International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

REVOLUTIONIZING DNA SPECIES CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK

[1]S. Bavankumar, [2]Dr. V. Rathikarani, [3]Dr. R. Santhoshkumar

[1]Research Scholar, [2]Assistant Professor, [3]Associate Professor

[1,2]Department of Computer Science and Engineering, Annamalai University, Chidambaram,Tamilnadu, India 608002,

[3]Dept. of CSE, Sree Dattha Group of Educational Institutions, Ibrahimpatnam, Telangana 501510, India

[1]sbavankumar55@gmail.com, [2]rathika_1982@rediffmail.com, rathika_198257@rediffmail.com, [3]santhoshkumar.aucse@gmail.com

Abstract:

Precise DNA species classification is essential in multiple domains, including forensics, medical diagnostics, and archaeology. Conventional classification methods, including sequence alignment techniques, frequently experience significant computational expenses and restricted scalability. This paper introduces a method for DNA species classification utilizing a Convolutional Neural Network (CNN), achieving an outstanding accuracy of 98.4% on established genomic datasets. Our model utilizes CNN's capacity to autonomously extract essential sequence features, facilitating accurate and efficient species identification. DNA sequences undergo preprocessing through one-hot encoding and k-mer embeddings to enhance feature representation. The experimental findings indicate that our CNN model substantially surpasses traditional methods in terms of accuracy and computational efficiency. This method's superior classification performance could transform DNA analysis in forensic investigations, medical research, and archaeological studies, offering a rapid and dependable solution for species identification.

Key Words: DNA Species, Convolutional Neural Network, Gene Family.

1. Introduction

In this era of the genome, where scientific progress has allowed us to probe the mysteries of life, one of the most striking features of molecular biology's recent evolution has been the explosion of biological data, leading to the rapid growth of a massive biological information database. It became apparent that we needed a way to derive practical insights from this deluge of data, so the field of bioinformatics was born. Bioinformatics combines math, computer science, and life science to get biological information from data and help with research that is related to it.

The initial step involves analyzing the DNA sequence of the genome to gain insights into the protein-coding region. The subsequent phase involves executing a simulation to formulate an informed hypothesis regarding the protein's three-dimensional configuration. The researchers utilize the protein's function as a foundation for the final drug design.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

There is more and more proof that the amount of biological data doubles every 18 months. In 1982, GenBank made its first database of nucleic acid sequences. It had 606 sequences and 680,000 nucleotide bases. Biological sequences totaling 162 million entries, or 150 billion nucleotide bases, were entered into the database as of February 2013. Extracting useful information from massive datasets is a major area of study in bioinformatics.

The enhancement of data processing capabilities and the generation of relevant biological information are among the numerous potential applications of machine learning currently being explored in the analysis of sequence data. This review examines data mining and machine learning methodologies applicable to DNA sequences. This section offers a concise history of the advancement of sequencing technology, delineates the structure of DNA sequence data, presents various sequence_encoding techniques employed in machine learning. Furthermore, The foundation of data mining in DNA sequences is sequence similarity, which we stress. All the main machine learning algorithms have been captured, and the data mining procedure has been examined in detail.

Common uses of machine learning with DNA sequence data include pattern mining, clustering, classification, and sequence alignment. The subsequent generalizations stem from our research. Parallel computation and distributed sequence alignment are poised to prevail in laboratory DNA sequence alignment. Identifying the most effective approach to represent sequence characteristics and assess DNA sequence classification poses a considerable challenge within the scientific community. Things are not easy in this case. Knowing how to extract different subsequences from a DNA sequence is crucial for effective clustering. The time and space needed to store the vastly increased number of possible sequence patterns that will result from mining DNA sequence patterns is enormous. The formulation of an effective search methodology and the eradication of redundant repetitions in sequence patterns will be a primary research emphasis in the forthcoming years.

2. Litrature Survey

Only around 1.19 millions of the estimated 8.69 millions species on Earth have undergone official taxonomic classification. Biodiversity loss remains a critical global concern, prompting ecologists to refine conservation strategies and improve natural resource management. A major obstacle to cataloging the planet's biodiversity is the taxonomic impediment, which restricts access to taxonomic knowledge for researchers lacking specialized expertise. An effective solution to this problem has been suggested using genetic data, specifically DNA sequencing.

DNA barcoding was initially suggested in 2003 as a way to identify species, with the COI gene serving as a DNA marker to differentiate between different types of animals. All the genetic information needed to identify a specimen's species is contained in a brief DNA barcode sequence. This method works for figuring out what species are and putting them into groups. Plants are now identified using markers like maturase K(matK) and chloroplast ribulose bisphosphate carboxylase(rbcL). Fungi are usually identified using internal transcribed spacers (ITSs) because COI barcodes are becoming more common in animal taxonomy. Genomic

Received: August 03, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

barcoding enables the accurate classification of unidentified specimens through a comprehensive reference sequence database.

Various species identification methodologies utilizing DNA barcodes have been proposed, broadly categorized into four types: There are a variety of approaches, including those based on trees (like neighbor joining), similarity (like BLAST), characters (like BLOG), and machine learning (like supervised learning classification). Using bioinformatics to identify species via DNA sequence analysis is a complicated process. Some of the classification algorithms that have been looked at for species identification using DNA barcodes are support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), naïve Bayes (NB), multilayer perceptron (MLP), random forest (RF), and hierarchical supervised classifiers.

In DNA barcode classification, a reference library is established utilizing DNA sequences from recognized species. When an unknown query sequence is introduced, it undergoes classification based on comparison with the reference dataset. The classification process follows a supervised learning framework, where reference sequences serve as training data and unknown specimens form the testing set. The training set includes labeled samples from recognized species, whereas the testing set contains sequences from unidentified species for classification purposes.

A classification model is made with the training set in supervised learning, and its accuracy is checked with the testing set. Recent advancements in deep learning and various machine learning methodologies have significantly enhanced numerous bioinformatics applications, including those related to images, survival analyses, signals, and sequences. Convolutional neural networks (CNNs) have shown the most promise for prediction tasks, especially those that involve image processing and sequences. Unlike traditional methods requiring manual feature extraction, CNNs autonomously learn relevant patterns from data through local connectivity, parameter sharing, pooling, and multilayer architectures.

This study presents a CNN model developed to extract DNA sequence patterns directly from training data for the purpose of predicting species identity. The model's performance was assessed in comparison to traditional supervised learning techniques executed with the the scikit-learn package, which includes the following classifiers: RF, DT, SVM, KNN, NB, and MLP. Species identification was conducted using both simulated data and actual sequences of DNA barcodes (ITS, COI, and rbcL genes). The findings demonstrate that our proposed CNN model attained enhanced accuracy and surpassed conventional supervised learning classifiers, establishing a robust framework for DNA barcode sequence analysis.

3. Proposed Methods

3.1 Dataset

To model DNA barcode sequences in Mesquite, we utilized a coalescent package and sourced our datasets from http://dmb.iasi.cnr.it/supbarcodes.php. Datasets from invertebrates, plants, and vertebrates were simulated following the Yule model. Using two key parameters—effective population size (Ne) and species divergence timing—a random ultrametric species tree was generated, representing fifty distinct species. For each species, twenty specimens were

Received: August 03, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

simulated using gene trees with N_e values of N_e-1,000, N_e -10,000, and N_e -50,000. To enhance datasets robustness, each dataset was duplicated 100 times, resulting in a total of 300 simulated datasets. The complexity of these datasets was controlled by varying the Ne parameter. Consistent with the standard length for DNA barcodes, sequences were set at 650 base pairs (bp). A summary of the simulated data is presented in Table 1.

Table 1. A synopsis of the outcome of the model.

N_e	Datasets	Individual	Seq. length	Species
1,000	Ne1000	20	650	50
10,000	Ne10000	20	650	50
50,000	Ne50000	20	650	50

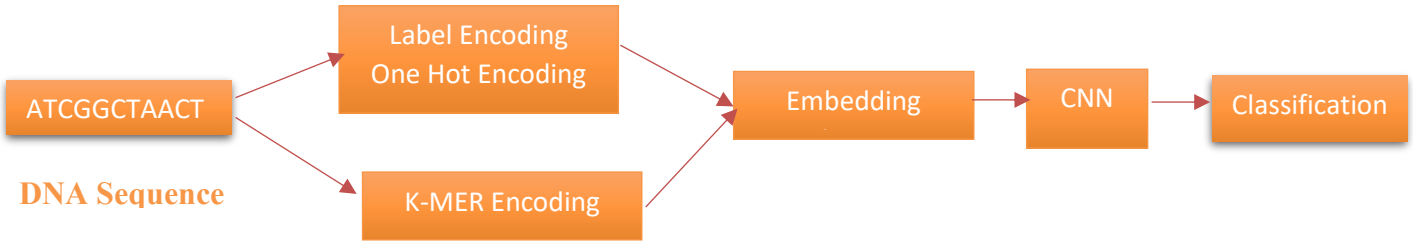


Fig 3.1 Block Diagram of Proposed Methods

3.2 Convolutional Neural Network Architecture (CNN)

- **CNN Introduction:** The text correctly attributes the initial introduction of CNNs to LeCun et al. in 1989. This is a foundational point in deep learning history.
- **Fundamental CNN Architecture:** It outlines the core layers of a basic CNN:
  - **Input Layer:** Where the data (e.g., images) enters the network.
  - **Convolutional Layer:** The heart of CNNs, responsible for feature extraction using filters.
  - **Pooling Layer:** Reduces the size of feature maps, which makes features stronger and uses less processing power.
  - **Fully Connected Layer:** For classification, it connects all the neurons in the previous layer to each neuron in this layer.
  - **Output Layer:** Produces the final prediction (e.g., class probabilities).
- **Marking up Training Data's :**
  - {X(n),y(n)} represents the training dataset.
  - X(n) is an n-dimensional matrix, indicating the input data.
  - X={x1,x2,...,xn} represents the individual data points.
  - y={y1,y2,...,yn} represents the corresponding labels.

Received: August 03, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- yi∈{1,2,...,m} indicates that each label belongs to one of m classes.

Further Considerations:

Simplification: This is a simplified description. Modern CNN architectures can be significantly more complex, involving multiple convolutional and pooling layers, residual connections, batch normalization, and various other techniques.

- **Image Focus:** While CNNs can be used for various data types, the mention of "input layer" and the general structure strongly imply an image processing context.
- **Deep Learning Context:** The text correctly positions CNNs within the broader field of deep learning.
- **Matrix Dimensionality:** While it states X is an n dimensional matrix, it would be more correct to state that X is a matrix containing n samples. The dimensionality of each sample depends on the type of data. For example, for a grayscale image, each sample would likely be a 2D matrix. For a color image, each sample would likely be a 3 dimensional tensor.

In essence, this passage provides a good starting point for understanding the basic structure and notation associated with CNNs.

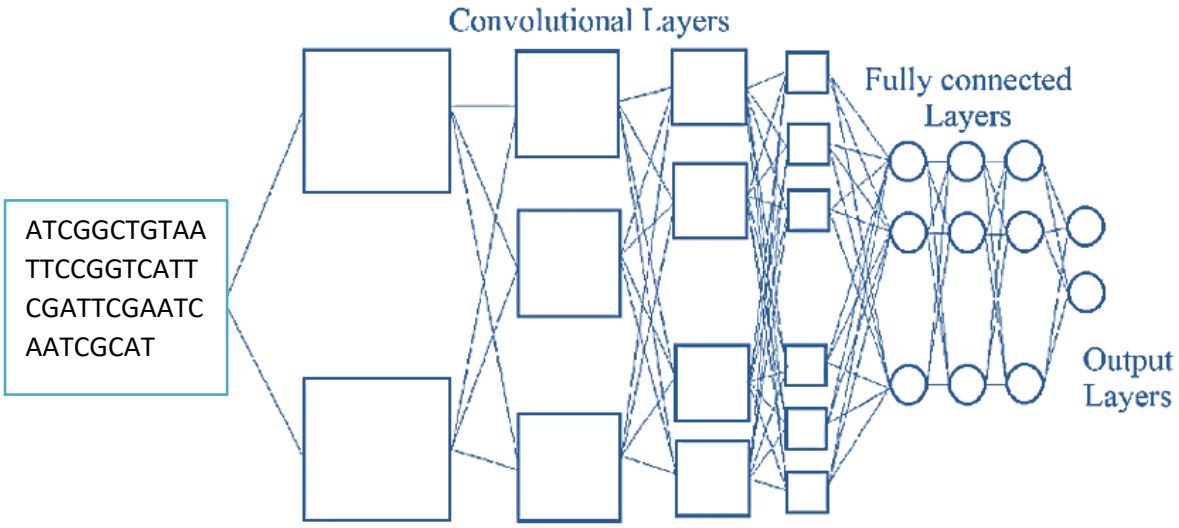


Fig 3.2 Architecture of CNN

3.2.1 Convolutional layer

For the purpose of distributing weight and extracting features, the convolutional layer is utilized as a filters.

3.2.2 Pooling layer

A max-pooling layer is used to make the output dimensions smaller than those of the previous convolutional layer. The max-pooling operation picks the highest value from a group of neuron clusters that are at the output of the convolutional layer.

3.2.3 Fully connected Layer

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The convolutional and pooling layers give the fully connected layer, also known as the Deep Neural Network (DNN), the best parameters. The Deep Neural Network (DNN) is made up of three layers that work together to classify things: the Input_layer, the Hidden_layer, and the Output_layer. The input layer of the Deep Neural Network (DNN) is changed to include the previous outputs of feature maps. To perform a nonlinear transformation utilizing the ReLU activation function, h neurons are utilized in the hidden layer. The output is evaluated by the softmax function by finding the probabilities of all possible outcomes. On the whole training dataset {X(n), Y(n)}, the cross-entropy loss function is minimized to obtain the optimal parameter for the neural network model. The Adam optimizer algorithm finds the best values for the loss function by using back-propagation.

3.2.4 Species taxonomy using DNA barcodes

Here we introduce DeepBarcoding, a deep learning framework for DNA barcode species classification. The DeepBarcoding architecture for species classification using DNA barcode sequences outlines the five-step workflow in Figure 1.

1. **Data Processing**: GeneBank is consulted for DNA barcode sequences, which are then aligned in MEGA software using the Muscle algorithm. Alignment guarantees consistency, which is important because sequence lengths differ based on GenBank accession numbers.
2. **Encoding DNA Sequences**: The four nucleotide bases (adenine (A), cytosine (C), guanine (G), and thymine (T)) are changed into a one-hot encoded format. This makes the sequence look like a structured one-dimensional image that can be processed further.
3. **Feature Extraction with Convolutional and Pooling Layers**: The one-hot encoded sequences are looked at using convolutional and max-pooling layers. This step uses n filters that are m*m in size and a pooling window that is l*l in size to get k feature matrices. The ReLU activation function is then used to flatten these matrices into a vector. For better feature extraction, the architecture may have many convolutional and pooling layers.
4. **Deep Neural Network for Feature Integration**: A fully connected deep neural network processes the features that were taken out and improves the classification performance. We use the Adam optimizer to optimize the loss function, and we use ReLU activation functions on each neuron.
5. **Species Classification with Softmax**: The Softmax function serves as the final classifier, calculating the probability distribution among species categories to guarantee precise classification.

DeepBarcoding was implemented using the **Keras** toolkit (https://keras.io/), built on **TensorFlow** in Python.

3.3.4 Hyperparameter settings

We used two layers in all of the datasets: the max-pooling layer and the convolutional layer. The hyperparameters for the first layer were set to 16 filters with a 5 × 5 kernel size, and for the second layer, they were set to 36 filters with the same kernel size. We also used pooling windows that were 2 × 2 and a dropout rate of 0.2. The three-layer deep neural network had a

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

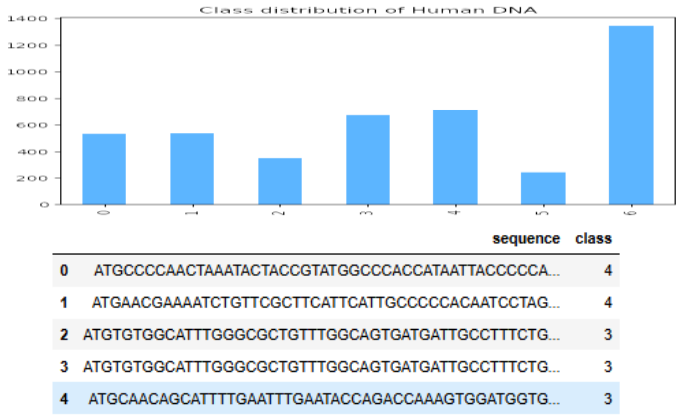
dropout rate of 0.4 and a total of 1024 neurons. For training, the mini-batch size was set at ten and the epoch size was set at one hundred. We used a grid search on the simulation datasets to find the parameters.

4. Results and Discussion

Table 2. Gene Family with numbers and class_label

Gene_family	Numbers	Class_label
Transcription factor	1343	6
Synthase	711	4
Synthetase	672	3
Tyrosine kinase	534	1
G protein coupled receptors	531	0
Tyrosine phosphatase	349	2
Ion channel	240	5

We have just retrieved it and identified its category; thus, it relates to the gene family. A sequence classified as class 0 signifies its membership in the gene family associated with G protein-coupled receptors. This number signifies that class label 0 is present at 531 and persists thereafter. A gene family called G protein-coupled receptors encompasses this specific sequence. K-mer counting was employed in this instance. The method known as k-mer counting is employed to transform DNA sequences into a linguistic format during DNA sequencing analysis. To transform those sequences into languages, we employ this specific technique. We accomplished this by utilizing six-letter words, referred to as hexamers due to their length. The conclusion is that this sequence can be deconstructed into four hexameric words.



	sequence	class
0	ATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCA...	4
1	ATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAG...	4
2	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
3	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
4	ATGCAACAGCATTTTGAATTTGAATACCAGACCAAAGTGGATGGTG...	3

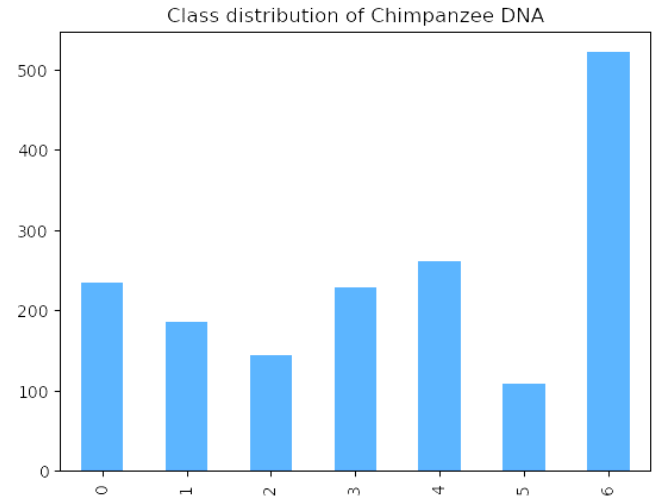
Fig 4.1 Distribution of Human DNA Classes

Received: August 03, 2025

International Journal of Applied Mathematics

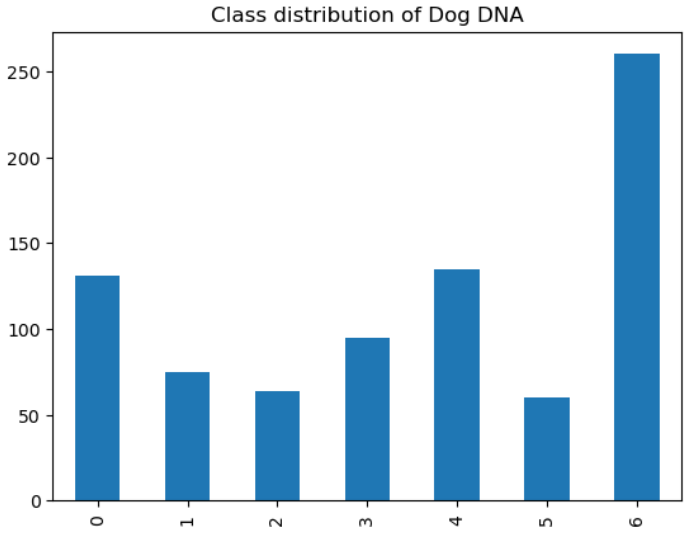
Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)



	sequence	class
0	ATGCCCCAACTAAATACCGCCGTATGACCCACCATAATTACCCCCA...	4
1	ATGAACGAAAATCTATTCGCTTCATTCGCTGCCCCCACAATCCTAG...	4
2	ATGGCCTCGCGCTGGTGGCGGTGGCGACGCGGCTGCTCCTGGAGGC...	4
3	ATGGCCTCGCGCTGGTGGCGGTGGCGACGCGGCTGCTCCTGGAGGC...	4
4	ATGGGCAGCGCCAGCCCGGGTCTGAGCAGCGTGTCCCCCAGCCACC...	6

Fig 4.2 Distribution of Chimpanzee DNA Classes



	sequence	class
0	ATGCCACAGCTAGATACATCCACCTGATTTATTATAATCTTTTCAA...	4
1	ATGAACGAAAATCTATTCGCTTCTTTCGCTGCCCCCTCAATAATAG...	4
2	ATGGAAACACCCTTCTACGGCGATGAGGCGCTGAGCGGCCTGGGCG...	6
3	ATGTGCACTAAAATGGAACAGCCCTTCTACCACGACGACTCATACG...	6
4	ATGAGCCGGCAGCTAAACAGAAGCCAGAACTGCTCCTTCAGTGACG...	0

Fig 4.3 Distribution of Dog DNA Classes

Received: August 03, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Currently, we can examine Figures 4.1, 4.2, and 4.3, which illustrate that all classes are nearly balanced. Nonetheless, other classes exhibit a relative balance, allowing for direct application, and enabling us to manage imbalanced datasets effectively. Certain data sets are deficient, while other categories are relatively balanced. In the event of an imbalanced dataset, one may consider employing oversampling techniques. Twenty percent of the overall test size will be designated for the Train test split. Subsequently, we conducted an accuracy assessment by implementing all of our classification algorithms on the three datasets at our disposal. A range of classification metrics is employed to facilitate the evaluation of these classification algorithms. All previously mentioned metrics are obtained from the confusion matrix. The model demonstrates exceptional performance regarding human data. The chimpanzee is similarly impacted by it. This is unsurprising, considering the established genetic relationship between humans and chimpanzees. Due to the lack of close genetic relation between humans and dogs, the model yielded unsatisfactory results when applied to canine data.

Confusion matrix for predictions on human test DNA sequence

Predicted	0	1	2	3	4	5	6
Actual							
0	99	0	0	0	1	0	2
1	0	104	0	0	0	0	2
2	0	0	78	0	0	0	0
3	0	0	0	124	0	0	1
4	1	0	0	0	143	0	5
5	0	0	0	0	0	51	0
6	1	0	0	1	0	0	263

accuracy = 0.984
precision = 0.984
recall = 0.984
f1 = 0.984

Confusion matrix for predictions on Chimpanzee test DNA sequence

Predicted	0	1	2	3	4	5	6
Actual							
0	232	0	0	0	0	0	2
1	0	184	0	0	0	0	1
2	0	0	144	0	0	0	0
3	0	0	0	227	0	0	1
4	2	0	0	0	254	0	5
5	0	0	0	0	0	109	0
6	0	0	0	0	0	0	521

accuracy = 0.993
precision = 0.994
recall = 0.993
f1 = 0.993

Confusion matrix for predictions on Dog test DNA sequence

Predicted	0	1	2	3	4	5	6
Actual							
0	127	0	0	0	0	0	4
1	0	63	0	0	1	0	11
2	0	0	49	0	1	0	14
3	1	0	0	81	2	0	11
4	4	0	0	1	126	0	4
5	4	0	0	0	1	53	2
6	0	0	0	0	0	0	260

accuracy = 0.926
precision = 0.934
recall = 0.926
f1 = 0.925

Received: August 03, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This paper proposes deep learning techniques to demonstrate the model's accuracy. The CNN method achieved an accuracy of 98.00 percent in machine learning, whereas the transfer learning algorithm attained an accuracy of 98.4 percent in deep learning. The accuracy achieved is derived from the classification of the seven distinct categories. We attained a chimpanzee accuracy of 99.3% and a dog detection accuracy of 92%. The previously presented experimental results unequivocally indicate that the proposed model is an efficacious method for classifying DNA sequences.

Table 3. Comparison of Species

Species	Accuracy	Precision	Recall	F1-Score
Human	98.45	98.45	98.45	98.45
Chimpanzee	99.3	99.4	99.3	99.3
Dog	92.6	83.4	92.6	92.5

5. **Conclusion**:

Deep barcoding is limited in its capacity to incorporate data from new species because it relies on known DNA sequences and a high-quality training set, which are limitations of supervised learning. This article delineates methods for classifying unidentified specimens into established species utilizing a trained model with DNA barcodes. This article delineates these methods. Notwithstanding progress in species classification via deep learning, significant challenges persist in its application to DNA barcode analysis. Researchers can enhance their study of DNA barcoding through techniques like deep barcoding, which progressively clarifies species identification using DNA barcodes. Forensic departments, medical analysis units, and archaeologists are typically the primary users of it.
The classification of DNA species is a potent instrument with significant applications across numerous domains. It is employed in the domains of forensics, medical analysis, and archaeology.

Forensics Department:

- **Identifying Unknown Samples:** When biological material is found at a crime scene, DNA species classification can determine if it's human or from another animal. This helps focus investigations and avoid wasting resources on irrelevant samples.
- **Wildlife Crimes:** Poaching and illegal trafficking of endangered species are serious problems. DNA analysis can identify the species from seized samples like meat, hides, or bones, helping to prosecute these crimes and conserve wildlife.
- **Food Fraud:** DNA species classification can detect mislabeling or adulteration of food products. For example, it can verify if expensive fish is being substituted with cheaper alternatives, protecting consumers and ensuring food safety.

Medical Analysis:

Received: August 03, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Identifying Pathogens:** Rapid and accurate identification of bacteria, viruses, or other pathogens is crucial for diagnosis and treatment of infectious diseases. DNA-based methods are faster and more precise than traditional techniques, leading to timely interventions and better patient outcomes.
- **Tracking Disease Outbreaks:** By analyzing the DNA of pathogens, scientists can trace the source and spread of outbreaks, helping to contain them and prevent future occurrences.
- **Monitoring Antibiotic Resistance:** DNA analysis can identify genes associated with antibiotic resistance, Assisting in monitoring the emergence and dissemination of drug-resistant strains and informing treatment strategies.

Archaeology:

- **Reconstructing Past Environments:** Analyzing DNA from ancient remains helps identify the species present in past ecosystems, providing insights into biodiversity, climate change, and human impact on the environment.
- **Understanding Human-Animal Interactions:** DNA analysis of animal remains from archaeological sites can reveal how humans interacted with animals in the past, including domestication, hunting practices, and the spread of livestock.
- **Tracing the Origins of Agriculture:** By analyzing the DNA of ancient crops and livestock, researchers can trace the origins and spread of agriculture, shedding light on human cultural development and technological advancements.

In conclusion, DNA species classification is a versatile technique with broad applications across various disciplines. It provides valuable information about the biological origin of samples, aiding in criminal investigations, disease diagnosis, and our understanding of the past.

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. 1690-701, 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?," PLOS Biology, vol. 9, no. 8, Art. no. e1001127, 2011.

[3]. 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. 313-21, 2003.

[4]. 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, Art. no. e312, 2004.

[5]. M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe, "Defining operational taxonomic units using DNA barcode data," Philos Trans R Soc Lond B Biol Sci, vol. 360, no. 1462, pp. 1935-43, 2005.

[6]. 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. 1043-6, 2013.

[7]. E. Weitschek, G. Fiscon, and G. Felici, "Supervised DNA Barcodes species classification: analysis, comparisons and results," BioData Min, vol. 7, no. 1, pp. 4, 2014.

Received: August 03, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[8]. A. Fiannaca, M. La Rosa, R. Rizzo, and A. Urso, "A k-mer based barcode DNA classification methodology based on spectral representation and a neural gas network," Artif Intell Med, vol. 64, no. 3, pp. 173-84, 2015.

[9]. 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, Art. no. 2, 2019.

[10]. G. Sohsah, A. R. Ibrahimzada, H. Ayaz, and A. Cakmak, "Scalable classification of organisms into a taxonomy using hierarchical supervised learners," bioRxiv, 2020.

[11]. D. S. Kermany, M. Goldbaum, and et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122-1131 e9, 2018.

[12]. C. H. Yang, S. H. Moi, O. Y. Fu, L. Y. Chuang, M. F. Hou, and Y. D. Lin, "Identifying risk stratification associated with a cancer for overall survival by deep learning-based CoxPH," IEEE Access, vol. 7, pp. 67708-67717, 2019.

[13]. A. Gharehbaghi, and M. Linden, "A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network," IEEE Trans Neural Netw Learn Syst, vol. 29, no. 9, pp. 4102 – 4115, 2017.

[14]. H. Zeng, and D. K. Gifford, "Predicting the impact of non coding variants on DNA methylation," Nucleic Acids Research, vol. 45, no. 11, pp. e99-e99, 2017.

[15]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-44, May 28, 2015.

[16]. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Computation, vol. 1, no. 4, pp. 541-551, 1989.

[17]. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[18]. C. P. Meyer, and G. Paulay, "DNA barcoding: error rates based on comprehensive sampling," PLoS Biology, vol. 3, no. 12, Art. no. e422, 2005.

[19]. 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.

[20]. 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.

Received: August 03, 2025