Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
VOLUME:8, ISSUE:2, AUGUST2025, PAPER NUMBER:7

Species Classification Using DNA Barcoding

Ishwarya A[1], Rathi S. M.E. , Ph.D.[2],

[1] Student and [2] Faculty
Dept. of Computer Science Engineering,
Government College of Technology,
Coimbatore, India.
ishwaryaarumugam2000@gmail.com

Abstract: *Recent advancements in sequencing technologies have revolutionized the field of species identification, amplifying the significance of DNA barcodes. This work presents "deep barcoding," a novel deep learning architecture designed to use DNA barcodes for species classification. Utilizing the speed and accessibility of DNA sequences from a wide variety of organisms, this method shows promise in resolving the difficulties associated with species identification. To improve its quality, the CYPRAEIDAE and INGA datasets are pre-processed using nucleotide augmentation. One-hot encoding was then used to convert the dataset into one-dimensional images. Next, a Convolution Neural Network (CNN) was used to classify species. Using stratified k-fold cross-validation, the model's performance was assessed. For CYPRAEIDAE and INGA, the accuracy scores were 96.79%, and 97.68% respectively. This model displays as a useful tool for researchers, conservationists, and other stakeholders as well as a way to address practical issues related to species classification. Through the integration of DNA sequencing and deep learning, this model presents an innovative approach for species.*
Keywords: *DNA Barcoding, CNN, Nucleotide Augmentation, Stratified K-Fold Cross-Validation.*

1.INTRODUCTION:

There are over 8.7 million species in total on Earth, but only about 1.2 million of those are completely assembled using taxonomic classification. The loss of biodiversity has been highlighted as a major environmental issue facing the entire world, and ecologists are constantly changing methods for protecting natural resources and conserving biological diversity. However, the taxonomic obstacles frequently serves as the primary barrier to accessing the biosphere's taxonomic classification. Taxonomists use morphological characteristics to identify biological specimens; still in some difficult circumstances, even experts cannot correctly identify and understand the taxonomic data. Consequently, it has been suggested that this issue can be solved by using genetic information, most especially DNA sequences[1][2] .

DNA barcoding was proposed by Hebert et al. [3] in 2003 as a method of species identification. Certain fragments, derived from brief segments of nuclear, plastid, and mitochondrial DNA, have been designated as DNA Barcodes and can serve as indicators for creatures of the primary domains of life. The Internal Transcribed Spacer (ITS) for fungi [4], rbcL and matK for plants [5], and cytochrome C Oxidase subunit I (COI) for animals [6] are the gene sections selected as barcodes.

Nowadays, it is widely recognised that a DNA barcode provides enough information to identify a specimen to a species, demonstrating significant variability even amongst closely related species [7, 8]. Consequently, starting in 2004, the International Barcode of Life initiative (IBOL) and the With the aim of establishing an online, publicly accessible sequence database, the Consortium for the Barcode of Life

(CBOL) has supported international activities devoted to the establishment of DNA barcoding as a global standard for the recognition of biological species (www.barcodinglife.org).
Species classification with DNA Barcode has been shown to be successful on a variety of organisms and is a method of classifying an unknown specimen to a recognised species by analysing its DNA Barcode sequence. There are numerous methods for identifying species using DNA barcodes. The methods can be divided into four categories: similarity-based taxonomic methods (like BLAST), tree-based taxonomic methods (like neighbour joining), and character-based taxonomy techniques (such as BLOG), and taxonomy techniques based on machine learning (e.g., supervised learning algorithms). In DNA barcode classification, a training set contains of known species is obtained from DNA-barcoded specimens. A test set contains of unknown species [9,10]. In this study a deep learning model called a convolutional neural network (CNN) has shown better results in a number of prediction tasks, particularly in image and sequence processing. When a CNN learns from training data instead of precisely build feature extraction/selection or preprocessing, it becomes more adaptive to the recognition job. CNNs use the features of local connection, parameter sharing, pooling, and the usage of multilayers to identify both local and global textures as important aspects from training data [11].

2. LITERATURE SURVEY:

P.D. Hebert et al., [12] discusses the challenges in identifying species due to a loss of taxonomic knowledge and proposes a method using DNA sequences as "barcodes" based on mitochondrial gene cytochrome c oxidase I (COI). They show that COI profiles can reliably place taxa in their phylum or order, leading to species-level assignments. Additionally, they demonstrate that thorough COI profiles can lead to species-level assignments, and a model profile was 100% accurate in identifying closely related lepidopteran species.

Dlamini et al., [13] analyzed the genomic signatures of the COVID-19 virus, SARS-CoV-2, using the extreme gradient boosting (XGBoost) model. It classified eight pathogenic species with 100% accuracy, revealing similarities between MERS-CoV and SARS-CoV-2. The study also revealed different dinucleotide patterns across six continents, with Oceania having a unique signature.

Shujaat et al., [14] presented the rice promoters are crucial for gene transcription, and accurate prediction is essential for understanding genetic networks and gene expression. A new model, Cr-Prom, uses convolutional neural networks to predict promoters with 99% accuracy. Compared to current rice-specific predictors, Cr-Prom showed promise on a different dataset, with a 98.57% sensitivity, 99.9% specificity, 99.1% accuracy, and 0.9839 MCC.

P.K. Meher et al., [15] discusses the importance of identifying unidentified fungal species for preserving fungal diversity. Traditional morphological identification is often impossible, but DNA barcoding can be efficient. A gapped base pair composition-based Random Forest predictor was developed, achieving over 85% accuracy with at least four sequences per species and 88% accuracy with at least seven sequences per species.
Aoki et al., [16] explores the use of Convolutional Neural Networks (CNNs) to categorize DNA sequences and identify sequence motifs, especially for non-coding RNA (ncRNA) sequences. CNNs are trained on distributed nucleotide representations and accurately cluster sequences through pairwise alignments. It combines map profiles of next-generation sequence reads with secondary structure information unique to ncRNAs and distributed RNA nucleotide representations, achieving up to 98% accuracy.

Srivastava et al., [17] discusses the use of dropout to prevent overfitting in deep neural networks with numerous parameters. Dropout removes units and connections during training,

preventing over adaptation. This technique significantly reduces overfitting and improves neural networks' performance in tasks like speech recognition, document classification, computational biology, and vision.

3. PROPOSED METHOD:

3.1 Dataset:

Datasets from http://dmb.iasi.cnr.it/supbarcodes.php were retrieved. The dataset was collected from the Gen Bank nucleotide database(http://dmb.iasi.cnr.it/supbarcodes.php). Two different types of datasets such that plant and invertebrate. The imbalanced datasets has splitted into two set such as training set and testing set. The training set are used to teach the model and the testing set are used to test the results and measure the quality of the trained model.

DATASETS	TYPE	NO.OF CLASS	TRAINING SET	TESTING SET
Cypraeidae	Invertebrates	211	1656	352
Inga	Plants	63	786	122

Table.1. Datasets Details

3.2 Data Preprocessing:

3.2.1 Nucleotide Augmentation:

Nucleotide augmentation techniques such as random mutations and reverse complementation are used to artificially improve the dataset. There are four nucleotide in the DNA sequence: adenine (A), thymine (T), guanine (G), and cytosine (C)[18].

3.2.2 One Hot Encoding:

The one hot encoding transforms categorical data into a numerical format. A binary vector is used to represent each category, with the majority of its elements set to 0 and only one element set to 1. In order to avoid unintentional ordinal relationships that might damage the accuracy of the model, one-hot encoding maintains category distinction without assuming any ordinal relationships.

3.2.3 Padding:

After one-hot encoding, padding is used to add padding elements to one-hot encoded sequences to ensure they all have the same length.

3.3 Convolution Neural Network:

The CNNs can learn to capture local patterns and features in the input sequences, making them useful for species classification. The CNN contains three types of layers such as input layer, hidden layer and output layer. There are many hidden layers, such as the convolution layer, max pooling layer, dropout layer, dense layer, and fully connected layer.

3.3.1 Convolution Layer:

CNNs are made up of several convolution layers, each of which processes the input image using a different set of learnable filters, or kernels. These filters create feature maps that capture specific details and patterns in the text by swiping over the input and performing element-wise multiplication and summation.

3.3.2 Max Pooling Layer:

Pooling layers are frequently used to down sample the spatial dimensions of feature maps after convolution layers. By pooling the resources, the network becomes more resilient to even tiny changes in the input and helps to lower the computational complexity.

3.3.3 Fully Connected Layer:

CNNs usually include one or more fully connected layers that combine the learned features to make predictions. These layers are frequently utilized in the network's last phases.

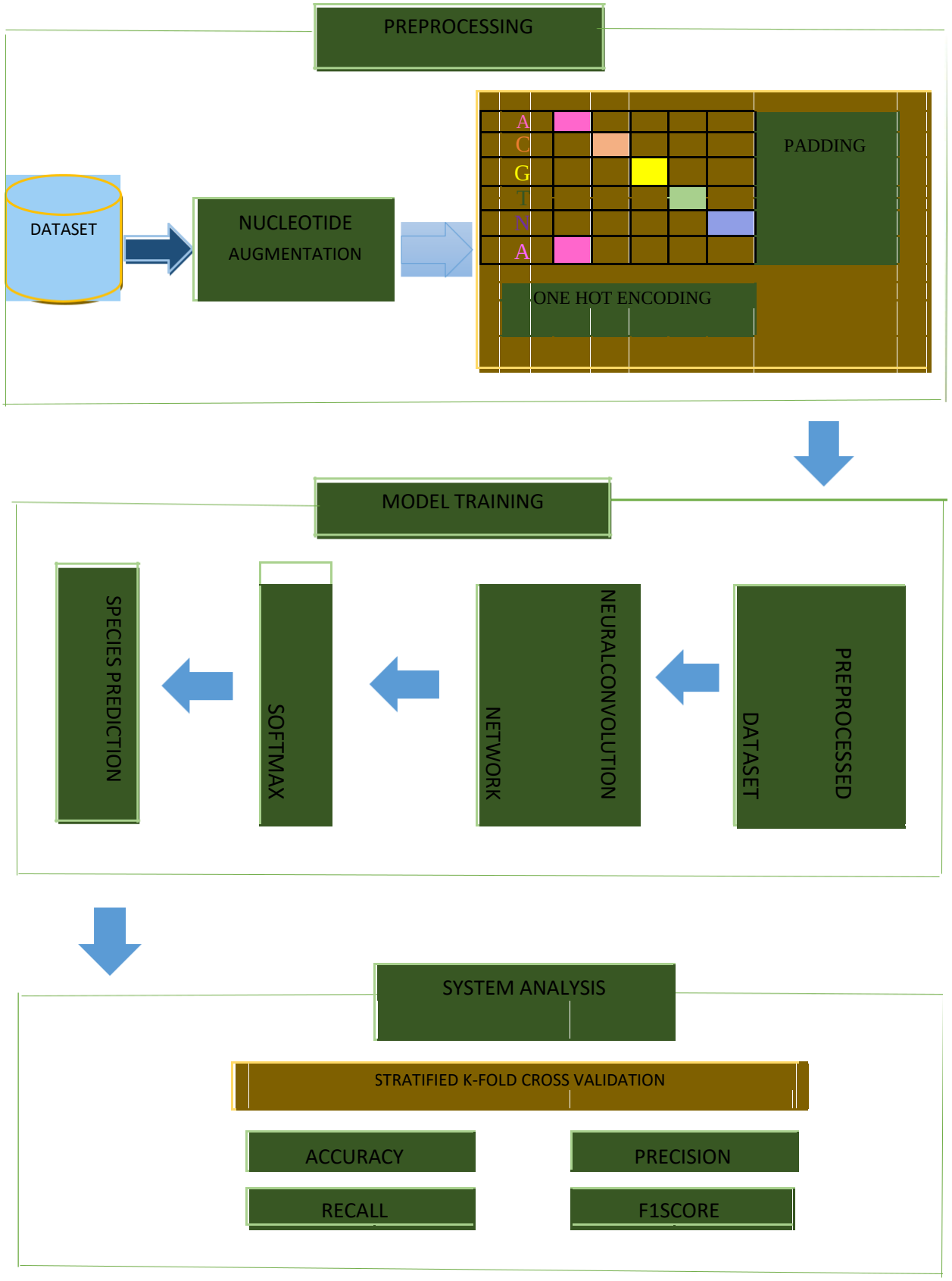

Fig.1 DNA Barcoding Framework

3.3.4 Activation Function:

The Rectified Linear Unit (ReLU) activation function is applied after every convolution and pooling layer. They provide the model non-linearity, which helps it pick up complex patterns and representations.

3.3.5 Dropout Layer:

In neural networks, such as Convolutional Neural Networks (CNNs)and other deep learning models, the Dropout layer is a regularization technique. It is intended to reduce over fitting, promote better generalization of models, and strengthen neural networks' resilience.

Fig 2. CNN Architecture

3.4 Classification Of Species:

We suggest a deep barcoding architecture for species classification based on DNA barcodes. Preprocessing steps are applied to the dataset in order to convert the data into a one-dimensional picture. A Convolutional Neural Network is used to retrieve the features. The categorical cross-entropy loss function is used for optimization. The Adam optimizer makes use of each neuron's. The Softmax function is used as the classifier after calculating the probability of each species as the output, and each species is then appropriately classified.

4. HYPERPARAMETER SETTING:

For all datasets, we used two sets of convolution and max-pooling layers and dropout layers, and the first convolution layer has 16 filters of 5×5 kernel size and the second convolution layer has 36 filters of 5×5 kernel size, respectively, with 2×2 windows of the pooling size at a dropout rate of 0.2. There are three fully connected layers with 1024 neurons and a dropout rate of 0.4. For training, the batch size and epoch sizes were set to 32 and 10, respectively.

5. RESULT ANALYSIS:

5.1 Stratified K-Fold Cross-Validation:

In stratified k-fold cross-validation, the dataset is partitioned into k folds so that the validation data has an equal number of instances of the desired class label. This ensures that, particularly during the dataset is imbalanced; neither the train nor the validation data exhibit an overrepresentation of a particular class. The final score is calculated as the average of the scores for each fold. Its performance is favorable for an unbalanced dataset[19]. In k-fold cross-validation, built accuracy, precision, recall, and f1 score are calculated.

5.1.1 Accuracy:

Accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables. On test data, accuracy is defined as the proportion of accurate predictions to all predictions. The classifier's accuracy score can range from 0 to 1, with 1 indicating that all of its predictions are accurate. The formula to calculate accuracy is,

$$\text{ACCURACY} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

5.1.2 Precision:

The proportion of instances that were correctly identified out of all actually classified instances is known as precision. Every instance of data that is classified as positive and has a precision value of 1 is a positive instance of data. The number of positive cases with the label negative that are projected as positive, it is vital to note, is unaffected by this.

$$\text{PRECISION} = \frac{TP}{TP+FP} \quad (2)$$

5.1.3 Recall:

The proportion of positively classified events among all positively occurring events is represented by recall, also known as sensitivity. It is defined as follows,

$$\text{RECALL} = \frac{TP}{TP+FN} \quad (3)$$

5.1.4 F1 Score:

Precision and recall are not thought to be accurate indicators of a classifier's performance. F1 has been rated more significant because it incorporates recall and precision and provides a score between 0 and 1. It represents the harmonic mean of recall and precision.

$$F1\ SCORE = 2\frac{\text{PRECISION}*\text{RECALL}}{\text{PRECISION}+\text{RECALL}} \quad (4)$$

Data sets	Folds	Accuracy	Precision	Recall	F1 score
CYPRAEIDAE	Fold1	96.14	96.31	96.14	95.66
	Fold2	96.98	97.10	96.98	96.45
	Fold3	96.92	97.24	96.92	96.42
	Fold4	96.80	97.21	96.80	96.47
	Fold5	97.10	97.01	97.10	96.69
	Average	**96.79**	**96.97**	**96.79**	**96.34**
INGA	Fold1	97.46	97.44	97.46	97.05
	Fold2	97.07	97.20	97.07	96.50
	Fold3	97.58	97.33	97.58	97.08
	Fold4	98.35	98.36	98.35	97.87
	Fold5	97.96	97.88	97.96	97.58
	Average	**97.68**	**97.64**	**97.68**	**97.22**

Table 2 .Performance analysis

6.CONCLUSION:

Since supervised learning is the foundation of deep barcoding, the expansion of our model to new species is limited by the availability, quality, and dependence on well-known DNA sequences. This research, however, offers techniques for classifying unknown specimens into known species through the use of DNA barcodes and a trained model. Deep learning has made species classification better, but there are still many issues with using it for DNA barcode analysis. Researchers can examine DNA barcoding in advance with the aid of techniques like deep barcoding, which can eventually clarify species identification based on DNA barcodes.

7.References:

[1] K. H. Chu, C. P. Li, and J. Qi, “Ribosomal RNA as molecular barcodes: A simple correlation analysis without sequence alignment,”Bioinformatics, vol. 22, no. 14, pp. 1690–1701, 2006.

[2] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm,“How many species are there on earth and in the ocean?,” PLoSBiol., vol. 9, no. 8, 2011, Art. no. e1001127

[3] Hebert PDN, Cywinska A, Ball SL, DeWaard J:”Biological identifications through DNA barcodes”. Proc R Soc B 2003, 270:313–321.

[4] Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium:Nuclear ribosomal internal transcribed spacer(ITS) region as a universal DNA barcode marker for Fungi.Proc Natl Acad Sci USA 2012, 109(16):6241–6246.

[5] CBOL Plant Working Group: A DNA barcode for land plants. Proc Natl Acad Sci U S A 2009, 106(31):12794–12797.

[6] Hebert PDN, Ratnasingham S, de Waard J: Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc B 2003, 270(Suppl 1):S96–S99.

[7] Hebert PDN, Gregory T: The promise of DNA barcoding for taxonomy. Syst Biol 2005, 54:852–859.

[8] Schindel D, Miller S: DNA barcoding a useful tool for taxonomists. Nature 2005, 435:17–17.

[9] 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," Mol.Ecol. Resour., vol. 13, no. 6, pp. 1043–1046, 2013.

[10] E. Weitschek, G. Fiscon, and G. Felici, "Supervised DNA Barcodes species classification: Analysis, comparisons and results," BioData Mining, vol. 7, no. 1, 2014, Art. no. 4

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature,vol. 521, no. 7553, pp. 436–444, May 28, 2015.

[12] P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard, "Biological identifications through DNA barcodes," Proc Biol Sci., vol. 270, no. 1512, pp. 313–321, 2003.

[13] Dlamini, Gciniwe S., Stephanie J. Müller, Rebone L. Meraba, Richard A. Young, James Mashiyane, Tapiwa Chiwewe, and Darlington S. Mapiye. "Classification of COVID-19 and other pathogenic sequences: a dinucleotide frequency and machine learning approach." *Ieee Access* 8 (2020): 195263-195273.

[14] Shujaat, Muhammad, Seung Beop Lee, Hilal Tayara, and Kil To Chong. "Cr-prom: A convolutional neural network-based model for the prediction of rice promoters." *IEEE Access* 9 (2021): 81485-81491.

[15] 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 Genet., vol. 20, no. 1, 2019, Art. no. 2

[16] Aoki, Genta, and Yasubumi Sakakibara. "Convolutional neural networks for classification of alignments of non-coding RNA sequences." *Bioinformatics* 34, no. 13 (2018): i237-i244.

[17] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15, no. 1 (2014): 1929-1958.

[18] Minot, Mason, and Sai T. Reddy. "Nucleotide augmentation for machine learning-guided protein engineering." *Bioinformatics Advances* 3, no. 1 (2023): vbac094.

[19] Bawankar, B. U., and Kotadi Chinnaiah. "Implementation of ensemble method on DNA data using various cross validation techniques." *3c Tecnología: glas de innovación aplicadas a la pyme* 11, no. 2 (2022): 59-69