Accuracy Responses in Species Identification varying DNA Barcode lengths with a Naïve Bayes Classifier: Efficacy of Mini-Barcode under A Supervised Machine Learning approach

Mohimenul Karim* and Md. Rashid Abid†
Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology,
Dhaka, Bangladesh.
Email: *1105097.mk@ugrad.cse.buet.ac.bd, †1018052031@grad.cse.buet.ac.bd

***Abstract*—Specific gene regions in DNA, like COI in case of animals, have been defined as DNA barcode and many studies proved that it can be used as an identifier to distinguish species. The standard length of a DNA barcode is approximately 650 bp. But because of the challenges in sequencing technologies or unavailability of high-quality genomic DNA, it is not always possible to get the full barcode sequence of an organism. As a result, recent studies suggest that mini-barcodes can provide a good contribution in the species identification process. Among various methods proposed for the identification task, supervised machine learning methods have been shown effective. In this study, we have analyzed the effect of different barcode lengths on species identification from the perspective of supervised machine learning and suggested a general approximation of required length of mini-barcode in this regard. We have implemented a Naïve Bayes classifier as our model and implied the effectiveness of mini-barcode by demonstrating the accuracy responses varying the length of DNA barcode sequences.**

***Index Terms*—DNA Barcoding; Species identification; Mini-barcode; Barcode length; Supervised Learning; Naïve Bayes**

I. Introduction

Species identification is naturally an elementary part in biological researches. According to the Catalogue of Life (CoL), which is a hierarchically ranked system based on expert and coherent opinions provided by more than 3,000 taxonomists, more than 1.6 million species have been identified throughout the world so far [1]. Still, the total number of species is estimated to be 8.7 millions and it is suggested that in spite of over 250 years of taxonomic classification, the existing records tend to cover only a small portion of species on earth (14%) and in the ocean (9%) while the rest are yet to be described [2]. From the taxonomic point of view, the process of species discovery can be said to be broadly comprised of three pivotal but onerous steps: specimen collection, species level sorting and lastly species identification/description [3]. In addition to being already exigent by virtue of their own nature of work and thereby creating taxonomic impediments, these steps bring morphological characteristics into account in the traditional 'evidence-based' taxonomic approach and thus inflict likewise challenges in the decision making process [4]. For instance, in the specimen collection phase due to adverse situations often organism-fragments, stomach contents, feces, saliva, skin-scraps, blood, pollen etc can be achievable as samples which lack substantial morphological characteristics. Additionally, in case of archival museum specimens, distinguishing morphological details are often absent. Also in the next phase of species level sorting, it is observed that many congeneric species possess morphological differences so negligible that they are often misclassified as a single species. Furthermore, in cases of species-complex and cryptic species, morphological differences are almost indiscernible. It might also be the case that due to sexual dimorphism, two sexes of same species can exhibit significantly disparate morphological attributes misleadingly appearing to be of different species. Various organisms can also develop conspicuous morphological changes in different stages of their life-cycle.

For addressing all the aforementioned issues, a more advanced, accurate, fast and efficient alternative approach in taxonomy was long yearned for which would attenuate morphological subservience in such cases. It is deemed that microgenomic approach involving DNA sequences has been successfully established as an extremely promising technique in recent years for species identification in this regard [5]. In this technique, a short portion of DNA from a specific gene or genes of an organism is used to uniquely identify a species in comparison with a reference library [6]. That specific portion is referred to as *DNA barcode* of that species as the process is similar to the process of identifying retail products with the help of machine readable barcodes. Different gene regions are used as DNA barcodes for different groups, namely: cytochrome-c-oxidase-I (*COI*/*COX1*) gene found in mtDNA for animals [7], *rbcL* and *matK* for plants [8] and the Internal Transcribed Spacer (*ITS*) rRNA for fungi [9]. Species identification with DNA Barcode can be treated as a classification problem where an unknown DNA barcode sequence is matched against a reference library comprising of DNA barcode sequences of already known species [10]. Various methods have been proposed for addressing this problem [11]

and supervised machine learning approach is a prominent one among them where the reference dataset of known species is used as *Train Set* for training corresponding machine learning models and barcode sequences requiring classifications are placed in *Test Set* for query [12].

Notwithstanding the prominence of DNA barcode in species identification, several shortcomings are however perceived in some specific cases. Although PCR amplification and sequencing is usually consistent in freshly collected and well-preserved specimens, obtaining a full-length ($\sim$650 bp) barcode in case of archival museum specimens which are preserved under sub-optimal or DNA-unfriendly conditions for years is often difficult [13]. This is also the case for processed biological materials (e.g. food products, pharmaceuticals etc.) and decayed tissues due to their degraded quality of DNA [14]. In such instances where obtaining a full-length barcode is hard, shorter barcode sequences ($\sim$100$-$300 bp), often referred to as *Mini-barcodes*, have been found efficacious for species-level identification owing to their comparatively better amplification gain [15]. For this reason, mini-barcode is also being popularly used in DNA metabarcoding strategies based on environmental DNA (eDNA) [16] and researching ecological biodiversity [17]. Mini-barcodes can also be advantageously applied in alternative next generation short-read sequencing platforms providing higher throughput but remaining more cost-effective at the same time [13].

II. Related Works

Since DNA Barcode was exercised as an effective molecular approach in species diagnosis by Hebert *et al.* [6], it has been used as a robust viewpoint in a wide range of studies [18]–[20]. *iBOL* and *CBOL* has also patronized the development initiatives for establishing DNA Barcode as a global standard for species identification [21]. Methodologies perceived in the literature review of species identification approaches with DNA barcoding can be divided into few major categories [11] [22] [23]. In *Distance Based* approaches, the reference library sequences are ranked with respect to their distances with the query barcode sequence on basis of the impression that small distances suggest two barcodes might be of specimens of same species while large distances evince their belonging to two different species. This distance, commonly referred to as *Barcode Gap* [24], might be calculated on the basis of pairwise p-distances [6] or K2P distance [22]. *TaxI* [25] uses pairwise distances as distance measures whereas *BOLD* [26] uses K2P distances by default. In *Similarity Based* approaches, the query sequences are usually assigned to species based on a *Similarity score* [27] indicating how much barcode characters they have in common. *BLAST* [28], *FASTA* [29], TaxonDNA [30], *NN* [22] are some of the prominent examples in this regard. *Phylogenetic Approaches* construct trees using hierarchical clustering [27] where distinct clusters represent distinct species [31]. It can adopt NJ [32], Parsimony (*PAR* [33]), Maximum Likelihood (*PhyML* [34]) or Bayesian Inference (*MRBAYES* [35]) as underlying methods. *Statistical Approaches* use population genetics assumptions [36] based on coalescent theory (*GMYC* [37]) which takes phylogenetic uncertainty (*SAP-NJ* [38]), likelihood ratio tests [39] etc. into account. *Character Based* approaches investigate the presence/absence of particular characters, known as *Diagnostic Sites*, instead of total sequence [23] imitating morphological approaches. *CAOS* [40], *BLOG* [10], *DNABar* [41], *BRONX* [27], *DOME-ID* [27] are some instances of this approach. There are also some *Alignment-Free tools* such as: *ATIM-TNT* (Tree-based) [27], *CVTreeAlpha1.0* (Component Vector based) [42], *Spectrum Kernel* [43], *Alfie* (python based) [44] as well as *Web Based Tools* such as: *Linker* [45], *iBarcode* [46], *BioBarcode* [47], *ConFind* [48] for DNA barcoding. Comparative analysis of these major classical barcoding approaches are also extensively studied [11] [27] but there is no general consensus about a single best method to analyze DNA barcode data in species identification [22] [23].

Machine learning methods have been introduced as state-of-the-art methods for their extensive applicability in species identification with DNA Barcoding [12]. Although supervised statistical classification methods (CART, RF and Kernel methods) were studied previously [22]; supervised machine learning algorithms e.g. SVM, Jrip, J48 (C4.5), Naïve Bayes were compared on WEKA platforms against those ad-hoc methods by Weitschek *et al.* and it was observed that SVM and Naïve Bayes outperform on average in both synthetic and empirical datasets [12]. Simple-logistic, IBK, PART, Attribute-selected Classifier, Bagging approaches were also implemented in another study [49] in this regard. SMO, BP-NN [50], RF [51], k-mer based approaches [52] [53] can also be perceived in recent studies. Naïve Bayes is also applied on COI [54] barcode database demonstrating misclassification rates and also on ribosomal databases [55] effectively.

With the exploration of the prospects of DNA minibarcode in species identification [56], the application of DNA barcode has also significantly broadened. It is observed that although full length barcodes perform best with 97% species resolution, 90% and 95% identification success can be obtained with 100bp and 250bp mini barcodes respectively [15]. In case of species diagnosis from damaged/degraded DNA samples, $\sim$100$-$300bp mini-barcodes have also shown promising results [14]. Tree-based methods, objective clustering, automatic barcode gap discovery (ABGD), Poisson Tree Process (PTP) were also examined for analyzing the utility of minibarcodes for species identification [3]. Although no simple straightforward formula is available through which necessary sequence lengths for ensuring species identification can be easily predicted [7], the general assumption of achieving superior performance by using full barcode length can be questioned as a saturation point is often seen beyond which the additional length-data create only a little impact in comparison to the increased cost overheads [3]. In this work of ours, we have tried to shed some lights on this question from the machine learning perspective by demonstrating the accuracy responses of species identification at different partial barcode lengths after implementing a Naïve Bayes classifier as our analyzing contrivance.

III. Datasets

A. Empirical Data

Weitschek *et al.* stated that for species identification, a full reference set with all possible nucleotide polymorphisms for the sequences of each species is required as well as a sufficient reference set is important to avoid over-fitting or under-fitting [12]. For this purpose, we included the datasets they used in their work of [12] in our analysis which is basically a collection of published empirical datasets and simulated DNA Barcode datasets. The public empirical datasets of *Bats*, *Birds*, *Cypraeidae*, *Drosophila* and *Fishes* and were chosen from GenBank Nucleotide Database for high phylogenetic diversity and available for download at dmb.iasi.cnr.it/supbarcodes.php in FASTA format [12]. The datasets referred here as *Bird2*, *Fish2* and *Butterfly2* were extracted from BOLD and available in CSV format at http://archive.dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges2007/. Another dataset was considered while performing our analysis which appears to be the first large-scale genetic assessment of the butterflies of Southern South America highlighting several cases of both deep intraspecific and shallow interspecific divergence [57]. COI gene is used as gene region in DNA barcode sequences of all these datasets. The detailed overview of these empirical datasets having COI gene as barcode sequence can be found in Table I.

TABLE I
SUMMARY OF THE EMPIRICAL DATASETS: COI GENE AS BARCODES

Dataset	No of sequences	Sequence Length	No of Species	Gene Region	Ref
Cypraeidae	2,008	614	211	COI	[58]
Drosophila	615	663	19	COI	[59]
Bats	826	659	82	COI	[26]
Fishes	626	702	82	COI	[60]
Birds	1,700	691	150	COI	[61]
Bird 2	2589	990	656	COI	[62]
Fish 2	754	901	211	COI	[63]
Butterfly	2028	658	417	COI	[57]
Butterfly 2	4267	901	561	COI	[56]

With a view to inspecting the overall scenario of species identification using DNA Barcoding and analyzing the accuracy responses with varying barcode lengths, we incorporated datasets of plants and fungi as well. These empirical datasets use rbcL and ITS gene regions respectively and are available for download at dmb.iasi.cnr.it/supbarcodes.php in FASTA format [12]. The detailed overview of these empirical datasets of plants and fungi can be found in Table II.

TABLE II
SUMMARY OF THE EMPIRICAL DATASETS OF FUNGI, INGA AND ALGAE

Dataset	No of sequences	Sequence Length	No of Species	Gene Region	Ref
Fungi	50	930	8	ITS	[9]
Inga	913	1,838	56	ITS	[64]
Algae	26	1,128	5	rbcL	[8]

B. Simulated Data

We have also used simulated real DNA Barcode datasets in our work for further analysis. These datasets were also available for downlaod at dmb.iasi.cnr.it/supbarcodes.php in FASTA format [12]. As stated in [11], these data were simulated considering time of species divergence and the effective population size (N_e). $N_e = 1000, 10000, 50000$ were used respectively in simulating gene trees and the dataset complexity was increased with population size. After that DNA Barcode sequences were simulated on the additive gene trees with 650 base sequence-length bearing similarity to the real-life size of a standard DNA Barcode. Each dataset consists of 50 species and 20 individuals per species. The detailed overview of simulated datasets is given in Table III.

TABLE III
SUMMARY OF THE SIMULATED DATASETS

Dataset	N_e	No of Individual	Sequence Length	No of Species	Ref
Ne1000	1000	20	650	50	[11]
Ne10000	10000	20	650	50	[11]
Ne50000	50000	20	650	50	[11]

IV. Methodology

A. Naïve Bayes

Naïve Bayes [65] is one of those methods which are often used when a large reference set is available. Naive Bayes is based on Bayes' Theorem and it is assumed that the predictors are independent of each other. If we have an instance S with n features which is represented by a feature vector $x = \{x_1, x_2, \ldots, x_n\}$, this model assigns to S a probability $P(C_m|x_1, x_2, \ldots, x_n)$. Here, C_m is the m^{th} class among all possible classes.
Using Bayes' theorem, we can calculate the posterior probability $P(C_m|x)$ from likelihood $P(x|C_m)$, prior probability $P(C_m)$ and evidence $P(x)$.

$$P(C_m|x) = \frac{P(x|C_m)P(C_m)}{P(x)} \tag{1}$$

Under naïve conditional assumption,

$$P(x_i|x_{i+1}, \ldots, x_n, C_m) = P(x_i|C_m) \tag{2}$$

So the posterior probability can be calculated as follows.

$$\frac{P(C_m)\prod_{i=1}^{n} P(x_i|C_m)}{P(x)} \tag{3}$$

As P(x) is constant for given input, we can only consider the numerator.
Using the naïve Bayes model, the classification rule can be implemented as follows where $\hat{y}$ is the class label.:

$$\hat{y} = \underset{m}{\operatorname{argmax}}\, P(C_m) \prod_{i=1}^{n} P(x_i|C_m) \tag{4}$$

B. Species Identification Using Naïve Bayes

Because of the efficacy of Naïve Bayes Algorithm in supervised prediction problem [54], we have used this model for species identification process. We can outline the identification approach as follows:

Let, b_k is the k^{th} sequence in the dataset where $k = 1, 2, \ldots, n_b$ and n_b= no of barcode entries in the dataset. $x_{1k}, x_{2k}, \ldots, x_{nk}$ are the nucleotides in a barcode sequence b_k of length n. If C_p is the p^{th} species among different classes of species, the probability that the barcode sequence b_k belongs to the species C_p is:

$$P(C_p|b_k) = \frac{P(b_k|C_p)P(C_p)}{P(b_k)} \tag{5}$$

The prior probability in eq. 5 can be calculated as follows:

$$\begin{aligned} P(b_k|C_p) &= P(x_{1k}, x_{2k}, \ldots, x_{nk}|C_p) \\ &= P(x_{1k}|x_{2k}, \ldots, x_{nk}, C_p)P(x_{2k}|x_{3k}, \ldots, x_{nk}, C_p) \\ &\ldots P(x_{nk}|C_p) \\ &= P(x_{1k}|C_p)P(x_{2k}|C_p)\ldots P(x_{nk}|C_p) \end{aligned} \tag{6}$$

In the above equation, x_{ik} refers to the i^{th} nucleotide position of k^{th} sequence in the dataset and $P(x_{ik}|C_p)$ means the probability of nucleotide base x_{ik} in species C_p at position i where $i = 1, \ldots, n$. So, if N_{c_p} and $N_{x_{ik}}$ are the number of samples of C_p and the number of occurrences of x_{ik} at position i respectively, then We calculate this probability as:

$$P(x_{ik}|C_p) = \frac{N_{x_{ik}}}{N_{C_p}} \tag{7}$$

If the proportion is equal to 0, then the posterior probability will become 0. To avoid this, we replaced zero with a very small fractional value α, where $0 < \alpha << 1$. This α can be considered as the mutation rate of nucleotides in the COI region [54]. $\alpha = 9.7X10^{-}8$ was proposed as the estimated value of this mutation rate in COI genes [66] and hence is also used here. In case of ambiguous base (e.g., *N*) or missing data (e.g., **-**) at any position, we have not considered those positions for identification purpose. In some datasets, there is a significant number of missing data (i.e., **-**) in a significant number of sequences because of sequence alignments. Ignoring these missing data in such cases can affect the performance of our model in identification process. For this reason, in these situations we replaced a missing value in a particular position with the most frequently occurring nucleotide in that position within a species.

C. Data Preparation

The empirical and simulated DNA Barcode datasets discussed in III are used to perform our analysis. Most of these empirical datasets were already split into 80% and 20% sequences per species in training and testing set respectively by biologists maintaining necessary polymorphism and species divergences [11]. In cases where only a single dataset file was provided, it was split into Train and Test set on the basis of 80%-20% random split. But it was ensured that whenever possible each dataset consists 4 or more representing sequences per species in the training set as it was found necessary for obtaining a reliable classification performance [12]. The simulated datasets were also split into Train and Test set in 80%-20% ratio where for every species 16 sequences were placed into Train Set and 4 sequences were placed into Test set. In cases where datasets were non-aligned, MAFFT [67] was used as multiple sequence alignemnt tool for aligning the barcode sequences. Most of the datasets were in FASTA format and a python program is written in order to convert these FASTA files into corresponding CSV files. In case of a large number of missing nucleotide in barcode sequences of a species, the most frequently occurring certain nucleotide within that species was considered as replacement for that missing value.

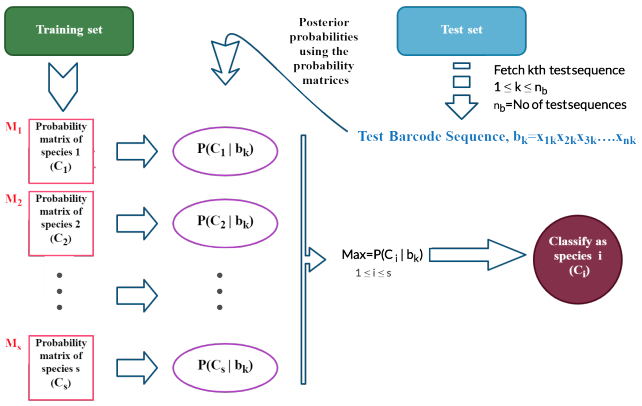

Fig. 1. Overview of Species Identification with Naïve Bayes Procedure

D. Experimental Arrangement

For the purpose of species identification with DNA Barcode, we developed a Naïve Bayes classifier from scratch in Java fitting our methodological necessities. In the training phase, a probability matrix for each known species is calculated. For accomplishing this task, a Hashmap is used where the key is the species name and the value is the probability matrix. The matrix basically stores the probability of each nucleotide bases at each position in the Barcode sequence for a species. A worked-out example can be found at Table IV where it is considered that there are four different species and each having three different barcode sequences of length five. The resulting probability matrix of *species* 1 is shown in Table V.

In the test phase, for a query sequence, we calculate it's likelihood of being a particular species. Our model calculates the maximum likelihood and classify the new sequence to that corresponding species. For example, if we get a query sequence TCCAC from test set, our model takes the probability of T, C, C, A, C in position 1, 2, 3, 4 and 5 respectively from probability matrix of species 1, 2 and so on. Finally if the highest probability value considering the whole length of query sequence comes from the matrix of species 2, the sequence will be classified as species 2. Figure 1 shows an overview of this classification process.

To analyze the impact of length of DNA Barcode sequences on species identification, we have run our model with different

TABLE IV
SAMPLE BARCODE SEQUENCES OF FOUR SPECIES

species 1	species 2	species 3	species 4
ATTGC	TCCAG	GTATG	CGAAT
ATTCC	TCCGG	GCATG	CGGAT
ATTGC	TCCGC	GTATC	CGAAT

sequence lengths on all the empirical as well as simulated datasets in a comprehensive manner.

TABLE V
PROBABILITY MATRIX OF SPECIES 1

	position 1	position 2	position 3	position 4	position5
A	3/3	0/3	0/3	0/3	0/3
T	0/3	3/3	3/3	0/3	0/3
G	0/3	0/3	0/3	2/3	0/3
C	0/3	0/3	0/3	1/3	3/3

V. RESULTS AND DISCUSSIONS

After implementing our Naïve bayes classifier, we needed to test the effectiveness of our method. For this reason, to ensure the efficacy of our model, we ran our Naïve bayes classifier on first five of the COI gene based empirical datasets mentioned in Subsection I considering their full-length barcode sequences. After comparing the results with the maximum accuracy mentioned in other related studies [12] [49], where comparative performance analysis of different supervised machine learning approaches on those same COI based empirical datasets are provided, we thereby find our Naïve Bayes classifier competent with their performances (shown in Fig. 2).

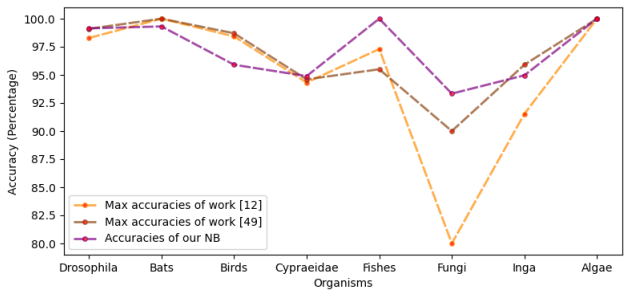

Fig. 2. Performance Competency of our Naïve Bayes Classifier

After testing our Naïve Bayes Classifier on the COI based first five empirical datasets, subsequently we ran our model on all of the COI based empirical datasets mentioned in Table I with a view to analyzing accuracy responses varying barcode sequence lengths. As all of the datasets in Table I had barcode sequence lengths >600 bp commonly, while varying lengths we chose to set the upper bound of sequence length at 600. Then the sequence lengths were decreasingly varied from this upper bound of 600 bp with an interval of 50 bases and corresponding accuracy scores were calculated in each case. As 50 bp falls too low as sequence length, we considered 70 base-length after 100 bp and tried to maintain this as the lower bound for varying sequence lengths throughout our study (with only a few exceptions). All these accuracy scores for varying DNA barcode lengths for COI based empirical datasets are represented in Table VI. These length-wise accuracy scores of Table VI are also plotted in Fig. 3 and demonstrated in the Heatmap of Fig. 4. If we analyze the pattern of these length-wise changes of accuracy scores in these plottings and heatmap, it seems that in most cases accuracy scores remain almost persistent in the range of 200bp-600bp. Although there is a gradual but trifling fall of accuracy in this range with the decrease of barcode sequence length, it significantly falls after <200bp barcode length.

TABLE VI
ACCURACY RESPONSES IN COI BASED EMPIRICAL DATASET

	Lengths											
	70	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**	**550**	**600**
Bats	93.06	97.22	99.31	99.31	99.31	99.31	99.31	99.31	99.31	99.31	99.31	99.31
Bird	90.54	97.48	98.42	98.42	98.42	97.48	97.16	97.48	97.16	97.16	97.16	97.16
Cypraeidae	89.77	91.19	93.75	94.60	94.03	94.60	95.17	95.17	95.17	95.45	95.45	95.45
Drosophila	96.55	98.28	98.28	98.28	98.28	99.14	99.14	99.14	99.14	99.14	99.14	99.14
Fish	88.29	93.69	99.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Butterfly	89.8	93.47	95.92	95.92	95.92	95.92	95.92	95.92	95.92	95.92	96.33	96.33
Bird 2	79.03	90.32	95.97	96.37	97.18	97.18	97.58	97.58	98.79	98.79	98.79	98.79
Butterfly 2	35.19	86.36	95.96	97.14	97.14	99.16	99.49	99.66	99.66	99.66	99.66	99.66
Fish 2	45.92	84.69	89.8	91.84	91.84	91.84	96.94	96.94	97.96	97.96	98.98	98.98

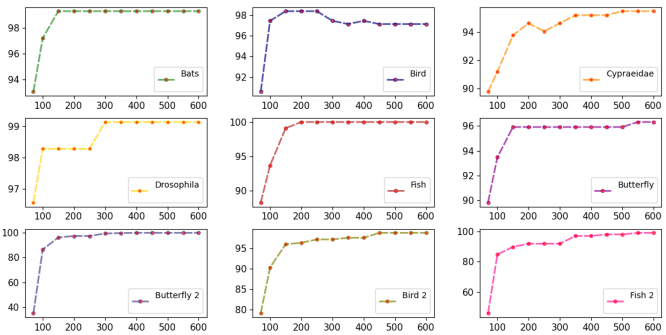

Fig. 3. Lengthwise Accuracy comparison for COI based Empirical Datasets

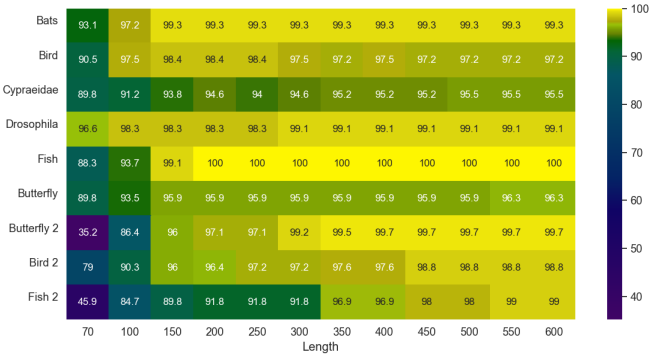

Fig. 4. Heatmap of Lengthwise Accuracy Comparison for COI based Empirical Datasets

This similar approach is applied for the simulated datasets mentioned in Table III with maintaing same upper bound of 600 bp and lower bound of 70bp. The length-wise accuracy responses for the simulated datasets are summarized in Table VII and demonstrated in Figure 5. From this Figure 5 it is also evident that, the accuracy score drops slightly and gradually with the reduction of sequence length and falls significantly when the sequence length is reduced to <200bp.

TABLE VII
ACCURACY RESPONSES IN SIMULATED DATASETS

	Lengths											
	70	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**	**550**	**600**
Ne1000	75.91	82.63	87.52	90.34	91.88	93.21	94.06	94.84	95.26	95.66	96.02	96.32
Ne10000	81.05	86.85	91.28	93.21	94.43	95.14	95.54	95.84	96.08	96.26	96.46	96.60
Ne50000	81.62	85.80	88.65	90.14	91.09	91.56	91.99	92.26	92.55	92.73	92.92	93.09

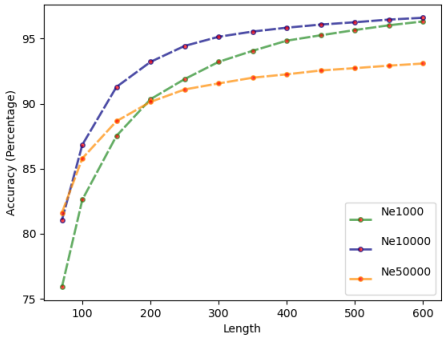

Fig. 5. Lengthwise Accuracy comparison for Simulated Datasets

Furthermore, in order to get a complete overview of the effect of length on the species identification process, the datasets mentioned in Table III, which are based on ITS and rbcL genes, are analyzed. For the dataset *Fungi* and *Algae*, we set upper bound and lower bound for varying sequence length as 850bp and 100bp respectively. For the dataset *Inga*, we set 1800bp and 1050bp as respective upper and lower bounds as the full lengths of barcode sequences in this dataset are >1800bp. In all three datasets, we decrease the length with an interval of 50 bases. The complete length-wise average accuracy analysis of these datasets is represented in Table VIII. This analysis is also depicted in Fig. 6.

TABLE VIII
LENGTHWISE ACCURACY RESPONSES IN FUNGI, INGA AND ALGAE DATASET

Fungi	bp	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850
	%	80.56	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89	88.89
Inga	bp	1050	1100	1150	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750	1800
	%	78.79	78.79	78.79	78.91	85.09	86.63	87.53	87.40	87.40	87.40	87.14	89.71	90.48	90.10	95.11	95.11
Algae	bp	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850
	%	90.0	90.0	90.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

From there it can be seen that although the contexts are different from the previous length wise accuracy patterns of COI based empirical datasets and simulated datasets, still *Algae* and

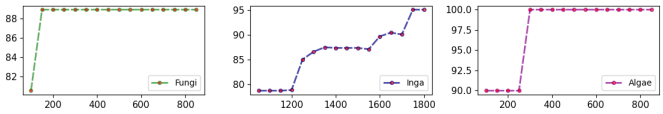

Fig. 6. Lengthwise Accuracy comparison in Fungi, Inga and Algae Dataset

Fungi shows a sharp fall in accuracy scores around sequence lengths below 200-250bp. But this corroboration cannot be extended in case of *Inga* as its length-bounds had to be chosen differently; but it still shows a decline in accuracy score like every other ones if the lengths are gradually decreased from full length. In order to bring these three datasets on a common ground, average accuracy responses were further calculated at various percentage levels of their full lengths. The results are represented in Table IX and depicted in Fig. 7. From these representations, we can observe that although no persistent range like previous COI based empirical datasets is evident from these ITS & rbCL based datasets, the length-wise accuracy of *Fungi* and *Algae* falls sharply at around 20%-30% of their full lengths whereas the accuracy in *Inga* dataset considerably falls first at around 70% and later again at 40% of its full length. It is worth mentioning in this regard that, in case of COI based datasets and simulated datasets, the length (∼200bp) beyond which a sharp decline was observed happens to be approx. 30% of the full length of barcode sequence. In both cases of Table VIII and Table IX, the first row denotes the length of the sequence and the second row denotes the corresponding accuracy considering that length respectively for each entry.

TABLE IX
ACCURACY RESPONSES IN FUNGI, INGA AND ALGAE DATASET AT DIFFERENT LENGTH-PERCENTAGE

		Percentages of Lengths									
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Fungi	bp	93	186	279	372	465	558	651	744	837	930
	%	79.44	93.33	93.33	93.33	93.33	93.33	93.33	93.33	93.33	93.33
Inga	bp	183	367	551	735	919	1102	1286	1470	1654	1838
	%	46.33	66.84	67.09	76.41	76.54	79.09	87.57	88.49	90.06	94.96
Algae	bp	112	225	338	451	564	676	789	902	1015	1128
	%	90.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

In support of our findings, we can allude to Fig. 8 which is adapted from the work of [15]. In this figure, the proportion of species resolution with different DNA Barcode size is depicted. The shapes in individual plottings of Fig. 3 also comport with the shape in Fig. 8 implying the consistency between these separate but similar findings. In other related studies regarding mini-barcodes, it was also observed that full length barcodes provide marginally better identification success rates than mini barcodes having 200-400bp sequence length while the identification ability is significantly compromised only when the barcode length falls too short <200bp [3]. In light of these above discussions, this proposition can

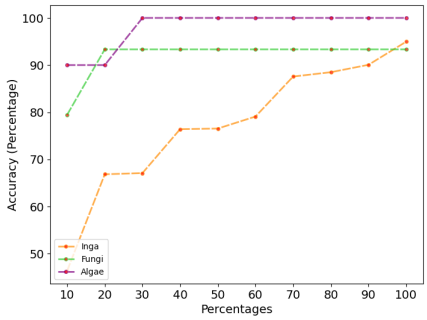

Fig. 7. Accuracy responses in Fungi, Inga and Algae dataset at different length-percentage

be made that an approximation of 200-350bp mini-barcode length can be suggested for using in the process of species identification with DNA barcode approached with a supervised machine learning model where it has a significant possibility to perform with a reasonable accuracy score.

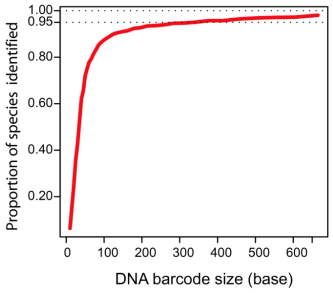

Fig. 8. Change of Species Resolution in different DNA mini-barcode lengths (adapted from [15])

VI. Conclusion

In this study, we tried to provide an approximation of required length of DNA Mini-barcode from the supervised learning perspective by analysing the length-wise accuracy in the species identification process. A Naïve Bayes classifier was implemented for the comprehensive analysis of the effect of varying DNA Barcode lengths on the identification accuracy. Our findings from this machine learning perspective bears consistency with the findings of other related biological works where DNA mini-barcodes were used and analyzed to facilitate the species identification process. Although we could perceive the approximate effective length of mini barcodes in case of COI gene based DNA sequences, further study is needed for plants and ITS gene barcode based organisms. Improving the quality of the sequence alignment by using better multiple sequence alignment tools/techniques might improve the identification accuracy in this regard. Other machine learning algorithms are also encouraged to be applied in this context to enrich the comprehensiveness of our finding.

References

[1] M. A. Ruggiero, D. P. Gordon, T. M. Orrell, N. Bailly, T. Bourgoin, R. C. Brusca, T. Cavalier-Smith, M. D. Guiry, and P. M. Kirk, "A higher level classification of all living organisms," *PloS one*, vol. 10, no. 4, p. e0119248, 2015.

[2] C. Mora, D. P. Tittensor, S. Adl, A. G. Simpson, and B. Worm, "How many species are there on earth and in the ocean?" *PLoS Biol*, vol. 9, no. 8, p. e1001127, 2011.

[3] D. Yeo, A. Srivathsan, and R. Meier, "Longer is not always better: Optimizing barcode length for large-scale species discovery and identification," *Systematic Biology*, 2020.

[4] B. Dayrat, "Towards integrative taxonomy," *Biological journal of the Linnean society*, vol. 85, no. 3, pp. 407–417, 2005.

[5] K. Armstrong and S. Ball, "Dna barcodes for biosecurity: invasive species identification," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1813–1823, 2005.

[6] P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. Dewaard, "Biological identifications through dna barcodes," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. 1512, pp. 313–321, 2003.

[7] P. D. Hebert, S. Ratnasingham, and J. R. De Waard, "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. suppl_1, pp. S96–S99, 2003.

[8] C. P. W. Group, P. M. Hollingsworth, L. L. Forrest, J. L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M. W. Chase, R. S. Cowan, D. L. Erickson *et al.*, "A dna barcode for land plants," *Proceedings of the National Academy of Sciences*, vol. 106, no. 31, pp. 12 794–12 797, 2009.

[9] C. L. Schoch, K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, F. B. Consortium *et al.*, "Nuclear ribosomal internal transcribed spacer (its) region as a universal dna barcode marker for fungi," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 6241–6246, 2012.

[10] E. Weitschek, R. Van Velzen, G. Felici, and P. Bertolazzi, "Blog 2.0: a software system for character-based species classification with dna barcode sequences. what it does, how to use it," *Molecular ecology resources*, vol. 13, no. 6, pp. 1043–1046, 2013.

[11] R. van Velzen, E. Weitschek, G. Felici, and F. T. Bakker, "Dna barcoding of recently diverged species: relative performance of matching methods," *PloS one*, vol. 7, no. 1, p. e30490, 2012.

[12] E. Weitschek, G. Fiscon, and G. Felici, "Supervised dna barcodes species classification: analysis, comparisons and results," *BioData mining*, vol. 7, no. 1, pp. 1–18, 2014.

[13] M. Hajibabaei, M. A. Smith, D. H. Janzen, J. J. Rodriguez, J. B. Whitfield, and P. D. Hebert, "A minimalist barcode can identify a specimen whose dna is degraded," *Molecular Ecology Notes*, vol. 6, no. 4, pp. 959–964, 2006.

[14] M. Hajibabaei and C. McKenna, "Dna mini-barcodes," in *DNA barcodes*. Springer, 2012, pp. 339–353.

[15] I. Meusnier, G. A. Singer, J.-F. Landry, D. A. Hickey, P. D. Hebert, and M. Hajibabaei, "A universal dna mini-barcode for biodiversity analysis," *BMC genomics*, vol. 9, no. 1, pp. 1–4, 2008.

[16] M. Staats, A. J. Arulandhu, B. Gravendeel, A. Holst-Jensen, I. Scholtens, T. Peelen, T. W. Prins, and E. Kok, "Advances in dna metabarcoding for food and wildlife forensic species identification," *Analytical and Bioanalytical Chemistry*, vol. 408, no. 17, pp. 4615–4630, 2016.

[17] D. L. Erickson, E. Reed, P. Ramachandran, N. A. Bourg, W. J. McShea, and A. Ottesen, "Reconstructing a herbivore's diet using a novel rbc l dna mini-barcode for plants," *AoB Plants*, vol. 9, no. 3, p. plx015, 2017.

[18] L. Frézal and R. Leblois, "Four years of dna barcoding: current advances and prospects," *Infection, Genetics and Evolution*, vol. 8, no. 5, pp. 727–736, 2008.

[19] P. Z. Goldstein and R. DeSalle, "Review and interpretation of trends in dna barcoding," *Frontiers in Ecology and Evolution*, vol. 7, p. 302, 2019.

[20] S. J. Adamowicz, J. S. Boatwright, F. Chain, B. L. Fisher, I. D. Hogg, F. Leese, D. A. Lijtmaer, M. Mwale, A. M. Naaum, X. Pochon *et al.*, "Trends in dna barcoding and metabarcoding," *Genome*, vol. 62, no. 3, pp. v–viii, 2019.

[21] D. E. Schindel and S. E. Miller, "Dna barcoding a useful tool for taxonomists," *Nature*, vol. 435, no. 7038, pp. 17–18, 2005.

[22] F. Austerlitz, O. David, B. Schaeffer, K. Bleakley, M. Olteanu, R. Leblois, M. Veuille, and C. Laredo, "Dna barcode analysis: a comparison of phylogenetic and statistical classification methods," *BMC bioinformatics*, vol. 10, no. S14, p. S10, 2009.

[23] M. Casiraghi, M. Labra, E. Ferri, A. Galimberti, and F. De Mattia, "Dna barcoding: a six-question tour to improve users' awareness about the method," *Briefings in bioinformatics*, vol. 11, no. 4, pp. 440–453, 2010.

[24] D. P. Little, "Dna barcode sequence identification incorporating taxonomic hierarchy and within taxon variability," *PLoS one*, vol. 6, no. 8, p. e20552, 2011.

[25] D. Steinke, M. Vences, W. Salzburger, and A. Meyer, "Taxi: a software tool for dna barcoding using distance methods," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1975–1980, 2005.

[26] S. Ratnasingham and P. D. Hebert, "Bold: The barcode of life data system (http://www. barcodinglife. org)," *Molecular ecology notes*, vol. 7, no. 3, pp. 355–364, 2007.

[27] D. P. Little and D. W. Stevenson, "A comparison of algorithms for the identification of specimens using dna barcodes: examples from gymnosperms," *Cladistics*, vol. 23, no. 1, pp. 1–21, 2007.

[28] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[29] W. Pearson, "Rapid and sensitive sequence comparison with fastp and fasta." *Methods in enzymology*, vol. 183, pp. 63–98, 1990.

[30] R. Meier, K. Shiyang, G. Vaidya, and P. K. Ng, "Dna barcoding and taxonomy in diptera: a tale of high intraspecific variability and low identification success," *Systematic biology*, vol. 55, no. 5, pp. 715–728, 2006.

[31] K. Munch, W. Boomsma, E. Willerslev, and R. Nielsen, "Fast phylogenetic dna barcoding," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1512, pp. 3997–4002, 2008.

[32] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[33] J. S. Farris, "Estimating phylogenetic trees from distance matrices," *The American Naturalist*, vol. 106, no. 951, pp. 645–668, 1972.

[34] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic biology*, vol. 52, no. 5, pp. 696–704, 2003.

[35] J. P. Huelsenbeck and F. Ronquist, "Mrbayes: Bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.

[36] Z. Abdo and G. B. Golding, "A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups," *Systematic biology*, vol. 56, no. 1, pp. 44–56, 2007.

[37] J. Pons, T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler, "Sequence-based species delimitation for the dna taxonomy of undescribed insects," *Systematic biology*, vol. 55, no. 4, pp. 595–609, 2006.

[38] K. Munch, W. Boomsma, J. P. Huelsenbeck, E. Willerslev, and R. Nielsen, "Statistical assignment of dna sequences using bayesian phylogenetics," *Systematic biology*, vol. 57, no. 5, pp. 750–757, 2008.

[39] M. V. Matz and R. Nielsen, "A likelihood ratio test for species membership based on dna sequence data," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1969–1974, 2005.

[40] I. N. Sarkar, P. J. Planet, and R. Desalle, "Caos software for use in character-based dna barcoding," *Molecular Ecology Resources*, vol. 8, no. 6, pp. 1256–1259, 2008.

[41] B. DasGupta, K. M. Konwar, I. Măndoiu, and A. A. Shvartsman, "Dnabar: distinguisher selection for dna barcoding," *Bioinformatics*, vol. 21, no. 16, pp. 3424–3426, 2005.

[42] K. H. Chu, M. Xu, and C. P. Li, "Rapid dna barcoding analysis of large datasets using the composition vector method," *BMC bioinformatics*, vol. 10, no. S14, p. S8, 2009.

[43] P. Kuksa and V. Pavlovic, "Efficient alignment-free dna barcode analytics," *BMC bioinformatics*, vol. 10, no. S14, p. S9, 2009.

[44] C. M. Nugent and S. J. Adamowicz, "Alignment-free classification of coi dna barcode data with the python package alfie," *Metabarcoding and Metagenomics*, vol. 4, p. e55815, 2020.

[45] M. Albu, H. Nikbakht, M. Hajibabaei, and D. A. Hickey, "The dna barcode linker," *Molecular ecology resources*, vol. 11, no. 1, pp. 84–88, 2011.

[46] G. A. Singer and M. Hajibabaei, "ibarcode. org: web-based molecular biodiversity analysis," in *BMC bioinformatics*, vol. 10, no. S6. Springer, 2009, p. S14.

[47] J. Lim, S.-Y. Kim, S. Kim, H.-S. Eo, C.-B. Kim, W. K. Paek, W. Kim, and J. Bhak, "Biobarcode: a general dna barcoding database and server platform for asian biodiversity resources," in *BMC genomics*, vol. 10, no. S3. Springer, 2009, p. S8.

[48] J. A. Smagala, E. D. Dawson, M. Mehlmann, M. B. Townsend, R. D. Kuchta, and K. L. Rowlen, "Confind: a robust tool for conserved sequence identification," *Bioinformatics*, vol. 21, no. 24, pp. 4420–4422, 2005.

[49] T. Kabir, A. S. Shemonti, and A. H. Rahman, "Species identification using partial dna sequence: A machine learning approach," in *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2018, pp. 235–242.

[50] T. He, L. Jiao, A. C. Wiedenhoeft, and Y. Yin, "Machine learning approaches outperform distance-and tree-based methods for dna barcoding of pterocarpus wood," *Planta*, vol. 249, no. 5, pp. 1617–1625, 2019.

[51] P. K. Meher, T. K. Sahu, S. Gahoi, R. Tomar, and A. R. Rao, "funbarrf: Dna barcode-based fungal species prediction using multiclass random forest supervised learning model," *BMC genetics*, vol. 20, no. 1, p. 2, 2019.

[52] P. K. Meher, T. K. Sahu, and A. Rao, "Identification of species based on dna barcode using k-mer feature vector and random forest classifier," *Gene*, vol. 592, no. 2, pp. 316–324, 2016.

[53] W. A. Kusuma, M. Nurilmala *et al.*, "Identification of tuna and mackerel based on dna barcodes using support vector machine (svm)," *Telkomnika*, vol. 14, no. 2, p. 778, 2016.

[54] M. P. Anderson and S. R. Dubnicka, "A sequential naive bayes classifier for dna barcodes," *Statistical applications in genetics and molecular biology*, vol. 13, no. 4, pp. 423–434, 2014.

[55] T. M. Porter, J. F. Gibson, S. Shokralla, D. J. Baird, G. B. Golding, and M. Hajibabaei, "Rapid and accurate taxonomic classification of insect (class insecta) cytochrome c oxidase subunit 1 (coi) dna barcode sequences using a naïve bayesian classifier," *Molecular Ecology Resources*, vol. 14, no. 5, pp. 929–942, 2014.

[56] M. Hajibabaei, D. H. Janzen, J. M. Burns, W. Hallwachs, and P. D. Hebert, "Dna barcodes distinguish species of tropical lepidoptera," *Proceedings of the National Academy of Sciences*, vol. 103, no. 4, pp. 968–971, 2006.

[57] P. D. Lavinia, E. O. Núñez Bustos, C. Kopuchian, D. A. Lijtmaer, N. C. García, P. D. Hebert, and P. L. Tubaro, "Barcoding the butterflies of southern south america: Species delimitation efficacy, cryptic diversity and geographic patterns of divergence," *PloS one*, vol. 12, no. 10, p. e0186845, 2017.

[58] C. P. Meyer and G. Paulay, "Dna barcoding: error rates based on comprehensive sampling," *PLoS biol*, vol. 3, no. 12, p. e422, 2005.

[59] M. Lou and G. B. Golding, "Assigning sequences to species in the absence of large interspecific differences," *Molecular Phylogenetics and Evolution*, vol. 56, no. 1, pp. 187–194, 2010.

[60] P. Bertolazzi, G. Felici, and E. Weitschek, "Learning to classify species with barcodes," *BMC bioinformatics*, vol. 10, no. S14, p. S7, 2009.

[61] P. D. Hebert, M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis, "Identification of birds through dna barcodes," *Plos biol*, vol. 2, no. 10, p. e312, 2004.

[62] K. C. Kerr, M. Y. Stoeckle, C. J. Dove, L. A. Weigt, C. M. Francis, and P. D. Hebert, "Comprehensive dna barcode coverage of north american birds," *Molecular ecology notes*, vol. 7, no. 4, pp. 535–543, 2007.

[63] R. D. Ward, T. S. Zemlak, B. H. Innes, P. R. Last, and P. D. Hebert, "Dna barcoding australia's fish species," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1847–1857, 2005.

[64] K. G. Dexter, T. D. Pennington, and C. W. Cunningham, "Using dna to assess errors in tropical tree identifications: How often are ecologists wrong and when does it matter?" *Ecological Monographs*, vol. 80, no. 2, pp. 267–286, 2010.

[65] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[66] D. R. Denver, K. Morris, M. Lynch, L. L. Vassilieva, and W. K. Thomas, "High direct estimate of the mutation rate in the mitochondrial genome of caenorhabditis elegans," *Science*, vol. 289, no. 5488, pp. 2342–2344, 2000.

[67] K. Katoh and D. M. Standley, "Mafft multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.