



Article CNN_FunBar: Advanced Learning Technique for Fungi ITS Region Classification

Ritwika Das, Anil Rai * D and Dwijesh Chandra Mishra D

Division of Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India

* Correspondence: Anil.Rai@icar.gov.in

Abstract: Fungal species identification from metagenomic data is a highly challenging task. Internal Transcribed Spacer (ITS) region is a potential DNA marker for fungi taxonomy prediction. Computational approaches, especially deep learning algorithms, are highly efficient for better pattern recognition and classification of large datasets compared to in silico techniques such as BLAST and machine learning methods. Here in this study, we present CNN_FunBar, a convolutional neural network-based approach for the classification of fungi ITS sequences from UNITE+INSDC reference datasets. Effects of convolution kernel size, filter numbers, *k*-mer size, degree of diversity and category-wise frequency of ITS sequences on classification performances of CNN models have been assessed at all taxonomic levels (species, genus, family, order, class and phylum). It is observed that CNN models can produce >93% average accuracy for classifying ITS sequences from balanced datasets with 500 sequences per category and 6-mer frequency features at all levels. The comparative study has revealed that CNN_FunBar can outperform machine learning-based algorithms (SVM, KNN, Naïve-Bayes and Random Forest) as well as existing fungal taxonomy prediction software (funbarRF, Mothur, RDP Classifier and SINTAX). The present study will be helpful for fungal taxonomy classification using large metagenomic datasets.

Keywords: CNN; fungi ITS; k-mer; KNN; Naïve-Bayes; random forest; SVM; taxonomy; topsis; UNITE

1. Introduction

Microorganisms are an inevitable part of the ecosystem. They play significant roles in organic matter decomposition [1,2], nutrient mineralization [3,4], atmospheric nitrogen fixation [5,6], bioremediation [7–10], etc. However, various harmful microbes are also present, which cause diseases in plants, animals as well as human beings. Although millions of diverse microbes exist in the environment as a highly interconnected network [11,12], only 1% of them can be isolated, cultured and studied through traditional wet lab experiments. Due to rapid advancement in sequencing methods, "Metagenomics" has emerged as a highly beneficial tool for researchers to characterize microbial community composition based on sequence data obtained directly from any environmental sample such as the human gut, infected tissue, crop rhizosphere, soil, ocean, etc. [13–16]. Taxonomy identification is the most crucial step of the metagenomic analysis pipeline [17–19]. Amplicon sequencing, i.e., DNA metabarcoding, has been found to be a highly successful approach [20–23]. In this method, a particular genomic region called marker gene is amplified using suitable primer pairs and sequenced, which can identify distinct species depending on the barcode gap. These marker genes vary according to the type of organisms. The 16S rRNA gene is the widely accepted marker gene for profiling prokaryotic (e.g., bacteria, archaea) microbes [24,25]. For eukaryotic species identification, several rRNA genes like large ribosomal subunit (LSU), small ribosomal subunit (SSU), the largest subunit of DNA-directed RNA polymerase II (RPB1), internal transcribed spacer (ITS region), etc. have been found as effective molecular markers [26].



Citation: Das, R.; Rai, A.; Mishra, D.C. CNN_FunBar: Advanced Learning Technique for Fungi ITS Region Classification. *Genes* **2023**, *14*, 634. https://doi.org/10.3390/ genes14030634

Academic Editor: Juan Francisco Martín

Received: 1 December 2022 Revised: 28 December 2022 Accepted: 9 January 2023 Published: 3 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). Fungi are the second most abundant eukaryotic microbial domain consisting of around 3.8 million species and playing diverse roles as symbionts [27,28], mutualists [29,30], decomposers [31,32], antibiotic producers [33,34] as well as pathogens [35,36]. This domain consists of nine major phyla, viz., *Ascomycota, Basidiomycota, Opisthosporidia, Chytridiomycota, Neocallimastigomycota, Blastocladiomycota, Zoopagomycota, Mucoromycota* and *Glomeromycota* [37] among which *Ascomycota* and *Basidiomycota* are the largest (~70 to 90% of total fungal species) [38] and second largest (~35,000 species) [39] phyla, respectively.

Various DNA markers such as mitochondrial cytochrome c oxidase subunit 1 (CO1), protein-coding genes, viz., the largest subunit of RNA polymerase II (RPB1), the second largest subunit of RNA polymerase II (RPB II), minichromosome maintenance protein (MCM7), 18S rRNA, internal transcribed spacer (ITS) region, D1/D2 region of large ribosomal subunit (LSU), etc. have been used for fungal species identification [40]. Among these, the ITS region consisting of ITS1 and ITS2 hypervariable regions separated by comparatively conserved 5.8S segment (Figure 1) is the universally accepted DNA marker due to high variability in sequence composition and length among species [40]. Protein-coding DNA markers are highly efficient for phylogenetic analysis and fungal species identification. However, these genes have lower PCR (Polymerase Chain Reaction) amplification success rates compared to the ITS region. Apart from some early diverging fungal lineages, viz., *Pezizomycotina, Rozella, Neocallimasgomycota*, etc. and cryptic fungi species, ITS region is capable of discriminating a broad range of closely related fungal species with a high PCR amplification success rate [40].



Figure 1. Schematic diagram of fungi ITS region consisting of ITS1 and ITS2 regions separated by 5.8S segment. ITS1-F and ITS4-B are widely used forward and reverse primers [41] to amplify the whole ITS region.

Various ITS reference databases have been developed like UNITE (https://unite.ut.ee/; accessed on 20 October 2022) [42], Warcup training dataset (accessed on 20 October 2022) [43], BOLD database (https://boldsystems.org/; accessed on 21 October 2022) [44] and ITS1 database of NCBI GenBank (https://www.ncbi.nlm.nih.gov/; accessed on 21 October 2022) [45]. UNITE is a web-based database of nuclear ribosomal ITS sequences of various eukaryotic microorganisms. It collects all eukaryotic ITS sequences from the INSDC database and clusters them based on various similarity thresholds into species-level Operational Taxonomic Units (OTUs). These OTUs are called Species Hypothesis (SH), and they are assigned distinct Digital Object Identifiers (DOIs) for facilitating precise identification and assembly [42]. If two or more ITS sequences are available for a particular SH, one of them is randomly chosen to represent that SH. Currently, the UNITE database contains >6.4 million ITS sequences belonging to >0.2 million fungal species. However, more than 90% of a wide range of fungal species are still unknown [40]. The lack of reference sequences poses a significant challenge in fungal species delineation. Fungal taxonomy assignment can be done by aligning the query sequence against reference sequences through BLAST [46]. However, it is time-consuming and inefficient to identify

novel species as compared to alignment-free machine learning-based algorithms. In recent years, a few machine learning-based fungal ITS classifiers have been developed, such as the RDP classifier [43,47], Mycofier [45], Mothur [48], SINTAX [49] and funbarRF [50]. The RDP classifier's taxonomy assignment is based on 8-mer frequency features using the Naïve-Bayes algorithm [47]. Another Naïve-Bayes classifier, Mycofier, can classify fungal ITS1 sequences using 5-mer features up to the genus level [45]. Both Mothur [48] and SIN-TAX [49] use 8-mer features for fungi classification by the *k*-nearest neighbor (KNN) and non-Bayesian approaches, respectively. Random forest algorithm has been implemented in funbarRF for fungal species classification based on *g*-spaced di-nucleotide frequency feature representation [50].

Although various machine-learning approaches have been developed, there is enough scope for further research. Performances of the RDP classifier and SINTAX were evaluated using the Warcup training dataset (8551 species from 1461 genera) [43,49]. Mycofier was developed using ITS1 datasets (1794 species from 510 genera) of NCBI GenBank [45]. funbarRF was trained using the BOLD database (3674 species from 777 genera), and its performance was evaluated using various simulated and real metagenome datasets [50]. However, UNITE dataset covering a broad range of fungal species has not been extensively explored to develop machine learning-based classification approaches. The recent version of UNITE database (Version 9.0; Dated: 16 October 2022) comprises around 6,441,764 fungal ITS reference sequences belonging to 290,922 UNITE fungal SH with DOIs at 1.5% threshold (https://unite.ut.ee/#main; accessed on 21 October 2022). Recently, deep learning has become an effective paradigm for the classification and clustering of big data [51,52]. Deep learning algorithms like CNN and DBN have been found to be highly efficient for bacterial taxonomic classification based on whole shotgun metagenome as well as 16S rRNA amplicon sequencing datasets as compared to machine learning methods [53].

In this present work, we have evaluated the effects of convolution kernel size, number of filters, different *k*-mer sizes, degree of diversities and varying category-wise data frequencies in each taxonomic level on the performance of the convolutional neural network-based algorithm, CNN_FunBar, for classifying fungi ITS sequences using UNITE+INSDC dataset. We have compared the classification performance of CNN with four machine learning-based classification algorithms, i.e., SVM, KNN, Naïve Bayes and Random Forest. A comparative analysis between CNN and four existing ITS classification software, i.e., funbarRF [50], Mothur [48], RDP Classifier [43] and SINTAX [49], has also been carried out.

2. Materials and Methods

2.1. Fungal Barcode Datasets

The recent release of the UNITE+INSDC dataset (Version: 9.0, Release Date: 16 October 2022) comprising 6,441,764 fungi ITS sequences belonging to 20 fungal phyla has been downloaded from https://unite.ut.ee/repository.php; accessed on 21 October 2022. From this dataset, 5,530,925 sequences (85.86%) belonging to the 2 most abundant phyla, viz., *Ascomycota* (41.95%) and *Basidiomycota* (43.91%), are considered for this study. In this dataset, class *Agaricomycetes* (44.98%), order *Agaricales* (18.19%), family *Russulaceae* (7.22%), genus *Russula* (5.90%) and species *Russula_sp.* (5.82%) are found to be the most abundant class, order, family, genus and species, respectively. Redundant sequences, as well as those containing ambiguous characters other than A, T, G and C, have been removed from this dataset. After these filtering steps, the remaining dataset consists of 4,504,529 sequences (Table 1).

Table 1. Taxonomic level-wise distribution of ITS sequences in the dataset after filtering for duplicates and ambiguous nucleotides.

Phylum	Class	Order	Family	Genus	Species	Total ITS Sequences
2	49	274	1049	6367	44,167	4,504,529

Among these, several ITS sequences that do not have clearly defined identifiers for one or more taxonomic levels have been found. In the FASTA header of such sequences, those identifiers are suffixed with "*Incertae_sedis*" for that particular category. For example, the ITS sequence with GenBank ID *UDB01488607* does not have a clearly defined class identity. So, in the dataset, its FASTA header is represented as:

>UDB01488607/k_Fungi;p_Basidiomycota;c_Basidiomycota_cls_Incertae_sedis;o_Basidiomycota_ord_Incertae_sedis;f_Basidiomycota_fam_Incertae_sedis;g_Basidiomycota_gen_Incertae_sedis;s_Basidiomycota_sp/SH1250579.09FU

Likewise, sequences corresponding to undefined order, family and genus are annotated as "_ord_Incertae_sedis", "_fam_Incertae_sedis" and "_gen_Incertae_sedis", respectively. ITS sequences for which genus names are known but distinct species names are not assigned, species identifiers of such sequences are represented as [genus_name]_sp. If the higher taxonomic level of a sequence is undefined, then all lower taxonomic levels are also found to be unknown. In this study, the performances of all classifiers have been evaluated in 6 taxonomic levels, and separate training datasets have been prepared for each level. To avoid ambiguity due to uncertainly defined identifiers at any taxonomic level, we have filtered those sequences from our dataset for that level. For example, to prepare the class-level training dataset, ITS sequences having class identifiers ending with "_cls_Incertae_sedis" are removed. Similarly, for order, family, genus, and species level training datasets preparation, sequence identifiers suffixed with "_ord_Incertae_sedis", "_fam_Incertae_sedis", "_gen_Incertae_sedis" and "_sp" for order, family, genus and species levels, respectively are removed (Figure 2).

For each of these six taxonomic-level datasets, the corresponding taxon is considered as the "category". We have observed that the category-wise frequencies of ITS sequences are not uniform in these datasets. For example, in the family-level dataset, some families have >1000 ITS sequences, whereas some have even <50 ITS sequences. Similar scenarios have been observed for class, order, genus and species-level datasets. For example, in the species-level dataset, 35,613 out of 38,365 species have <20 ITS sequences. Balanced datasets with higher class frequencies are preferable for training the deep learning classifiers for better pattern recognition and higher model accuracies. Therefore, separate balanced datasets have been prepared for each taxonomic level using the stratified sampling method. For instance, to prepare family level-balanced dataset with 100 ITS sequences per category, 100 ITS sequences are randomly sampled from the family-level filtered ITS sequence dataset where 453 families have been found as strata with the minimum number of ITS sequences >100. So, the final dataset contains 45,300 sequences. In the same way, balanced datasets have been prepared for other taxonomic levels (Figure 2). The effects of diversity degrees, as well as category-wise data frequencies, on the performance of the CNN model, have been explored in this study. Hence, separate collections of balanced datasets with varying (Figure 2) and invariant unique categories have been prepared for each taxonomic level. Detailed information about these datasets is summarized in Tables 2 and 3.

As only two phyla (i.e., *Ascomycota* and *Basidiomycota*) are considered in the present work, datasets having variation in the unique category numbers cannot be prepared for the phylum level (Table 2). Therefore, in-house python scripts have been used for data filtering as well as sampling to create these datasets.



Figure 2. Diagrammatic representation of the pipeline for filtering and construction of balanced ITS sequence datasets at each taxonomic level.

Taxonomic Level	Datasets	No. of Unique Categories	No. of Sequence per Category	Total No. of ITS Sequences
	S1	2752	20	55,040
	S2	939	50	46,950
Creasian	S3	429	100	42,900
species	S4	148	250	37,000
	S5	117	300	35,100
	S6	64	500	32,000
	G1	1682	50	84,100
Ō	G2	1293	100	129,300
Genus	G3	841	250	210,250
	G4	580	500	290,000
	F1	520	50	26,000
Family	F2	453	100	45,300
Failiny	F3	369	250	92,250
	F4	301	500	150,500
0.1	O1	149	100	14,900
Order	O2	106	500	53,000
	C1	31	100	3100
Class	C2	25	500	12,500
Phylum	P1	2	500	1000

Table 2. Balanced datasets of ITS sequences with varying numbers of categories for all taxonomic levels.

Table 3. Balanced datasets of ITS sequences with invariant category numbers for different taxonomic levels (from Species to Class levels).

Taxonomic Level	No. of Unique Categories	No. of Sequence per Category	Total No. of ITS Sequences
		100	6400
Species	64	250	16,000
		500	32,000
		100	58,000
Genus	580	250	145,000
		500	290,000
		100	30,100
Family	301	250	75,250
		500	150,500
		100	10,600
Order	106	250	26,500
		500	53,000
		100	2500
Class	25	250	6250
		500	12,500

2.2. Feature Vector Generation

ITS sequences are strings of A, T, G and C letters. Machine learning and deep learning models cannot process these data as such. In previous studies, *k*-mer composition features have been considered for classifier development [45,47–50]. Here, we have considered *k*-mer frequencies for training and evaluation of all the classifiers. For a DNA sequence of length L, the total number of distinct feature descriptors for a particular value of *k* will be 4^k and the total number of extracted *k*-mers will be (L - k + 1). In this present work, three different values for *k*, i.e., *k* = 4, 5 and 6, have been considered to evaluate the classification performances of the CNN model as well as other machine learning algorithms. The total numbers of distinct feature descriptors are 256, 1024 and 4096 in our training datasets corresponding to 4-mers, 5-mers and 6-mers, respectively. Each ITS sequence is represented as frequencies of each feature descriptor for a particular *k* value. Normalization of *k*-mer frequencies has been performed so that features in the dataset can be assumed to follow

the normal distribution with zero mean and unit variance. Normalized features can be calculated as follows:

7.

$$_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{1}$$

Here, z_{ij} = normalized value for *j*'th instance of *i*'th feature, x_{ij} = actual value for *j*'th instance of *i*'th feature, μ_i = mean of *i*'th feature and σ_i = standard deviation of *i*'th feature for all $i \in 4^k$ and j = total number of ITS sequences in the dataset.

The functions oligonucleotideFrequency from R-package Biostrings [54] and Standard-Scaler from Python library Scikit-Learn [55] have been used to compute *k*-mer frequencies followed by the data normalization.

2.3. Supervised Classifiers

We have considered one deep learning model, CNN and four machine learning algorithms, i.e., SVM, KNN, Naïve-Bayes and Random Forest, in this study. Brief descriptions of each classifier are presented in this section. Finally, RDP Classifier [47] is demonstrated in short.

2.3.1. Convolutional Neural Network (CNN)

It is a deep learning algorithm that can efficiently classify image data. CNN model is also found to be effective for text and DNA sequence classification problems. It consists of an input layer, an output layer and four types of hidden layers: convolution layer, activation layer, pooling layer and fully connected layer. Features are given as inputs to the model through the input layer, and these are sequentially passed through four hidden layers before the final classification or prediction results are obtained from the output layer. Convolution kernels transform the input features and extract useful information from them [56]. In the activation layer, non-linear activation functions, i.e., ReLU, Sigmoid, tanh, etc., are applied to convolved features to handle the non-linearity present in the data. ReLU [57] is the most useful activation function, and its value ranges between 0 and 1. It converts any negative input from the previous layer to 0 so that the corresponding neuron is not activated. The pooling layer reduces the dimension of the input feature space. Finally, the input features are flattened to be represented in the fully connected layer for classification. The number of nodes in the output layer is the same as the number of classes in the input dataset. In this study, a 1D-CNN architecture derived from the original LeNet-5 [58] having two convolution layers of varying kernel sizes and filter numbers, followed by a ReLU activation layer and max pooling layer of size 2×1 with valid padding have been used (Figure 3).

In the hidden layer of the fully connected layer, 500 nodes have been considered with the Softmax activation function [59]. For hyperparameter tuning, the Adam optimizer [60] and Categorical crossentropy [61] loss function has been used. For the training of CNN, 100 epochs and 20 batch sizes have been found to be optimum.

2.3.2. Support Vector Machine (SVM)

SVM [62] is highly efficient in classifying the input dataset into two distinct classes using two parallel hyperplanes. If the distance between these two parallel hyperplanes increases, the classifier becomes more accurate. Multiclass problems are considered a set of several binary classification cases, and classification is done iteratively. Kernel functions like gaussian, radial basis function, tanh, etc., are used for the non-linear transformation of input features. In this study, SVM with RBF kernel function having two parameters as C = 1000 and gamma = 0.0001 has been used.



Figure 3. Architecture of the proposed CNN model. Here, m = size of convolution kernel, $n_1 =$ number of kernels in the first convolution layer and $n_2 =$ number of kernels in the second convolution layer.

2.3.3. K-Nearest Neighbor (KNN)

KNN [63,64] is a simple and non-parametric supervised machine learning classifier for multi-class data. It checks for the similarity between the input feature instance and available data points and classifies the new data into the suitable class. Although KNN is effective and robust for large and noisy training datasets, it is prone to overfitting due to the inability to determine the optimum number of neighbors (*K*). In this study, GridSearch method has been applied to identify the optimum value of *K* for different datasets at various taxonomic levels.

2.3.4. Naïve-Bayes Method

It is a probabilistic supervised classifier suitable for text classification based on the Bayes theorem [65]. The classification probability of a data point in a particular class is defined as the posterior probability, which is calculated as:

$$P(c_i|x) = \frac{P(x|c_i) \times P(c_i)}{P(x)}$$
(2)

For all i = 1, 2, ..., C; C = Total number of classes. $P(c_i|x)$ is the posterior probability of input data x to be classified as a member of i'th class. $P(x|c_i)$ = likelihood of x to be classified in c_i class. $P(c_i)$ and P(x) are prior probabilities for input data x and class c_i respectively. The assigned class for that particular data point is the one for which this posterior probability is found to be the highest. In this study, Gaussian Naïve-Bayes function has been used.

2.3.5. Random Forest (RF)

Random forest [66] is an ensemble machine-learning classifier. Initially, a large number of candidate trees are generated from the bootstrap replicate sample of the input training dataset, and final class is decided based on the majority voting method. In this study, we have considered 500 initial candidate trees and gini entropy function.

2.3.6. RDP Classifier

RDP Classifier [47] uses the Naïve-Bayes algorithm for rapid taxonomy assignment to microbial rRNA sequences based on 8-mer frequency features. It is a Java-based classifier,

and source codes are available at https://github.com/rdpstaff/classifier (accessed on 20 October 2022). Before classifying new datasets, the classifier is needed to be retrained using available reference datasets from https://sourceforge.net/projects/rdp-classifier/files/RDP_Classifier_TrainingData/ (accessed on 20 October 2022) or using reference training datasets as per the choice of the researcher.

2.4. Training and Evaluation

Each dataset described in Tables 2 and 3 is initially split into training and test dataset using unique categories as strata in a 90:10 ratio. Models are trained with the training dataset in a stratified 10-fold cross-validation procedure. In each fold, 90% of the training dataset is used for training the model, and the remaining 10% is used for validation. Values of evaluation metrics have been computed by taking averages of all 10 folds for the validation dataset. Final model performances are evaluated using the test dataset (Figure 4).



Figure 4. Pipeline for training and evaluation of CNN models starting from feature matrix generation from FASTA file of ITS sequences.

For the comparative study between CNN and other machine learning models, datasets with 100 sequences per category for class, order, family, genus and species levels and 500 sequences per category for phylum level are taken under consideration (Table 2). Training of machine learning models is done in a similar procedure mentioned before. Performances of all supervised classifiers have been evaluated based on the following metrics [67,68]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
(3)

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN}$$
(5)

$$F1 \ Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$
(6)

$$MCC \ Score = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$
(7)

where *TP*, *FP*, *TN* and *FN* are the number of True Positive, False Positive, True Negative and False Negative predictions, respectively. In this study, balanced datasets have been used. Hence, all the above-mentioned performance matrices have been found to perform equally well. Furthermore, the Multi-Criteria Decision-Making (MCDM) approach [69] and TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) [70] have been applied to rank all the learning algorithms based on the aforementioned five evaluation metrics. R package Topsis [71] has been used for this analysis.

2.5. Comparison of CNN with Existing Fungi ITS Classification Software

The classification performance of CNN has been compared with RDP Classifier based on datasets having 100 sequences per category at each taxonomic level (Table 2). For CNN, 6-mer features have been considered. RDP Classifier has been retrained with the training dataset (90%) and evaluated using the testing dataset (10%). 10-fold cross-validation has been carried out.

Comparison of CNN model with existing four fungal barcode classification software, namely, funbarRF, RDP Classifier, Mothur and SINTAX, have also been conducted. ITS dataset of 13,630 sequences corresponding to 1363 fungal species (10 sequences per class) from funbarRF paper [50] has been used here. Evaluation has been done based on accuracy as well as the species identification success rate (*SISR*) [50]. *SISR* is calculated as follows:

$$SISR = \frac{\sum_{h=1}^{H} n_h}{\sum_{h=1}^{H} N_h}$$
(8)

Here, h = total number of distinct species/classes in the dataset = 1, 2, ..., H. N_h is the total number of sequences belonging to hth class, and n_h the number of sequences correctly classified into hth class.

2.6. Implementation Details

Training and validation of all classifiers have been performed in Python 3.8.7 using Keras [72] library with Tensorflow backend [73]. RDP Classifier is downloaded from https://github.com/rdpstaff/classifier (accessed on 21 October 2022) [47]. The dataset used for comparing CNN with existing software is available at http://cabgrid.res.in: 8080/funbarrf/dataset/ (accessed on 17 November 2022) [50]. All computational analyses have been carried out in a high-performance computing cluster of ASHOKA bio-computing resources at ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, with the following configuration:

- CPU: 64-bit Intel(R)-Xeon(R), 2.9 GHz, 1 TB;
- RAM: 128 GB;
- OS: Linux RedHat.

3. Results and Discussion

3.1. Effect of Convolution Kernel Size and Filter Numbers on CNN Model Performance

Species-level balanced dataset with 4-mer frequency features having category-wise frequency 100 (Table 2) has been used to compare the performance of various CNN model architectures for varying convolution kernel sizes and filter numbers. The dataset contains 42,900 ITS sequences belonging to 429 unique fungal species. Initially, the CNN architecture (i.e., CNN-1 in Table 4), as described in previous studies [53,74], is applied to the dataset. Afterward, the kernel size has reduced in each of these two convolution layers from five to three, and the number of filters in each layer has increased. In the second convolution layer, the number of filters is kept as double of the first convolution layer. It is observed that the improvements in all the evaluation metrics scores (i.e., accuracy, precision, recall, F1 score and MCC score) for three CNN architectures (i.e., CNN-2, CNN-3 and CNN-4 in Table 4) are very less, i.e., only ~1–2% increase in performance metrics scores as compared to the CNN-1 model architecture (Table 4) proposed earlier [53,74]. It is possibly due to very high sequence similarity (~99%) among 429 species and a lesser amount of data per category. Another possible reason may be the consideration of a lower *k* value. For larger *k* values, such differences will be more prominent due to the availability of more training data per category with the increased number of feature descriptors.

However, the CNN-3 model with two convolution layers of kernel size = 3 and numbers of filters = 32 and 64 in the first and second convolution layers, respectively, have been observed to generate the best classification performances in terms of accuracy, precision, recall, *F1 score* and *MCC score* on the test dataset (Table 4). Further increase in filter numbers has resulted in reduced model performance. Hence, the result indicates that the CNN-3 model is the optimum architecture for fungi ITS classification. This architecture has been used for other comparative analyses.

CNN Model							
Name	Kernel Size (m)	No. of Filters in 1st and 2nd Convolution Layers (n ₁ , n ₂)	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	MCC Score (%)
CNN-1	5	5, 10	87.88	87.74	87.86	87.84	87.85
CNN-2	3	16, 32	88.39	89.10	88.37	88.23	88.37
CNN-3	3	32, 64	89.14	89.78	89.13	89.01	89.11
CNN-4	3	64, 128	88.16	88.69	88.15	87.98	88.13

Table 4. Effect of kernel size and filter numbers on CNN model performance for classifying 100 sequences per species dataset with 4-mer features. Highest values are highlighted in bold.

3.2. Impact of Diversity and k-mer Sizes on CNN Model Performance

Table 2 shows that when the category-wise data frequency is the lowest, the number of unique categories is the highest in our balanced datasets for all taxonomic levels. However, an increase in the unique category numbers leads to an increase in diversity levels for all taxonomic levels. Therefore, these datasets have been used to train and evaluate the CNN model with 4-mer, 5-mer and 6-mer frequency features to study the impact of diversity degrees as well as *k*-mer sizes. Results are presented in Table 5.

Table 5. Average classification accuracy (%) of CNN model with varying *k*-mer sizes for balanced datasets at all taxonomic levels. Accuracy values obtained from the test dataset are presented in the table, and highest accuracies at each level are highlighted in bold.

Tavanamia Laval	Datasata	Total No. of Lovala	Accuracy	(%) for Different k-	mer Sizes
Taxonomic Level			k = 4	<i>k</i> = 5	<i>k</i> = 6
	S1	2752	84.70	85.34	86.12
	S2	939	86.69	88.31	89.18
Species	S3	429	89.14	89.67	90.00
species	S4	148	89.65	89.81	89.46
	S5	117	88.60	89.14	89.94
	S6	64	93.03	93.03	93.28
	G1	1682	92.98	94.23	94.33
	G2	1293	94.78	95.37	95.39
Genus	G3	841	96.74	97.00	97.10
	G4	580	97.59	97.54	97.83
	F1	520	92.35	93.73	95.23
Family	F2	453	95.32	96.11	96.64
ганшу	F3	369	96.89	97.25	97.49
	F4	301	97.37	97.82	98.21
0.1	O1	149	94.43	95.64	96.17
Order	O2	106	98.00	98.33	98.19
	C1	31	95.48	97.42	97.09
Class	C2	25	98.32	98.88	98.80
Phylum	P1	2	99.00	100.00	100.00

The result has suggested that the highest model accuracy has been obtained at the phylum (100%) level, followed by class (98.88%), order (98.33%), family (98.21%), genus (97.83%) and species (93.28%) levels, respectively (Table 5; Figure 5). It is due to lesser inter-class sequence similarities in higher taxonomic levels (e.g., phylum, class) compared to lower taxonomic levels (e.g., genus, species). The highest average accuracies (Table 5) have been obtained for training datasets containing 500 data per category (i.e., S6, G4, F4, O2, C2 and P1) in each taxonomic level (Table 2), but diversities due to unique categories in these datasets are found to be the lowest among others. The result also indicates that CNN can efficiently classify fungi ITS sequences with >94% accuracy up to the genus level

and >89% accuracy at the species level using 6-mer feature vectors even when the diversity level is very high. Still, the category-wise data frequency is very low (e.g., 50 sequences per category). A previous study [74] revealed that the CNN model (CNN-1 model architecture in Table 4) could classify 1035 yeast species based on 6-mer frequency features belonging to *Ascomycota* and *Basidiomycota* phyla with an average accuracy of 82.84%. This model was found to be the most efficient classifier as compared to the other three classifiers, i.e., Deep Belief Network (DBN) (avg. accuracy = 80.40%), RDP Classifier (avg. accuracy = 71.45%) and BLAST Classifier (avg. accuracy = 81.12%). However, it is observed in the present study that CNN (CNN-3 model architecture in Table 4) has produced >86% accuracy for classifying 2752 different fungal species (i.e., more than twice the number of species classified earlier) with only 20 ITS sequences per class (Table 5; Figure 5).

In four out of six taxonomic levels (species, genus, family and phylum), the highest classification accuracies have been obtained with 6-mer features. It is due to the fact that with higher values of k, the number of distinct feature descriptors is also increasing, resulting in better training of the model. In order and class level, 5-mer features have generated better results than 6-mers, but differences in accuracies are not significant. We have also found similar results for other evaluation metrics in order and class levels (Table S1). The possible reason may be the sampling fluctuation that occurred during the dataset preparation step. Hence, it can be inferred that CNN performs best with hexanucleotide frequency features. However, it has the potential to produce high classification accuracy (e.g., 93.03%) with lower values of k, i.e., k = 4, if class numbers are less (e.g., 64) or category-wise data frequencies (e.g., 500 sequences per class) are high even at the species level (Table 5).

3.3. Classification Performance Analysis Based on Varying Class Frequencies

In the previous analysis, it has been found that CNN performs better for datasets with category-wise higher frequencies. But in those datasets, unique category numbers were not constant. Hence, a separate set of datasets (Table 3) with invariant category numbers has been used for each taxonomic level. In addition, the performances of CNN models have been evaluated based on 6-mer frequency features in all taxonomic levels (Table 6).

It is found from the above results that at higher taxonomic levels (i.e., genus, family, order, class and phylum), an increase in all the performance metrics scores has been observed with the increase in category-wise data frequencies from 100 to 500. At the species level, average scores of all the performance metrics initially decreased when the category-wise data frequency was increased from 100 to 250. The species level is the lowest among all the taxonomic levels, and inter-species sequence similarity is very high (~99%), which indicates that the degree of heterogeneity is much less in species-level datasets. This is probably because the CNN model could not correctly recognize true patterns from the input features when the category-wise data frequency is increased from 100 to 250. Hence the increase in performance metrics scores may not have been observed with the increase in the category-wise data frequencies at the species level compared to other higher taxonomic levels. However, an increase in all the performance scores was observed when the category-wise data frequency was increased from 250 to 500 in the case of species-level datasets (Table 6; Figure 6). Overall, it can be seen that the average values of performance metrics scores are the highest for the largest value of category-wise data frequencies at all levels (Table 6; Figure 6). Due to the presence of more data points per class, better training of model parameters has been possible while capturing the maximum variability present in training datasets. In addition, larger data points are also available in respective validation and test datasets, leading to better model evaluation.



Figure 5. Average accuracy % of CNN for different datasets with varying *k*-mer sizes at all taxonomic levels.

Taxonomic Level	Sequence per	Average Values of Performance Metrics (%)				
(N = No. of Unique Categories)	Category	Accuracy	Precision	Recall	F1	МСС
Species (N = 64)	100 250 500	91.87 91.31 93.28	92.25 91.37 93.36	91.86 91.31 93.28	91.67 91.23 93.22	91.76 91.18 93.18
Genus (N = 580)	100 250 500	96.48 97.18 97.83	96.76 97.32 97.88	96.48 97.18 97.83	96.44 97.17 97.83	96.48 97.17 97.83
Family (N = 301)	100 250 500	96.44 97.42 98.21	96.69 97.55 98.24	96.44 97.42 98.20	96.43 97.42 98.20	96.43 97.41 98.20
Order (N = 106)	100 250 500	96.70 97.32 98.19	97.10 97.48 98.25	96.69 97.32 98.19	96.66 97.32 98.19	96.67 97.30 98.17
Class (N = 25)	100 250 500	97.60 98.40 98.80	97.84 98.52 98.86	97.60 98.41 98.80	97.63 98.42 98.80	97.51 98.34 98.75

Table 6. Performance evaluation of CNN classifiers for datasets with unchanged diversity levels and hexanucleotide (k = 6) frequency features at various taxonomic levels. Highest values are highlighted in bold.

Species (N = 64)



90



Figure 6. Cont.



Accuracy Precision Recall F1 Score

MCC Score



Figure 6. Average values (%) of evaluation metrics obtained from CNN models for balanced datasets at all taxonomic levels with invariant diversity levels and hexamer nucleotide frequency features. In each level, three datasets with 100, 250 and 500 data points per category are considered.

3.4. Comparison among Classification Performances of CNN and Other Machine Learning Algorithms

The classification performance of CNN is compared with SVM, KNN, Naïve-Bayes and Random Forest models using datasets with 100 sequences per category for species to class levels and 500 sequences per category at the phylum level (Table 2) while retaining a large number of classes. Average accuracies produced by these five classifiers for varying k values are presented in Table 7.

Table 7. Average accuracies (%) obtained by CNN, SVM, KNN, Naïve-Bayes and Random Forest classifiers with different *k*-mer sizes. For each taxonomic level (from Class to Species), the training dataset consists of 100 sequences per category. Phylum level dataset consists of 500 sequences per category. Highest values are highlighted in bold.

Taxonomic Level	Malara of L	Average Accuracy (%)				
(N = No. of Unique Categories)	value of <i>k</i>	CNN	SVM	KNN	Naïve-Bayes	Random Forest
Constant	4	89.14	85.8	86.71	74.68	89.12
Species $(N = 420)$	5	89.67	87.35	88.57	75.17	89.34
(1N = 429)	6	90.00	88.02	88.58	76.01	89.68
Conus	4	94.78	92.65	93.1	83.33	94.57
(N = 1293)	5	95.37	92.98	93.52	84.00	95.01
	6	95.39	93.07	93.64	84.12	95.17
Family	4	95.32	94.13	93.07	85.29	94.56
Family	5	96.11	95.22	94.15	85.79	95.58
(1N = 455)	6	96.64	95.31	93.71	85.91	96.64
Order	4	94.43	92.69	91.54	85.40	94.16
Order (N = 140)	5	95.64	94.02	93.69	87.00	95.63
(1N = 149)	6	96.17	95.27	93.73	88.54	96.04
Class	4	95.48	93.62	92.52	88.38	95.45
Class (NJ 21)	5	97.42	95.19	93.44	89.16	97.06
(1N = 51)	6	97.09	95.84	92.11	89.29	97.38
Dharles	4	99.00	98.88	99.00	95.00	99.00
(NL 2)	5	100.00	99.55	99.01	96.00	99.00
(1N = 2)	6	100.00	99.55	100.00	96.52	100.00

Similar to the previous study [74], it is also found that CNN has outperformed all machine learning classifiers at most taxonomic levels, especially at the species level (Table 7; Figure 7). Performances of CNN and Random Forest are found to be at par for all *k*-mer

sizes. It can be observed that all five classifiers have produced the best classification results with hexanucleotide features in the majority of levels (Table 7; Figure 7). Some exceptions have been found for KNN; it obtained ~1% more accuracy with 5-mer features at class and family levels (Table 7). CNN also performed better with 5-mer frequencies at the class level. The average accuracies of CNN models are >2% higher than SVM and KNN classifiers up to the class level with hexanucleotide frequency features. However, it can be noticed that at the species level, CNN has obtained 4% and 3% more accuracies than SVM and KNN, respectively, with tetra-nucleotide frequencies (Table 7). The difference between the accuracy values of CNN and the Naïve-Bayes algorithm is >10% at this level. The Naïve-Bayes classifier developed the lowest accuracies among all classifiers at all levels. However, at higher levels, these accuracy gaps are found to be narrowed down, and the other three machine learning-based classifiers, as well as CNN, have obtained >93% classification accuracies. So, the result suggests that the stability of accuracy is highest for CNN and Random Forest and lowest for the Naïve-Bayes method. Similar results have been observed for mean precision, recall, F1 scores and MCC scores (Table S2). Instead of the Naïve-Bayes algorithm, CNN, SVM, KNN and Random Forest models are further ranked based on the Topsis analysis. It can be observed from the results (Table 8) that apart from the class level, CNN has been ranked first among all four classification algorithms, with the highest score obtained at the other five taxonomic levels. At the phylum level, the same ranks (rank = 1) have been obtained from CNN, KNN and Random Forest with the highest scores of 1.0000.

Therefore, it can be inferred from the above results that CNN is the most efficient algorithm for classifying fungi ITS sequences compared to machine learning-based supervised classifiers.



Figure 7. Average accuracies (%) obtained by CNN, SVM, KNN, Naïve-Bayes and Random Forest classifiers with 6-mer features at various taxonomic levels.

Taxonomic Level	Model	Score	Rank
	CNN	1.0000	1
Smaatias	SVM	0.1701	4
Species	KNN	0.2335	3
	Random Forest	0.7877	2
	CNN	0.9978	1
Carrier	SVM	0.1201	4
Genus	KNN	0.2171	3
	Random Forest	0.9043	2
	CNN	1.0000	1
Family	SVM	0.6319	3
	KNN	0.2310	4
	Random Forest	0.9939	2
	CNN	1.0000	1
Orden	SVM	0.6253	3
Order	KNN	0.2011	4
Older	Random Forest	0.9204	2
	CNN	0.9418	2
Class	SVM	0.6928	3
Class	KNN	0.3414	4
	Random Forest	1.0000	1
	CNN	1.0000	1
Phylum	SVM	0.7824	2
1 ftyfulli	KNN	1.0000	1
	Random Forest	1.0000	1

Table 8. Ranking of CNN, SVM, KNN and Random Forest classifiers based on Topsis analysis at various taxonomic levels. Performances of all models have been evaluated based on five evaluation metrics (accuracy, precision, recall, *F1 score* and *MCC score*) with equal weightage using 6-mer frequency features. Highest ranks are highlighted in bold.

3.5. Comparison of CNN with Existing ITS Classification Software

The classification performances of the RDP Classifier (k = 8 mer) [47] and CNN (k = 6 mer) have been compared using datasets with 100 sequences per category (Table 2). Table 9 and Figure 8a shows that the accuracy of the CNN model is ~10% more than RDP Classifiers from species to class level. At the phylum level, CNN has obtained 4% more accuracy than RDP Classifier.

Table 9. Average classification accuracy (%) of CNN and RDP Classifier.

Classifier	Taxonomic Levels					
Classifier	Species	Genus	Family	Order	Class	Phylum
CNN (k = 6)	90	95.39	96.64	96.17	97.09	100
RDP Classifier $(k = 8)$	76.18	85.50	86.20	88.69	90.13	96.41

However, the training time of the CNN model was much higher than RDP Classifier. Apart from this time complexity, it can be inferred that CNN is more efficient than RDP classifier as it can produce higher accuracy scores with fewer features at all taxonomic levels. Moreover, with the availability of advanced computational facilities such as powerful processors and efficient memory allocation, the time complexity problem can be overcome.

Further, the test dataset of the funbarRF paper [50] has been used to evaluate and compare the performance of the CNN model with funbarRF [50], Mothur [48], RDP Classifier [43] and SINTAX [49] in terms of accuracy score and species identification success rates



(SISR) (Table 10; Figure 8b). The scores of the performance matrices for the existing four software are obtained from the above-mentioned paper.



Figure 8. (a) Comparison between CNN and RDP Classifier based on average accuracy (%) at various taxonomic levels, (b) Comparison between CNN and existing software in terms of SISR (%) scores for funbarRF dataset [50].

From the above result, it can be noticed that with 6-mer features, the CNN model could correctly predict 148 more species than the funbarRF tool resulting in a high SISR of 89.89% (Table 10; Figure 8b). The average accuracy of CNN is also 2% higher than funbarRF software for this dataset. The previous study [50] revealed that funbarRF is better software for fungi species classification than Mothur, RDP Classifier and SINTAX in terms of both accuracy scores and the number of correct predictions. Hence, the above comparative

analysis suggests that the CNN model (i.e., CNN_FunBar) is the most efficient fungal taxonomic classifier compared to all the existing software.

Table 10. Comparison among CNN model and other fungal ITS sequence classification software based on average accuracy (%) and SISR (%) values at the species level. Highest values of average accuracy (%), number of correct predictions and SISR (%) score are highlighted in bold.

Algorithm/ Software	Feature Type	Avg. Accuracy (%)	Correct Predictions	SISR (%)
	k = 4 NF	89.21	12,063	88.50
CNN	k = 5 NF	90.75	12,176	89.33
	k = 6 NF	91.27	12,252	89.89
FunbarRF (Random Forest)	<i>g</i> -spaced di-nucleotide features (g = 1 + 2 + 3 + 4 + 5)	~89	12,104	88.80
Mothur (KNN)	<i>k</i> = 8 NF	~89	12,062	88.49
RDP Classifier (Naïve-Bayes)	k = 8 NF	~87	11,864	87.04
SINTAX (Non-Bayesian)	<i>k</i> = 8 NF	~87	11,887	87.21

4. Conclusions

In this study, impacts of category diversity, *k*-mer size, and category-wise frequencies of ITS sequences on fungal ITS sequence classification by CNN have been assessed based on various performance metrics. UNITE+INSDC dataset has been used to cover a broad range of fungal taxa, and some important data filtering steps have been applied to avoid ambiguities. It is found that CNN architecture with two convolution layers of kernel size 3 and filter numbers 32 and 64 in the first and second layers, respectively, is the optimum one. Our study revealed that CNN had produced the highest classification accuracies for datasets with the highest number of sequences per category based on 6-mer frequency features. However, high accuracies have been obtained even when the dataset contains a large number of unique categories comprising a small number of sequences in each of them. The proposed CNN model architecture has been found to classify more than twice the number of fungal species with $\sim 4\%$ higher classification accuracy as compared to the previously used CNN model, which was found to be the best classifier among CNN, DBN, RDP Classifier and BLAST Classifier [74]. Our model has produced 86.12% average accuracy in classifying 2752 species based on hexamer nucleotide frequency features. The classification accuracy at the species level can further be improved by removing closely related species, cryptic fungi species and species complexes, as ITS sequence is not the most efficient DNA marker for their identification [40,74]. A comparative study has revealed that CNN can outperform machine learning algorithms for classifying fungi ITS barcodes even at the species level. CNN_FunBar is also found to be the most efficient taxonomy classification method among all existing software. Currently, ITS sequences from <1% of 3.8 million fungal species are available in the public domain, and many of these sequences are of poor quality [74]. This study highlights the necessity of retraining the developed algorithm with the availability of authenticated reference barcode datasets having higher category-wise data frequencies for developing better taxonomic classification algorithms.

Supplementary Materials: The following supporting information can be downloaded at: https: //www.mdpi.com/article/10.3390/genes14030634/s1, Table S1: Classification performances of CNN model with different *k* values for balanced datasets at all taxonomic levels; Table S2: Classification performances of CNN, SVM, KNN, Naïve-Bayes and Random Forest models for different *k* values at all taxonomic levels.

Author Contributions: Conceptualization, R.D., A.R. and D.C.M.; Data curation, R.D.; Investigation, R.D.; Methodology, R.D.; Formal Analysis, R.D. and D.C.M., Validation, R.D. and D.C.M.; Visualization, R.D.; Writing—original draft, R.D.; Writing—review & editing, R.D., A.R. and D.C.M.; Supervision, A.R.; Funding acquisition, A.R. All authors have read and agreed to the published version of the manuscript.

Funding: Authors are thankful to Indian Council of Agricultural Research, Ministry of Agriculture and Farmers' Welfare, Govt. of India for providing financial assistance from the Network Project on Agricultural Bioinformatics and Computational Biology to carry out this present research work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are collected from the public domain, viz., https://unite.ut.ee/#main (accessed on 20 October 2022) and http://cabgrid.res.in:8080/funbarrf/dataset/ (accessed on 17 November 2022).

Acknowledgments: Authors duly acknowledge the support and assistance from the Dean, PG School, ICAR-Indian Agricultural Research Institute, and the Director, ICAR-Indian Agricultural Statistics Research Institute. Authors are also thankful to the Indian Council of Agricultural Research, Ministry of Agriculture and Farmers' Welfare, Govt. of India for the creation of the Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility where the work was carried out.

Conflicts of Interest: Authors declare no conflicts of interest.

References

- 1. Ferreira, V.; Elosegi, A.; Tiegs, S.D.; Von Schiller, D.; Young, R. Organic Matter Decomposition and Ecosystem Metabolism as Tools to Assess the Functional Integrity of Streams and Rivers—A Systematic Review. *Water* **2020**, *12*, 3523. [CrossRef]
- Bhattacharyya, S.S.; Ros, G.H.; Furtak, K.; Iqbal, H.M.N.; Parra-Saldívar, R. Soil carbon sequestration—An interplay between soil microbial community and soil organic matter dynamics. *Sci. Total. Environ.* 2022, *815*, 152928. [CrossRef] [PubMed]
- 3. Chukwuneme, C.; Ayangbenro, A.; Babalola, O. Metagenomic Analyses of Plant Growth-Promoting and Carbon-Cycling Genes in Maize Rhizosphere Soils with Distinct Land-Use and Management Histories. *Genes* **2021**, *12*, 1431. [CrossRef] [PubMed]
- 4. Enebe, M.; Babalola, O. The Influence of Soil Fertilization on the Distribution and Diversity of Phosphorus Cycling Genes and Microbes Community of Maize Rhizosphere Using Shotgun Metagenomics. *Genes* **2021**, *12*, 1022. [CrossRef] [PubMed]
- Aasfar, A.; Bargaz, A.; Yaakoubi, K.; Hilali, A.; Bennis, I.; Zeroual, Y.; Kadmiri, I.M. Nitrogen Fixing Azotobacter Species as Potential Soil Biological Enhancers for Crop Nutrition and Yield Stability. Front. Microbiol. 2021, 12, 628379. [CrossRef] [PubMed]
- 6. Bloch, S.E.; Ryu, M.-H.; Ozaydin, B.; Broglie, R. Harnessing atmospheric nitrogen for cereal crop production. *Curr. Opin. Biotechnol.* **2020**, *62*, 181–188. [CrossRef]
- Dixit, R.; Wasiullah; Malaviya, D.; Pandiyan, K.; Singh, U.B.; Sahu, A.; Shukla, R.; Singh, B.P.; Rai, J.P.; Sharma, P.K.; et al. Bioremediation of Heavy Metals from Soil and Aquatic Environment: An Overview of Principles and Criteria of Fundamental Processes. *Sustainability* 2015, 7, 2189–2212. [CrossRef]
- Behera, B.K.; Chakraborty, H.J.; Patra, B.; Rout, A.K.; Dehury, B.; Das, B.K.; Sarkar, D.J.; Parida, P.K.; Raman, R.K.; Rao, A.R.; et al. Metagenomic Analysis Reveals Bacterial and Fungal Diversity and Their Bioremediation Potential from Sediments of River Ganga and Yamuna in India. *Front. Microbiol.* 2020, *11*, 556136. [CrossRef]
- Behera, B.K.; Das, A.; Sarkar, D.J.; Weerathunge, P.; Parida, P.K.; Das, B.K.; Thavamani, P.; Ramanathan, R.; Bansal, V. Polycyclic Aromatic Hydrocarbons (PAHs) in inland aquatic ecosystems: Perils and remedies through biosensors and bioremediation. *Environ. Pollut.* 2018, 241, 212–233. [CrossRef]
- Pal, A.K.; Singh, J.; Soni, R.; Tripathi, P.; Kamle, M.; Tripathi, V.; Kumar, P. The role of microorganism in bioremediation for sustainable environment management. In *Bioremediation of Pollutants*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 227–249. [CrossRef]
- Steele, J.A.; Countway, P.D.; Xia, L.; Vigil, P.D.; Beman, J.M.; Kim, D.Y.; Chow, C.-E.T.; Sachdeva, R.; Jones, A.C.; Schwalbach, M.S.; et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.* 2011, *5*, 1414–1425. [CrossRef]
- 12. Barberán, A.; Bates, S.T.; Casamayor, E.O.; Fierer, N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* 2012, *6*, 343–351. [CrossRef] [PubMed]

- 13. Schmieder, R.; Edwards, R. Insights into antibiotic resistance through metagenomic approaches. *Futur. Microbiol.* **2012**, *7*, 73–89. [CrossRef] [PubMed]
- 14. Berendsen, R.L.; Pieterse, C.M.J.; Bakker, P.A.H.M. The rhizosphere microbiome and plant health. *Trends Plant Sci.* 2012, 17, 478–486. [CrossRef] [PubMed]
- 15. Igiehon, N.; Babalola, O. Rhizosphere Microbiome Modulators: Contributions of Nitrogen Fixing Bacteria towards Sustainable Agriculture. *Int. J. Environ. Res. Public Health* **2018**, *15*, 574. [CrossRef] [PubMed]
- Miller, S.; Chiu, C. The Role of Metagenomics and Next-Generation Sequencing in Infectious Disease Diagnosis. *Clin. Chem.* 2021, 68, 115–124. [CrossRef] [PubMed]
- 17. Niu, S.-Y.; Yang, J.; McDermaid, A.; Zhao, J.; Kang, Y.; Ma, Q. Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief. Bioinform.* **2017**, *19*, 1415–1429. [CrossRef]
- Breitwieser, F.P.; Lu, J.; Salzberg, S.L. A review of methods and databases for metagenomic classification and assembly. *Briefings Bioinform.* 2017, 20, 1125–1136. [CrossRef]
- 19. Navgire, G.S.; Goel, N.; Sawhney, G.; Sharma, M.; Kaushik, P.; Mohanta, Y.K.; Mohanta, T.K.; Al-Harrasi, A. Analysis and Interpretation of metagenomics data: An approach. *Biol. Proced. Online* **2022**, *24*, 18. [CrossRef]
- Poretsky, R.; Rodriguez-R, L.M.; Luo, C.; Tsementzi, D.; Konstantinidis, K.T. Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS ONE* 2014, 9, e93827. [CrossRef]
- Liu, Y.-X.; Qin, Y.; Chen, T.; Lu, M.; Qian, X.; Guo, X.; Bai, Y. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* 2021, 12, 315–330. [CrossRef]
- Tonge, D.P.; Pashley, C.H.; Gant, T.W. Amplicon –Based Metagenomic Analysis of Mixed Fungal Samples Using Proton Release Amplicon Sequencing. *PLoS ONE* 2014, 9, e93849. [CrossRef] [PubMed]
- 23. Mbareche, H.; Veillette, M.; Bilodeau, G.; Duchaine, C. Comparison of the performance of ITS1 and ITS2 as barcodes in amplicon-based sequencing of bioaerosols. *Peerj* 2020, *8*, e8523. [CrossRef] [PubMed]
- 24. Woese, C.R.; Fox, G.E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **1977**, 74, 5088–5090. [CrossRef] [PubMed]
- 25. Schleifer, K.H. Classification of Bacteria and Archaea: Past, present and future. Syst. Appl. Microbiol. 2009, 32, 533–542. [CrossRef]
- 26. Chakraborty, C.; Doss, C.G.P.; Patra, B.C.; Bandyopadhyay, S. DNA barcoding to map the microbial communities: Current advances and future directions. *Appl. Microbiol. Biotechnol.* **2014**, *98*, 3425–3436. [CrossRef]
- Igiehon, N.O.; Babalola, O.O. Below-ground-above-ground Plant-microbial Interactions: Focusing on Soybean, Rhizobacteria and Mycorrhizal Fungi. Open Microbiol. J. 2018, 12, 261–279. [CrossRef]
- 28. Hol, W.H.G.; Bezemer, T.M.; Biere, A. Getting the ecology into interactions between plants and the plant growth-promoting bacterium Pseudomonas fluorescens. *Front. Plant Sci.* **2013**, *4*, 81. [CrossRef]
- 29. Hartley, S.E.; Gange, A.C. Impacts of Plant Symbiotic Fungi on Insect Herbivores: Mutualism in a Multitrophic Context. *Annu. Rev. Entomol.* **2009**, *54*, 323–342. [CrossRef]
- 30. Martin, F.M.; Uroz, S.; Barker, D.G. Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria. *Science* 2017, 356, eaad4501. [CrossRef]
- Lindahl, B.D.; Tunlid, A. Ectomycorrhizal fungi—Potential organic matter decomposers, yet not saprotrophs. New Phytol. 2015, 205, 1443–1447. [CrossRef]
- Singh, T.; Singh, A.P. White and Brown Rot Fungi as Decomposers of Lignocellulosic Materials and Their Role in Waste and Pollution Control. In *Fungal Applications in Sustainable Environmental Biotechnology*; Springer: Cham, Switzerland, 2016; pp. 233–247. [CrossRef]
- Kück, U.; Bloemendal, S.; Teichert, I. Putting Fungi to Work: Harvesting a Cornucopia of Drugs, Toxins, and Antibiotics. *PLoS Pathog.* 2014, 10, e1003950. [CrossRef] [PubMed]
- Silber, J.; Kramer, A.; Labes, A.; Tasdemir, D. From Discovery to Production: Biotechnology of Marine Fungi for the Production of New Antibiotics. *Mar. Drugs* 2016, 14, 137. [CrossRef] [PubMed]
- McCartney, H.A.; Foster, S.J.; Fraaije, B.A.; Ward, E. Molecular diagnostics for fungal plant pathogens. *Pest Manag. Sci.* 2003, 59, 129–142. [CrossRef] [PubMed]
- Thambugala, K.M.; Daranagama, D.A.; Phillips, A.J.L.; Kannangara, S.D.; Promputtha, I. Fungi vs. Fungi in Biocontrol: An Overview of Fungal Antagonists Applied Against Fungal Plant Pathogens. *Front. Cell. Infect. Microbiol.* 2020, 10, 604923. [CrossRef]
- 37. Naranjo-Ortiz, M.Á.; Gabaldón, T. Fungal evolution: Diversity, taxonomy and phylogeny of the Fungi. *Biol. Rev.* 2019, 94, 2101–2137. [CrossRef] [PubMed]
- Grządziel, J.; Gałązka, A. Fungal Biodiversity of the Most Common Types of Polish Soil in a Long-Term Microplot Experiment. Front. Microbiol. 2019, 10, 6. [CrossRef] [PubMed]
- Nagy, L.G.; Szöllősi, G. Fungal Phylogeny in the Age of Genomics: Insights into Phylogenetic Inference from Genome-Scale Datasets. Adv. Genet. 2017, 100, 49–72. [CrossRef] [PubMed]
- Schoch, C.L.; Seifert, K.A.; Huhndorf, S.; Robert, V.; Spouge, J.L.; Levesque, C.A.; Chen, W.; Bolchacova, E.; Voigt, K.; Crous, P.W.; et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. USA* 2012, 109, 6241–6246. [CrossRef] [PubMed]

- 41. Toju, H.; Tanabe, A.; Yamamoto, S.; Sato, H. High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. *PLoS ONE* **2012**, *7*, e40863. [CrossRef]
- Kõljalg, U.; Nilsson, R.H.; Abarenkov, K.; Tedersoo, L.; Taylor, A.F.S.; Bahram, M.; Bates, S.T.; Bruns, T.D.; Bengtsson-Palme, J.; Callaghan, T.M.; et al. Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* 2013, 22, 5271–5277. [CrossRef] [PubMed]
- Deshpande, V.; Wang, Q.; Greenfield, P.; Charleston, M.; Porras-Alfaro, A.; Kuske, C.R.; Cole, J.R.; Midgley, D.J.; Tran-Dinh, N. Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia* 2016, 108, 1–5. [CrossRef] [PubMed]
- Ratnasingham, S.; Hebert, P.D.N. BARCODING: BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol. Ecol. Notes 2007, 7, 355–364. [CrossRef] [PubMed]
- Delgado-Serrano, L.; Restrepo, S.; Bustos, J.R.; Zambrano, M.M.; Anzola, J.M. Mycofier: A new machine learning-based classifier for fungal ITS sequences. *BMC Res. Notes* 2016, 9, 402. [CrossRef] [PubMed]
- Altschul, S. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 1997, 25, 3389–3402. [CrossRef]
- Wang, Q.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* 2007, 73, 5261–5267. [CrossRef]
- Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* 2009, 75, 7537–7541. [CrossRef]
- 49. Edgar, R.C. SINTAX: A Simple Non-Bayesian Taxonomy Classifier for 16S and ITS Sequences. bioRxiv 2016, 074161. [CrossRef]
- 50. Meher, P.K.; Sahu, T.K.; Gahoi, S.; Tomar, R.; Rao, A.R. funbarRF: DNA barcode-based fungal species prediction using multiclass Random Forest supervised learning model. *BMC Genet.* **2019**, *20*, 2. [CrossRef]
- Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q.S. Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. AAPS J. 2018, 20, 58. [CrossRef]
- 52. Kotsiopoulos, T.; Sarigiannidis, P.; Ioannidis, D.; Tzovaras, D. Machine Learning and Deep Learning in smart manufacturing: The Smart Grid paradigm. *Comput. Sci. Rev.* **2021**, *40*, 100341. [CrossRef]
- 53. Fiannaca, A.; La Paglia, L.; La Rosa, M.; Lo Bosco, G.; Renda, G.; Rizzo, R.; Gaglio, S.; Urso, A. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinform.* **2018**, *19*, 198. [CrossRef]
- 54. Pagès, H.; Aboyoun, P.; Gentleman, R.; DebRoy, S. Biostrings: Efficient Manipulation of Biological Strings. 2022. Available online: https://bioconductor.org/packages/Biostrings.%0A%0A (accessed on 25 October 2022).
- 55. Pedregosa, F.; Gaël, V.; Alexandre, G.; Vincent, M.; Bertrand, T.; Olivier, G.; Mathieu, B.; Al, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, *12*, 2825–2830.
- 56. Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z. A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals. *Sensors* **2017**, *17*, 425. [CrossRef] [PubMed]
- 57. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th international conference on machine learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814. [CrossRef]
- 58. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, 86, 2278–2324. [CrossRef]
- Lücke, J.; Sahani, M. Generalized Softmax Networks for Non-linear Component Extraction. In Artificial Neural Networks—ICANN 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 657–667. [CrossRef]
- 60. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. arXiv 2014, arXiv:1412.6980.
- 61. Zhang, Z.; Sabuncu, M.R. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *arXiv* 2018, arXiv:1805.07836.
- 62. Vapnik, V. The Nature of Statistical Learning Theory; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1995.
- 63. Crisci, C.; Ghattas, B.; Perera, G. A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Model.* **2012**, 240, 113–122. [CrossRef]
- Hota, S.; Pathak, S. KNN classifier based approach for multi-class sentiment analysis of twitter data. *Int. J. Eng. Technol.* 2018, 7, 1372. [CrossRef]
- 65. Schulman, P. Bayes' Theorem—A Review. Cardiol. Clin. 1984, 2, 319–328. [CrossRef]
- 66. Breiman, L. Random forests. *Mach. Learn.* 2001, 45, 5–32. [CrossRef]
- Yacouby, R.; Axman, D. Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online, 20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 79–91. [CrossRef]
- 68. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
- 69. Triantaphyllou, E. Multi-criteria decision making methods. In *Multi-Criteria Decision Making Methods: A Comparative Study;* Springer: Boston, MA, USA, 2000; pp. 5–21. [CrossRef]
- 70. Lai, Y.-J.; Liu, T.-Y.; Hwang, C.-L. TOPSIS for MODM. Eur. J. Oper. Res. 1994, 76, 486–500. [CrossRef]

- 72. Gulli, A.; Pal, S. Deep Learning with Keras; Packt Publishing Ltd.: Birmingham, UK, 2017.
- 73. TensorFlow. 2021. Available online: https://imagej.net/libs/tensorflow (accessed on 20 October 2022).
- 74. Vu, D.; Groenewald, M.; Verkley, G. Convolutional neural networks improve fungal classification. *Sci. Rep.* **2020**, *10*, 12628. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.