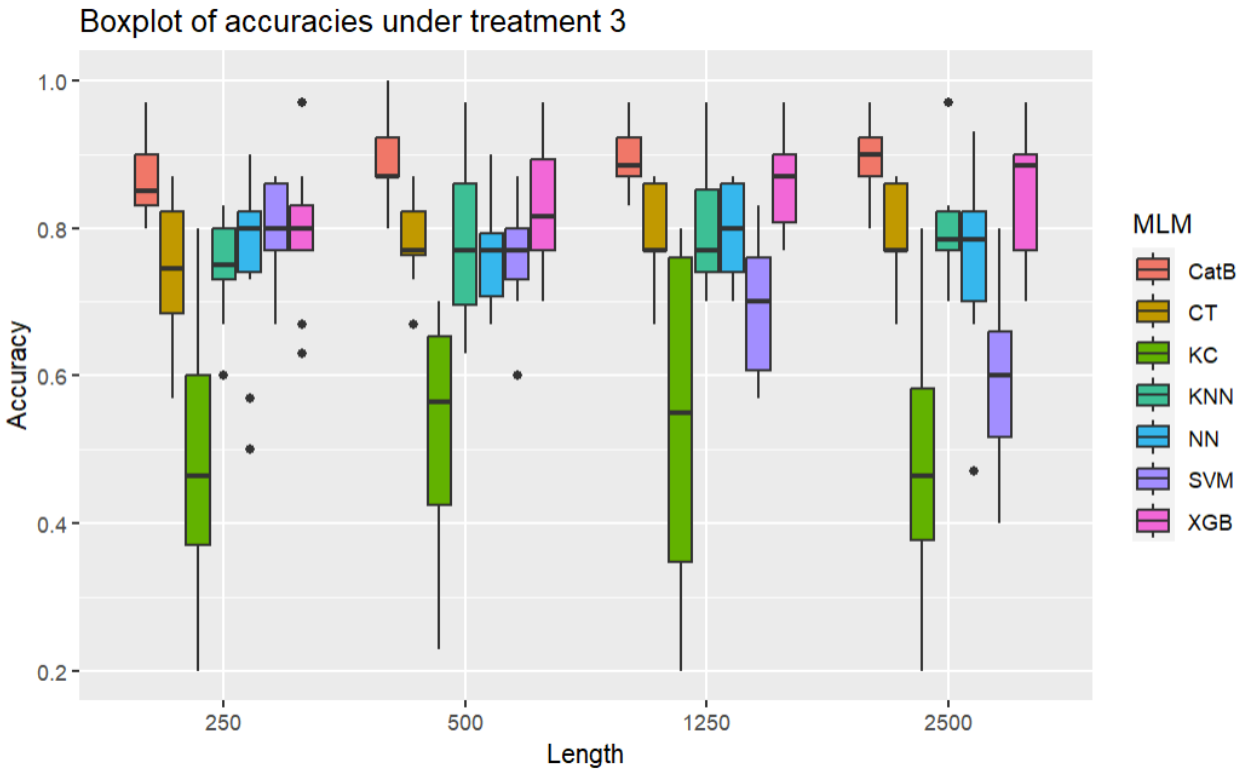


Figure 13: Boxplot of accuracies of treatment 3

Table 4: Average accuracy of treatment 3

Length	CatB	KC	KNN	NN	CT	SVM	XGB
250	0.866	0.487	0.746	0.757	0.746	0.795	0.794
500	0.887	0.526	0.785	0.764	0.784	0.764	0.827
1250	0.890	0.537	0.798	0.797	0.789	0.690	0.861
2500	0.887	0.489	0.797	0.750	0.789	0.597	0.844

4.2.4 Treatment 4

Figure 14 is the boxplot of accuracy under treatment 4, and Table 5 is the average accuracy of treatment 4 when fitting different machine learning models under different treatments. From Figure 14, we can find that the range of accuracy of KC and NN is larger than any other machine learning method. From Table 5, we can easily find that the results are very similar to treatment 1.

The accuracy is increased but not obviously. Gradient boosting also has the highest accuracy and KC's accuracy is increased from 0.6 to 0.7 compared to treatment 1.

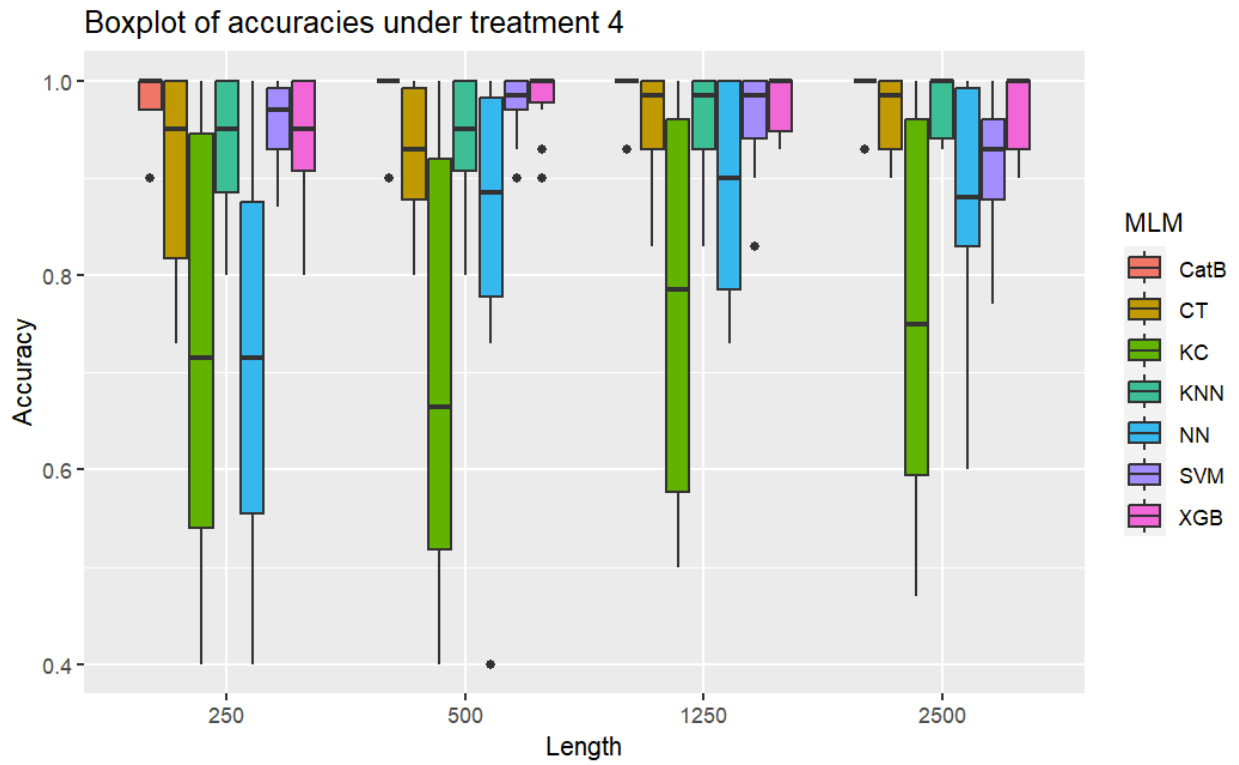


Figure 14: Boxplot of accuracies of treatment 4

Table 5: Average accuracy of treatment 4

Length	CatB	KC	KNN	NN	CT	SVM	XGB
250	0.981	0.727	0.930	0.702	0.910	0.954	0.933
500	0.990	0.704	0.940	0.840	0.923	0.974	0.980
1250	0.993	0.764	0.959	0.890	0.956	0.960	0.979
2500	0.993	0.768	0.976	0.872	0.963	0.913	0.969

4.2.5 Treatment 5

In treatment 5, compared to treatment 1, we keep the value θ of species C and species H are same, but we change the value θ to 0.1 except for the red branches. In theory, it will increase the

difficulties to delimit the species. Figure 15 and Table 6 show the boxplot of accuracy and average accuracy of different machine learning methods under different length gene sequences, separately. We can find that as the length of gene sequences increased, the accuracy also increased except for the SVM. Compared to the results of treatment 1, the accuracy is decreased which means that it is hard to find the boundary of four species which has the same answer as our design.

Table 6: Average accuracy of treatment 5

Length	CatB	KC	KNN	NN	CT	SVM	XGB
250	0.917	0.450	0.694	0.774	0.700	0.860	0.871
500	0.949	0.449	0.701	0.816	0.812	0.800	0.946
1250	0.960	0.492	0.734	0.871	0.856	0.674	0.970
2500	0.946	0.429	0.741	0.857	0.874	0.589	0.963

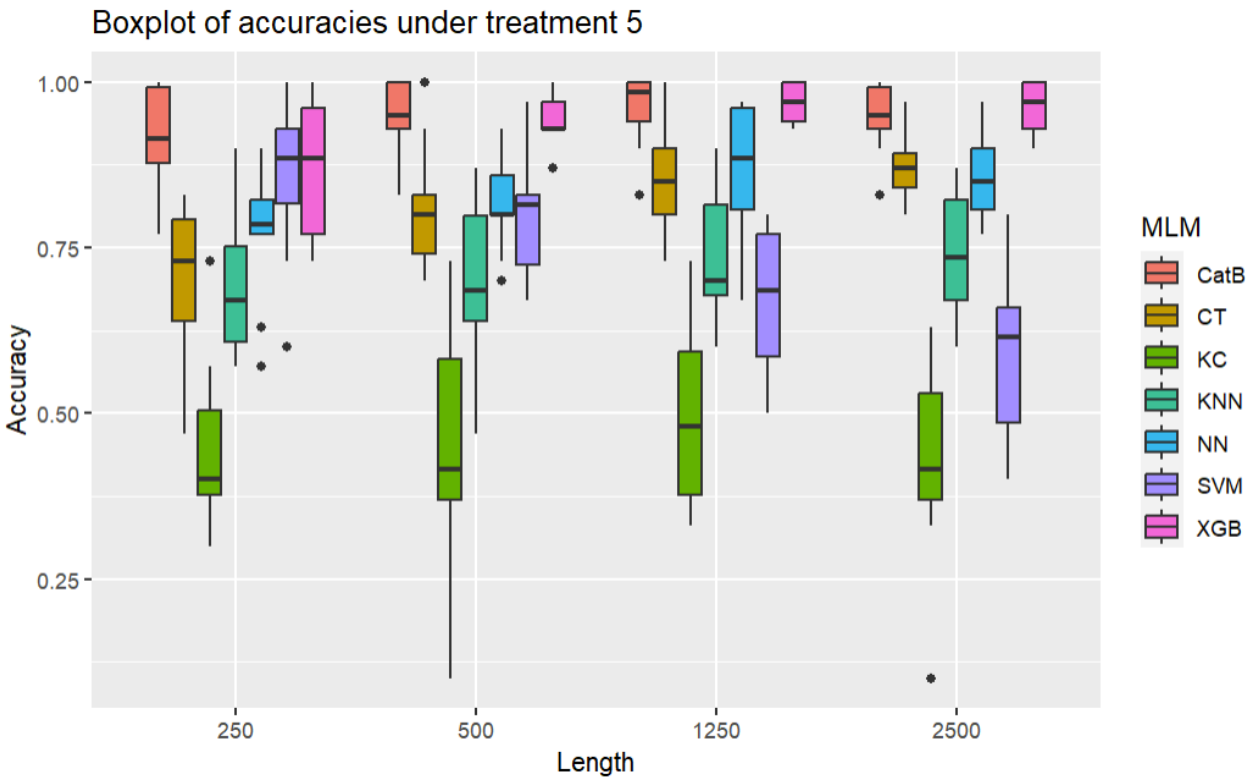


Figure 15: Boxplot of accuracies of treatment 5

4.3 Empirical Datasets

We use the *Oxystele* dataset to fit the machine learning models, and the results are shown in Table 7. From the results, we can find that Catboost has the highest accuracy. The XGboost's accuracy is below 0.5 which means that the dataset didn't have a good fit in these two models. Compared to the raw and updated dataset, we can find that after we pairwise sequence, the accuracy is increased compared to the raw dataset, especially in KC, KNN, and NN methods. However, based on the results, it has the same results from the literature that it is difficult to delimit the species of *Oxystele*.

Table 7: The accuracy of the empirical dataset

Dataset	CatB	KC	KNN	NN	CT	SVM	XGB
Raw	0.979	0.412	0.647	0.765	0.706	0.529	0.177
Updated	0.990	0.706	0.882	0.824	0.471	0.588	0.235

4.4 Discussion

From the above results, we can find that SML methods have high accuracies compared to the UML method (K-means Clustering) and DL method (Neural Network). Also, Gradient Boosting, Classification Tree, and KNN have better fitting than SVM. The gene sequences have significant effects on the results. When the gene sequences increase, the accuracy of accuracy increases in most cases because when the gene sequences increase, this means that these gene sequences include multiple lotus of data which will make it easier to find the difference between four species. However, in our design, the UML method doesn't have a good fit which means that if we don't know the species and directly use a machine learning model, it will be hard to get the clear species boundary and delimit the species. In my future work, we may determine to fit more UML methods

and explore which UML methods have a good fit in species delimitation. Furthermore, with the wide use of DL methods, fitting other deep learning models may also be a potential topic in our research.

CHAPTER 5

CONCLUSIONS

Our main goal is to compare utilizing different machine learning methods and to find which has the good results to find the boundary of four species and delimit the species. Compared to statistical model-based methods, machine learning methods are computationally highly efficient compared to the model-based methods which require long computational time and hard to solve large-size dataset problems (Pang, 2021).

In our design, we consider using SML methods (K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Classification Tree, and Gradient Boosting (Catboost and XGboost)), UML methods (K-means Clustering), and DL (Neural Network) to fit the different DNA sequence dataset and delimit the species boundary.

The five different treatments are considered in our design by changing the value θ on different branches of the species tree which can get five different 4-taxon species trees. For each species tree, we generate 10 gene trees which generate 100 gene sequences (each species has 25 gene sequences) which have 250 based pairs length. Considering we want to simulate multi-lotus data, the 1, 2, 5, and 10 lotus DNA sequences for each species tree are utilized to be our raw dataset.

Based on our design, treatments 2 and 4 will be easier to delimit the species, and treatments 3 and 5 will be harder to delimit the species compared to treatment 1. From the results of the accuracy of different machine learning models, the results are the same as our design which means the results are reasonable.

From Table 2 to Table 6, we can find that the UML method (K-means Clustering) does not have high accuracy compared to the SML methods (Gradient Boosting (Catboost and XGboost), SVM, KNN, and Classification Tree). Among SML methods, Gradient Boosting has the highest accuracy in all five treatments compared to the other methods.

In order to validate our methods, we randomly choose one of the empirical datasets (*Oxystele*) and fit our machine learning models. Because the DNA sequences are not the same length. We determine to pairwise the dataset. The results showed that after we do the pairwise, the accuracy is increased, especially in NN, KNN, and KM.

REFERENCES

Avise, J.C., 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Oxford surveys in evolutionary biology, 7, pp.45-67.

Chen, C.P. and Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information sciences, 275, pp.314-347.

Dash, M., Liu, H. and Yao, J., 1997, November. Dimensionality reduction of unsupervised data. In Proceedings ninth ieee international conference on tools with artificial intelligence (pp. 532-539). IEEE.

Deng, L., 2014. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA transactions on Signal and Information Processing, 3, p.e2.

Derkarabetian, S., Castillo, S., Koo, P.K., Ovchinnikov, S. and Hedin, M., 2019. A demonstration of unsupervised machine learning in species delimitation. Molecular phylogenetics and evolution, 139, p.106562.

Edwards, S.V., Xi, Z., Janke, A., Faircloth, B.C., McCormack, J.E., Glenn, T.C., Zhong, B., Wu, S., Lemmon, E.M., Lemmon, A.R. and Leaché, A.D., 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Molecular phylogenetics and evolution, 94, pp.447-462.

El Naqa, I. and Murphy, M.J., 2015. What is machine learning? (pp. 3-11). Springer International Publishing.

Friedman, J.H. and Meulman, J.J., 2004. Clustering objects on subsets of attributes (with discussion). Journal of the Royal Statistical Society Series B: Statistical Methodology, 66(4), pp.815-849.

Fujisawa, T. and Barraclough, T.G., 2013. Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. Systematic biology, 62(5), pp.707-724.

Grimmer, J., Roberts, M.E. and Stewart, B.M., 2021. Machine learning for social science: An agnostic approach. Annual Review of Political Science, 24, pp.395-419.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S., 2016. Deep learning for visual understanding: A review. Neurocomputing, 187, pp.27-48.

Hebert, P.D., Cywinska, A., Ball, S.L. and DeWaard, J.R., 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1512), pp.313-321.

Heled, J. and Drummond, A.J., 2009. Bayesian inference of species trees from multilocus data. Molecular biology and evolution, 27(3), pp.570-580.

Islam, M., Chen, G. and Jin, S., 2019. An overview of neural network. American Journal of Neural Networks and Applications, 5(1), pp.7-11.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A. and Flouri, T., 2017. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. Bioinformatics, 33(11), pp.1630-1638.

Kiangala, S.K. and Wang, Z., 2021. An effective adaptive customization framework for small

manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. Machine Learning with Applications, 4, p.100024.

Liu, L. and Yu, L., 2010. Phybase: an R package for species tree analysis. Bioinformatics, 26(7), pp.962-963.

Malehi, A.S. and Jahangiri, M., 2019. Classic and bayesian tree-based methods. In Enhanced Expert Systems (pp. 1-2). IntechOpen.

Martin, B.T., Chafin, T.K., Douglas, M.R., Placyk Jr, J.S., Birkhead, R.D., Phillips, C.A. and Douglas, M.E., 2021. The choices we make and the impacts they have: machine learning and species delimitation in North American box turtles (Terrapene spp.). Molecular ecology resources, 21(8), pp.2801-2817.

Mitchell, T.M., 2006. The discipline of machine learning (Vol. 9, p. 3). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Naeem, S., Ali, A., Anam, S. and Ahmed, M.M., 2023. An Unsupervised Machine Learning Algorithms: Comprehensive Review. Int. J. Comput. Digit. Syst.

Nasteski V. An overview of the supervised machine learning methods. Horizons. b. 2017 Dec 1;4:51-62.

Ntampaka, M., Avestruz, C., Boada, S., Caldeira, J., Cisewski-Kehe, J., Di Stefano, R., Dvorkin, C., Evrard, A.E., Farahi, A., Finkbeiner, D. and Genel, S., 2019. The role of machine learning in the next decade of cosmology. arXiv preprint arXiv:1902.10159.

Pandey, A. and Jain, A., 2017. Comparative analysis of KNN algorithm using various normalization techniques. International Journal of Computer Network and Information Security, 11(11), p.36.

Pang, Q., 2021. Species Delimitation by Neural Networks Using Multi-locus Sequence Data (Doctoral dissertation, University of Georgia).

Pei, J., Chu, C., Li, X., Lu, B. and Wu, Y., 2018. CLADES: A classification-based machine learning method for species delimitation from population genetic data. Molecular ecology resources, 18(5), pp.1144-1156.

Perez, M.F., Bonatelli, I.A., Romeiro-Brito, M., Franco, F.F., Taylor, N.P., Zappi, D.C. and Moraes, E.M., 2022. Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. Molecular Ecology Resources, 22(3), pp.1016-1028.

Pinheiro, F., Dantas-Queiroz, M.V. and Palma-Silva, C., 2018. Plant species complexes as models to understand speciation and evolution: a review of South American studies. Critical Reviews in Plant Sciences, 37(1), pp.54-80.

Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. and Vogler, A.P., 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Systematic biology, 55(4), pp.595-609.

Prieto, A., Prieto, B., Ortigosa, E.M., Ros, E., Pelayo, F., Ortega, J. and Rojas, I., 2016. Neural networks: An overview of early research, current frameworks and new challenges. Neurocomputing, 214, pp.242-268.

Puillandre, N., Lambert, A., Brouillet, S. and Achaz, G.J.M.E., 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular ecology, 21(8), pp.1864-1877.

Pyron, R.A., 2023. Unsupervised machine learning for species delimitation, integrative taxonomy, and biodiversity conservation. Molecular Phylogenetics and Evolution, 189, p.107939.

Pyron, R.A., O'Connell, K.A., Duncan, S.C., Burbrink, F.T. and Beamer, D.A., 2023. Speciation hypotheses from phylogeographic delimitation yield an integrative taxonomy for Seal

Salamanders (Desmognathus monticola). Systematic Biology, 72(1), pp.179-197.

Rambaut, A. and Grass, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics, 13(3), pp.235-238.

Rannala, B. and Yang, Z., 2020. Species delimitation. Self published.

Salles, M. and Domingos, F., 2023. Towards the next generation of species delimitation methods: an overview of Machine Learning applications.

Saryan, P., Gupta, S. and Gowda, V., 2020. Species complex delimitations in the genus Hedychium: A machine learning approach for cluster discovery. Applications in Plant Sciences, 8(7), p.e11377.

Sen, P.C., Hajra, M. and Ghosh, M., 2020. Supervised classification algorithms in machine learning: A survey and review. In Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018 (pp. 99-111). Springer Singapore.

Siddamsetty, S., Vangala, R.R., Reddy, L. and Vattipally, P.R., 2021. Restaurant Revenue Prediction using Machine Learning. International Research Journal of Engineering and Technology (IRJET) e-ISSN, pp.2395-0056.

Sinaga, K.P. and Yang, M.S., 2020. Unsupervised K-means clustering algorithm. IEEE access, 8, pp.80716-80727.

Smith, M.L. and Carstens, B.C., 2018. Disentangling the process of speciation using machine learning. BioRxiv, p.356345.

Smith, M.L. and Carstens, B.C., 2020. Process-based species delimitation leads to identification of more biologically relevant species. Evolution, 74(2), pp.216-229.

Tarca, A.L., Carey, V.J., Chen, X.W., Romero, R. and Drăghici, S., 2007. Machine learning and its applications to biology. PLoS computational biology, 3(6), p.e116.

Van Der Bank, H., Herbert, D., Greenfield, R. and Yessoufou, K., 2013. Revisiting species delimitation within the genus Oxystele using DNA barcoding approach. ZooKeys, (365), p.337.

Yang, Z., 2015. The BPP program for species tree estimation and species delimitation. Current Zoology, 61(5), pp.854-865.

Zapata, F. and Jiménez, I., 2012. Species delimitation: inferring gaps in morphology across geography. Systematic biology, 61(2), p.179.

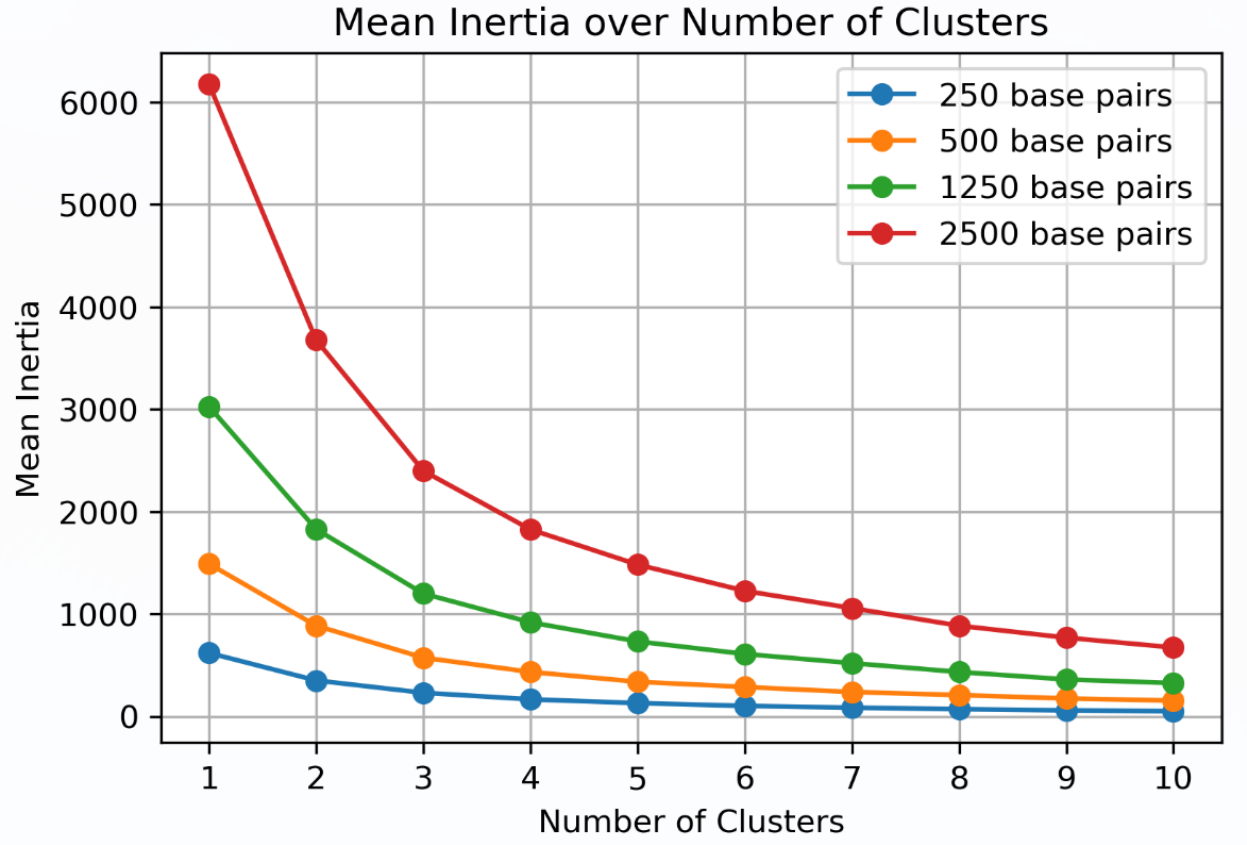
Zhang, C., Zhang, D.X., Zhu, T. and Yang, Z., 2011. Evaluation of a Bayesian coalescent method of species delimitation. Systematic biology, 60(6), pp.747-761.

Zhang, J., Kapli, P., Pavlidis, P. and Stamatakis, A., 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics, 29(22), pp.2869-2876.

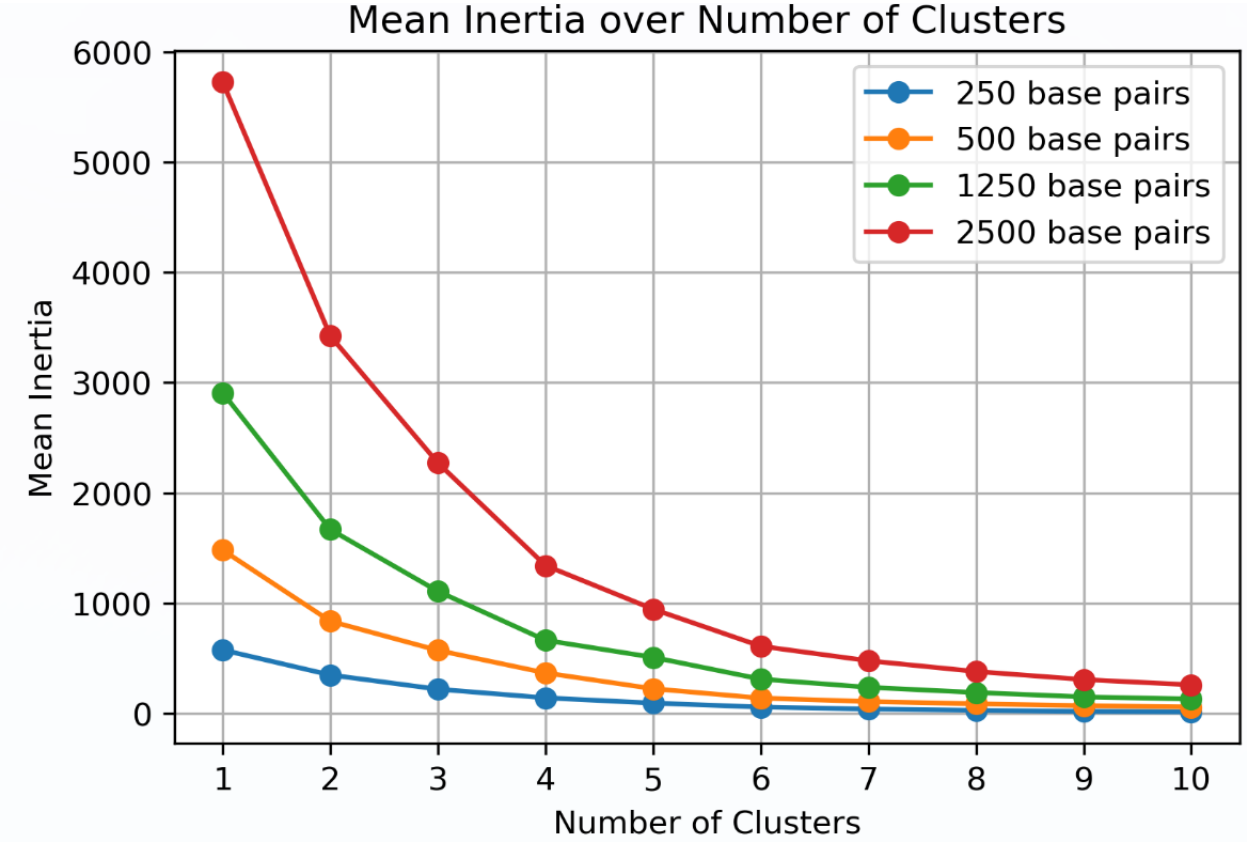
Zhang, Y. and Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, pp.308-324.

APPENDICES

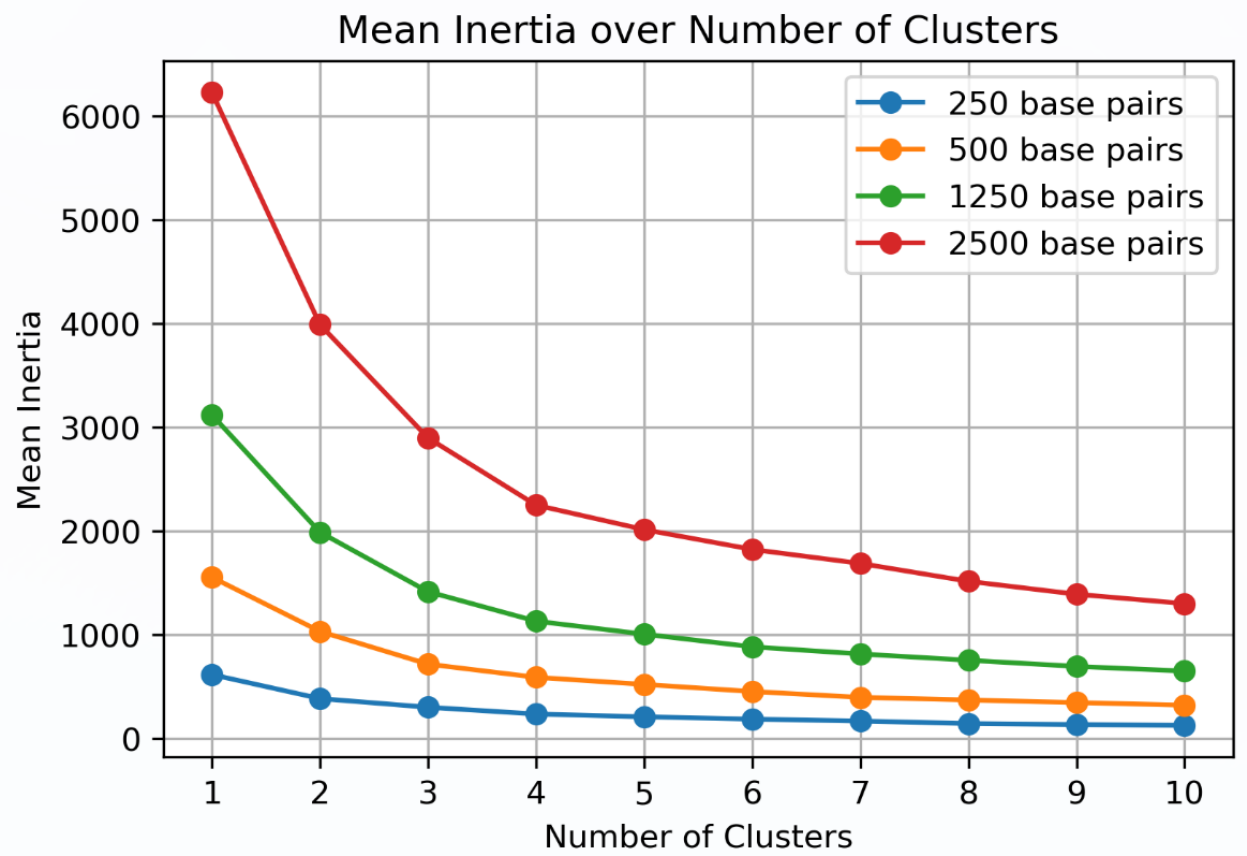
A. Simulation Datasets



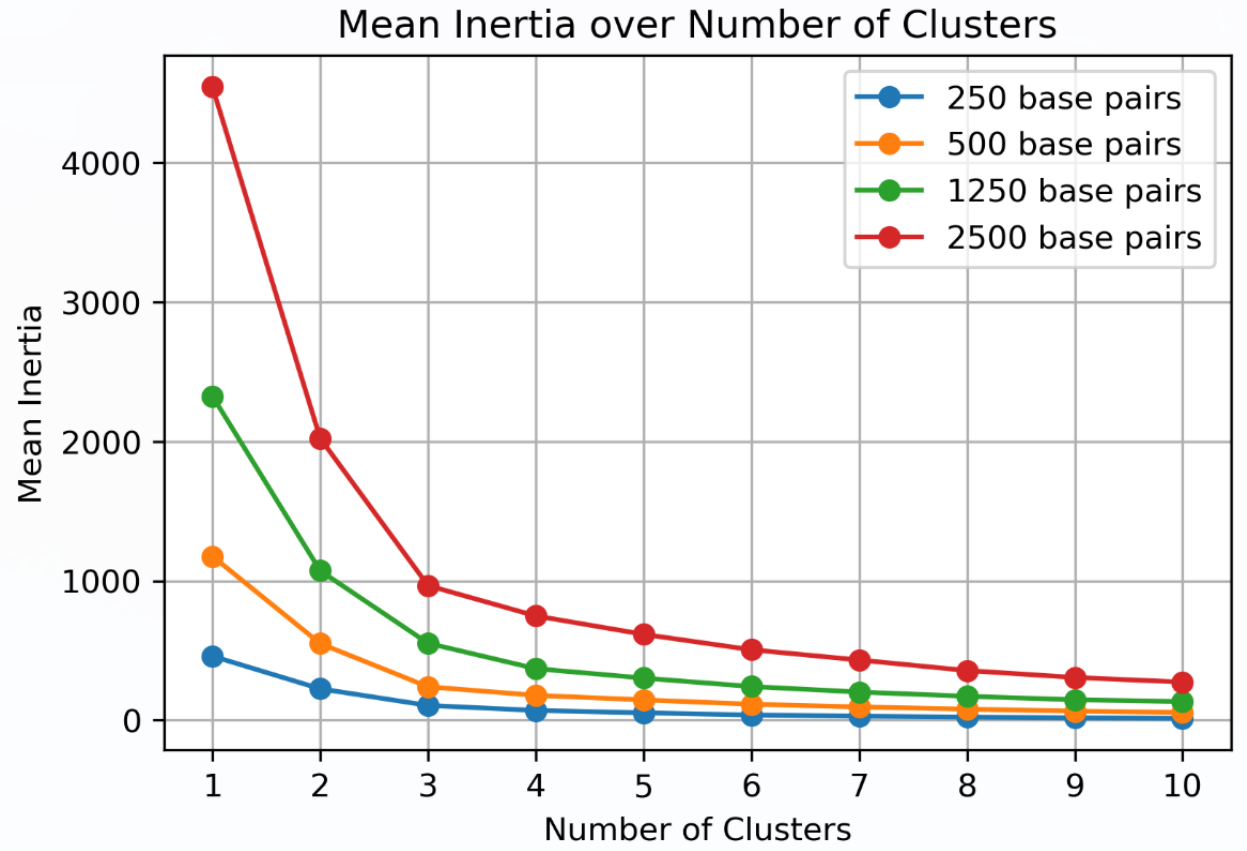
(a) Treatment 1



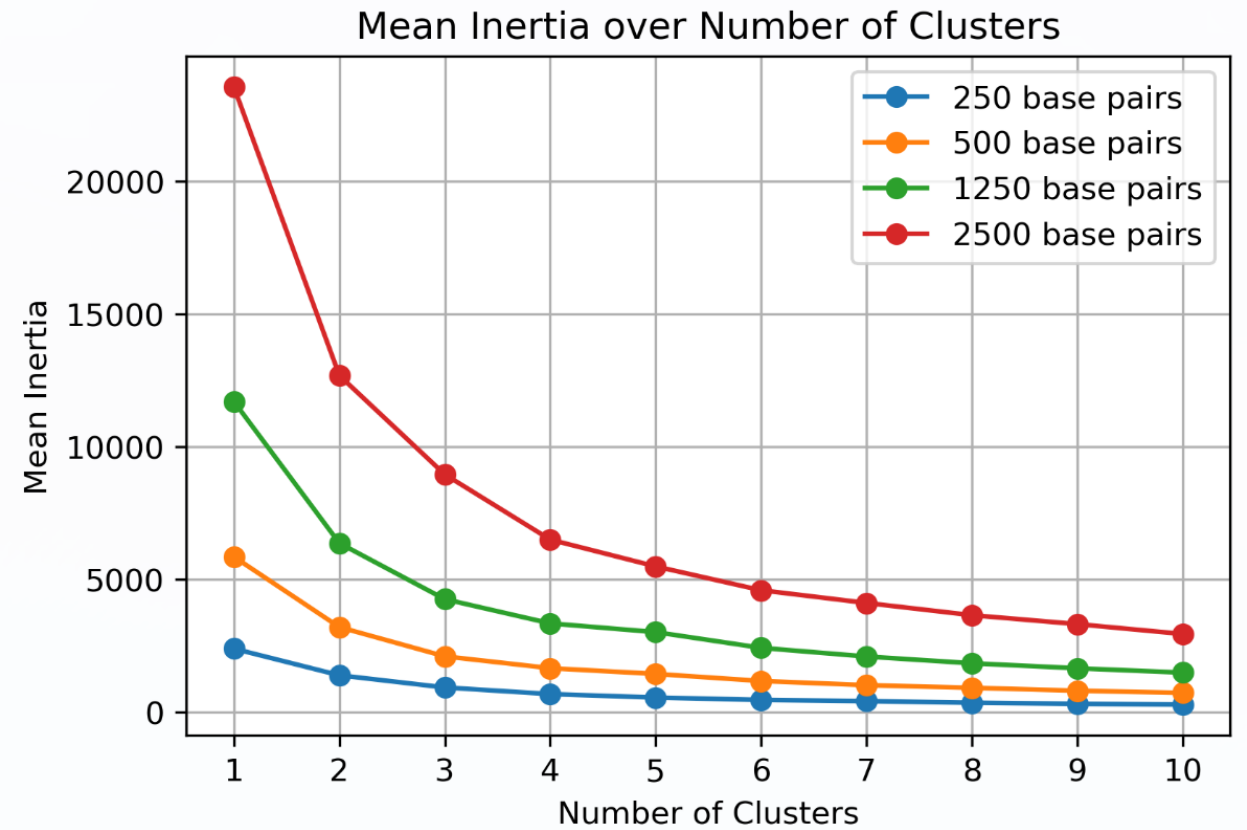
(b) Treatment 2



(c) Treatment 3

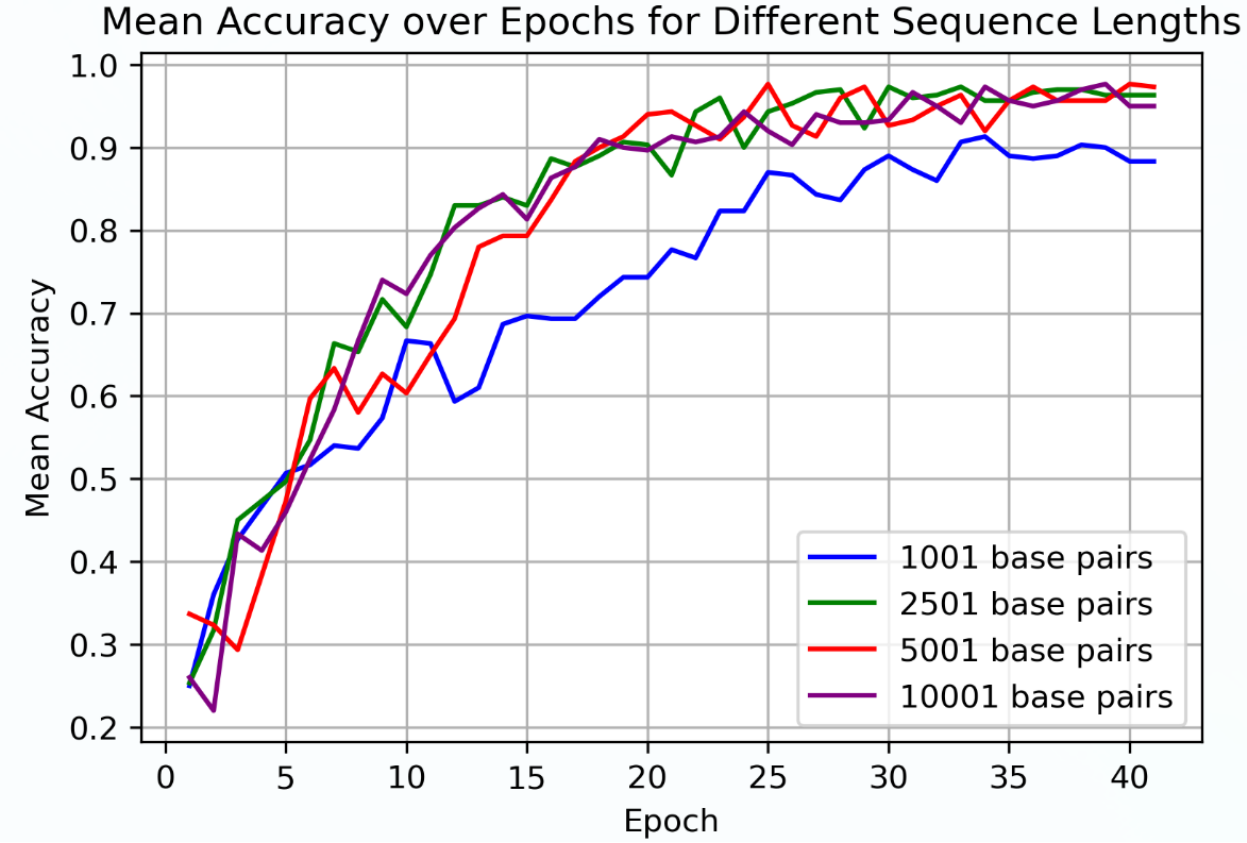


(d) Treatment 4

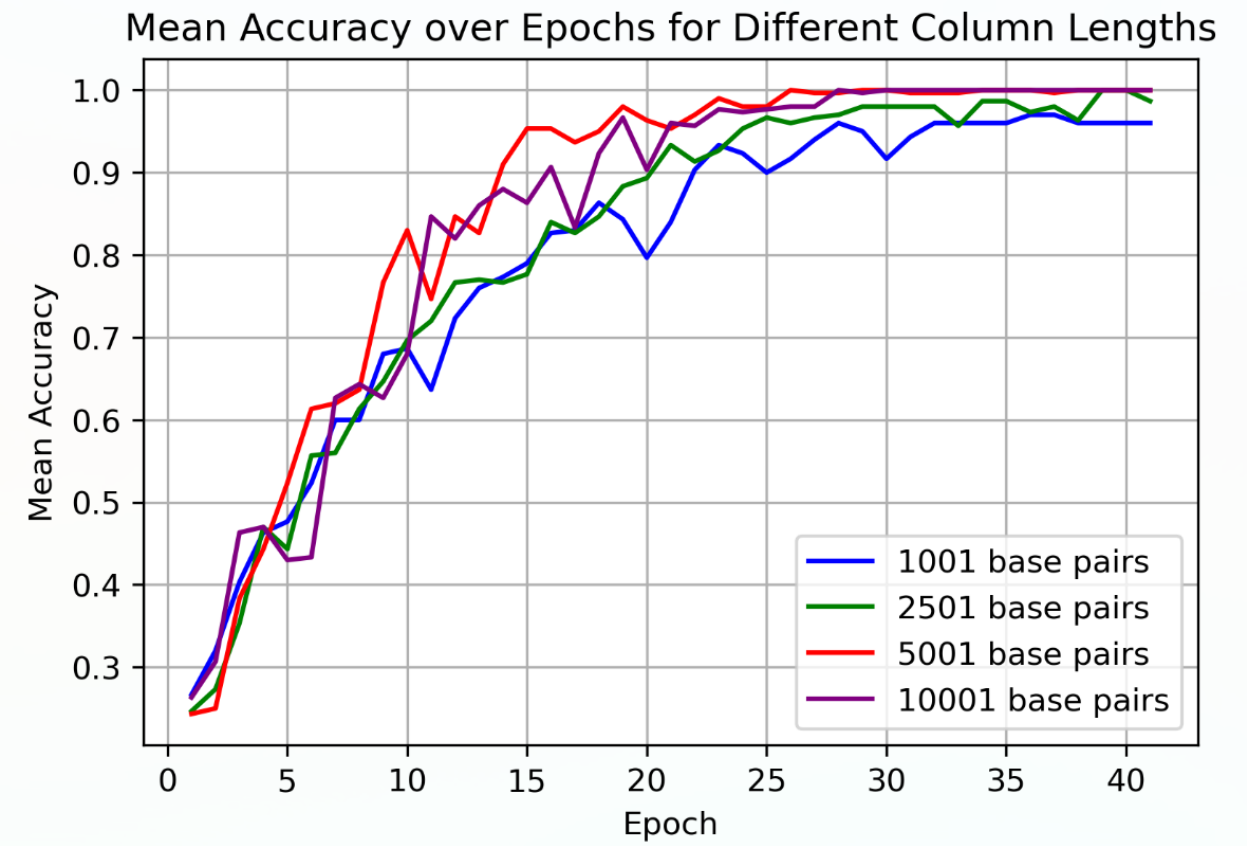


(e) Treatment 5

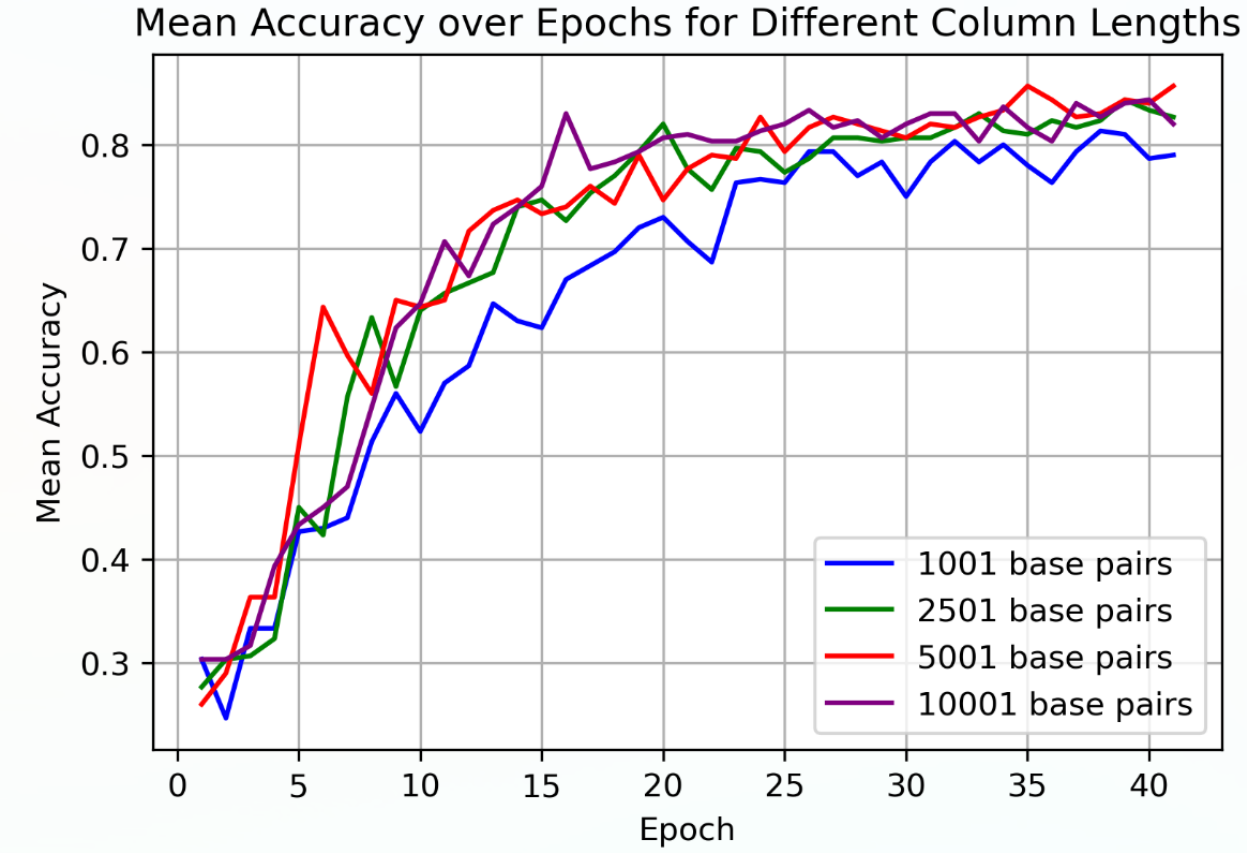
Figure A.1: Mean inertia values for each iteration under KC in simulation datasets. (a)-(e) are treatment 1-5.



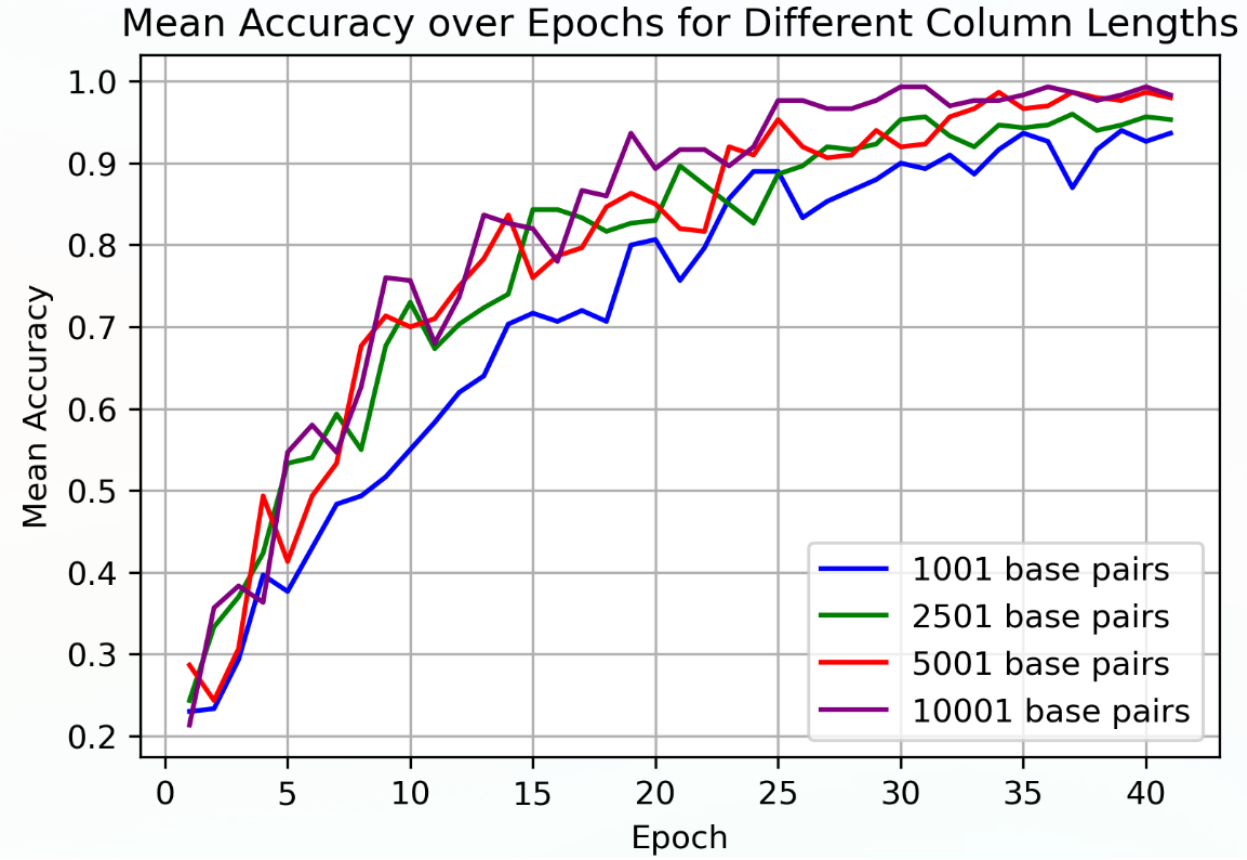
(a) Treatment 1



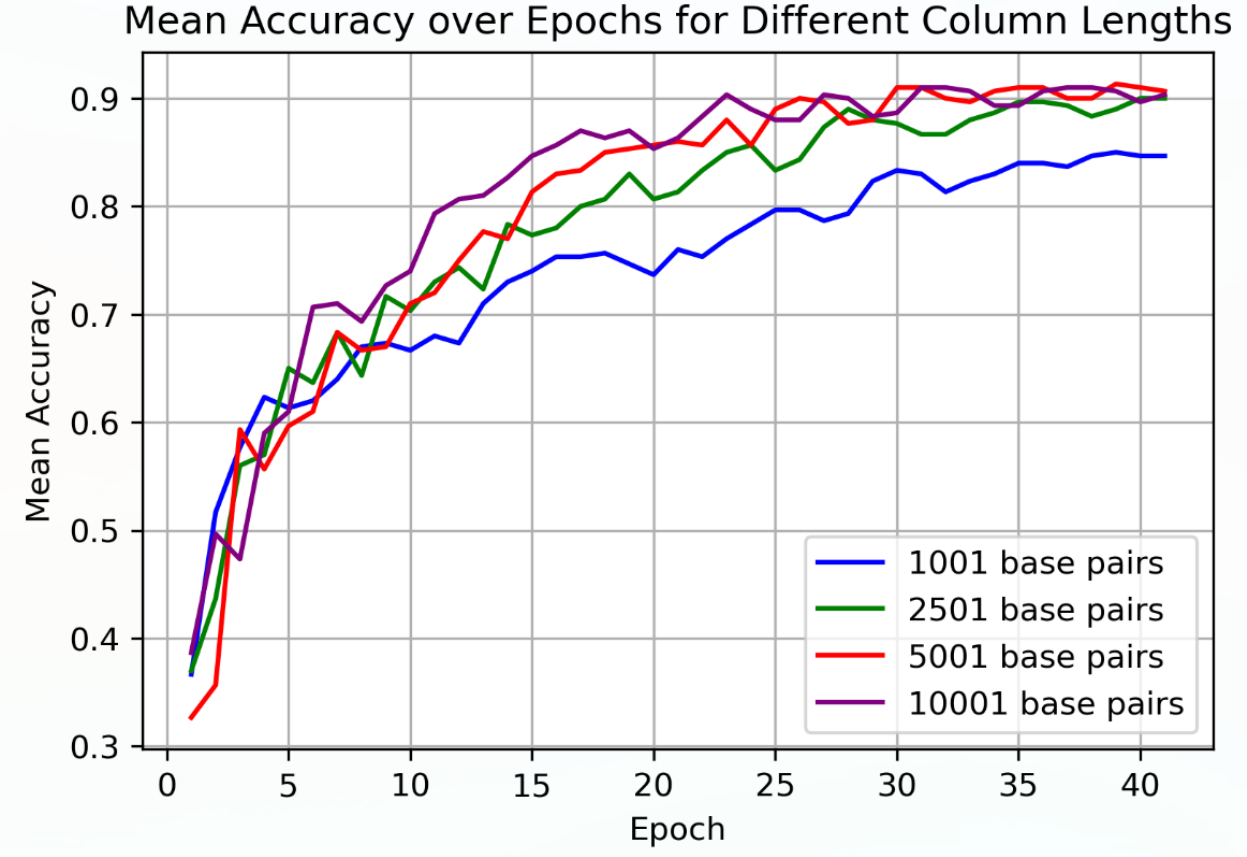
(b) Treatment 2



(c) Treatment 3



(d) Treatment 4



(e) Treatment 5

Figure A.2: Mean accuracies for each epoch NN in simulation datasets. (a)-(e) are treatment 1-5.

B. Empirical Datasets

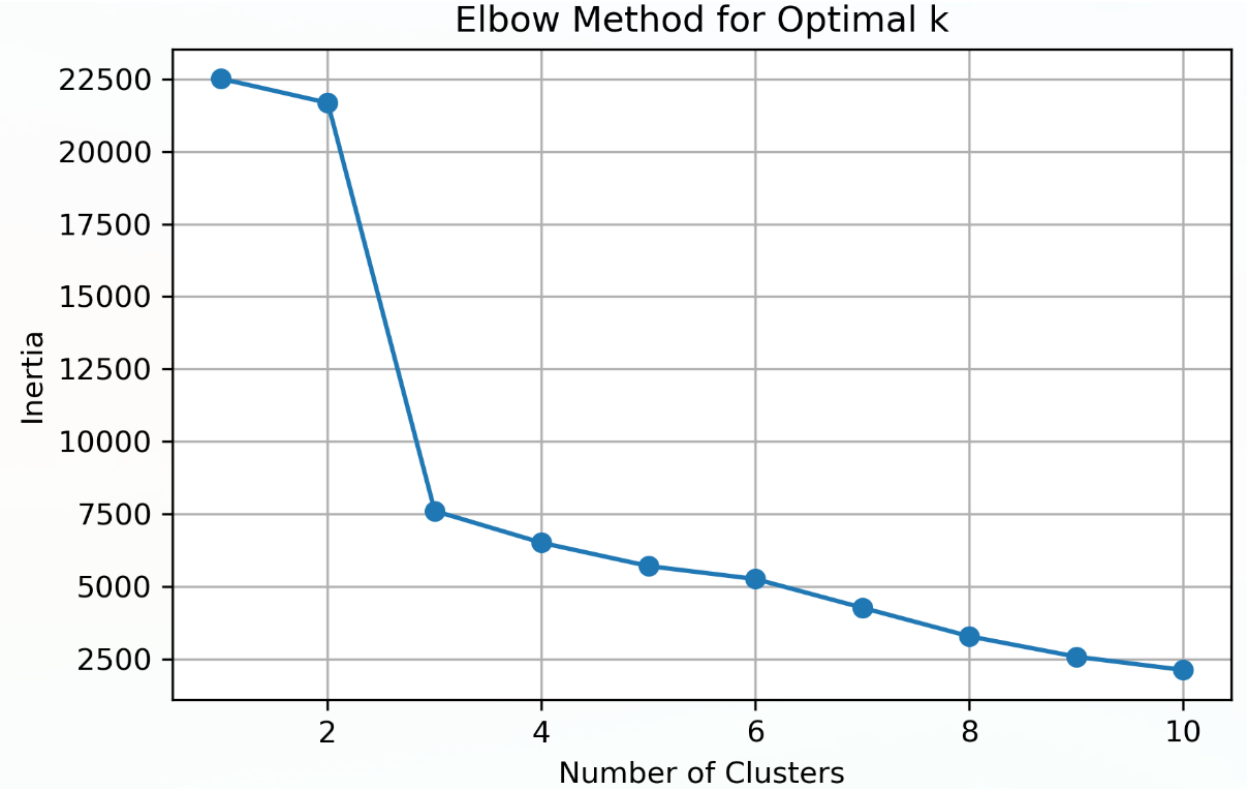


Figure A.3: Inertia values for each iteration under KC in the empirical raw dataset.

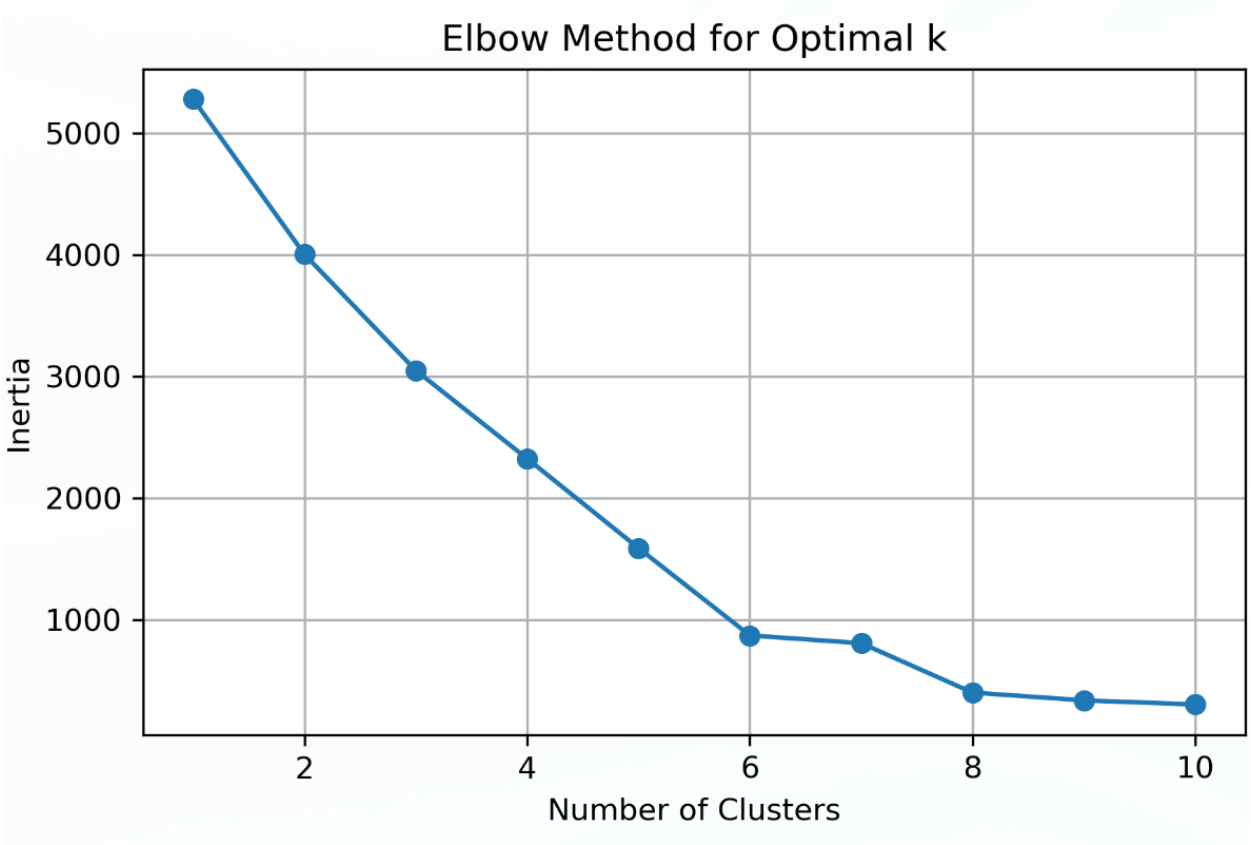


Figure A.4: Inertia values for each iteration under KC in the empirical updated dataset.

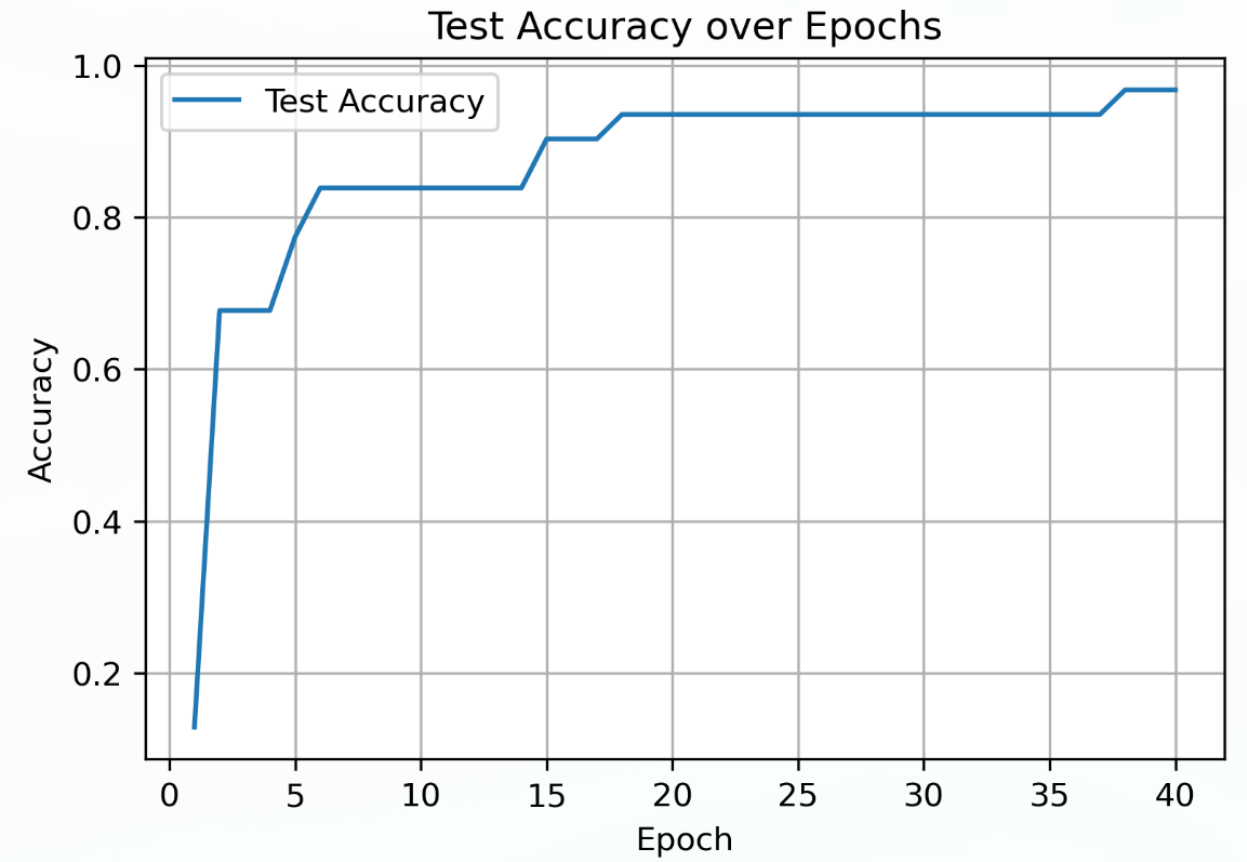


Figure A.5: Accuracies for each epoch NN in the empirical raw dataset.

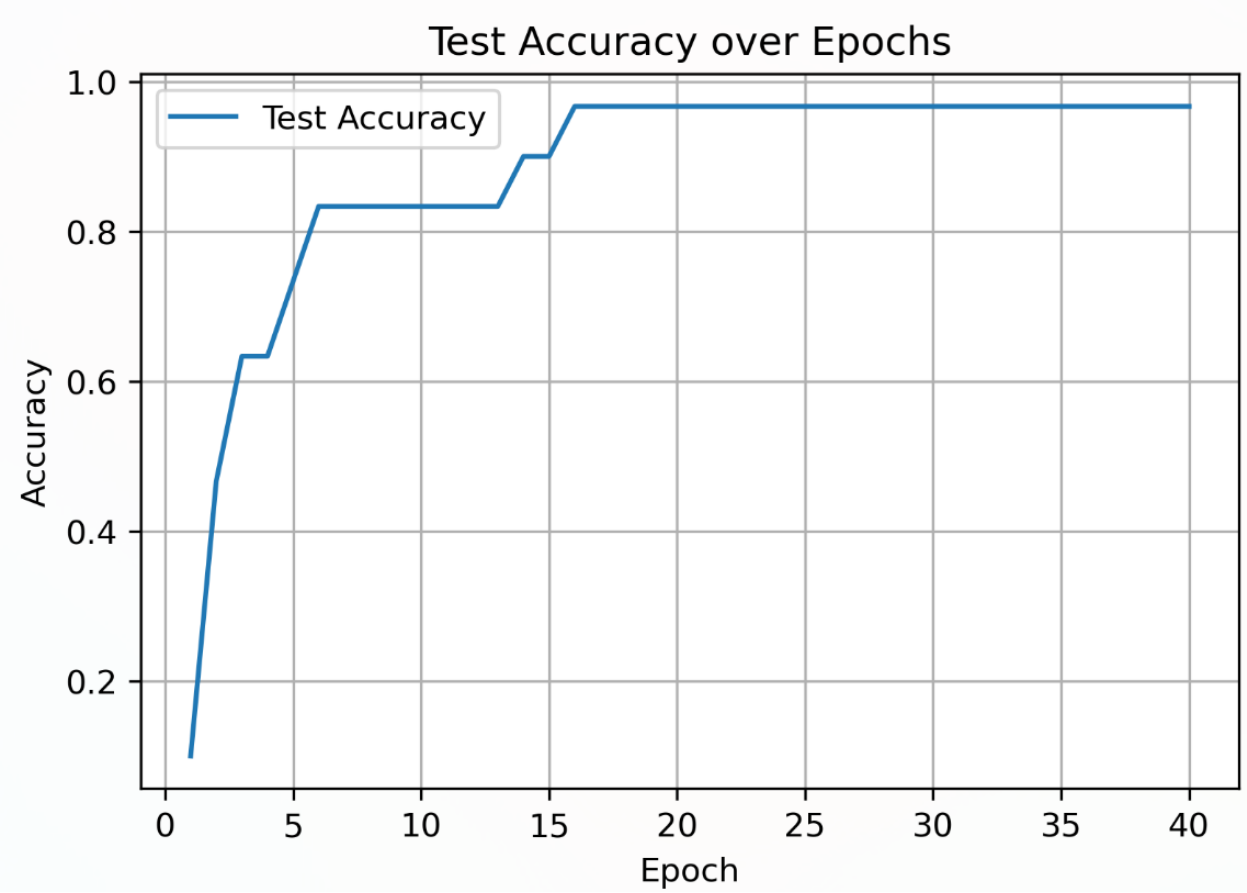


Figure A.6: Accuracies for each epoch NN in the empirical updated dataset.

C. Code

The GitHub link related to the Code is as follows:

https://github.com/jw94216/UGA-Thesis.git