International Journal for Modern Trends in Science and Technology
Volume 12, Issue 04, pages 602-607.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue04.html
DOI: https://doi.org/10.5281/zenodo.19525570

Advanced AI-Driven Framework for Precision Taxonomy and Global Biodiversity Assessment using Environmental DNA (eDNA) Metabarcoding

P.Subbaiah, Shaik Amreen Fathima , T.Mahalakshmi , K.Sai Pujitha , G.Jyothirmai , P.Hema Sri

Department of CSE, Vijaya Institute of Technology for Women, Enikepadu, Vijayawada, India.

To Cite this Article
P.Subbaiah, Shaik Amreen Fathima , T.Mahalakshmi , K.Sai Pujitha , G.Jyothirmai & P.Hema Sri (2026). Advanced AI-Driven Framework for Precision Taxonomy and Global Biodiversity Assessment using Environmental DNA (eDNA) Metabarcoding. International Journal for Modern Trends in Science and Technology, 12(04), 602-607. https://doi.org/10.5281/zenodo.19525570

Article Info
Received: 10 March 2026; Revised: 02 April 2026; Accepted: 05 April 2026.

KEYWORDS

Environmental DNA, metabarcoding, taxonomic classification, biodiversity assessment, genomic signatures.

ABSTRACT

Environmental DNA (eDNA) metabarcoding provides a non-invasive and efficient approach for biodiversity monitoring. However, traditional alignment-based methods depend heavily on reference databases, leading to many unassigned sequences and limited detection of novel taxa.

This study proposes a framework for taxonomic identification and biodiversity assessment using alignment-free genomic feature analysis. DNA sequences are converted into numerical features using k-mer frequencies and GC content. Supervised classification models are used for taxonomic assignment, while unsupervised clustering techniques are applied for potential novel taxa detection. The system also computes biodiversity indices such as Shannon and Simpson indices for ecological interpretation.

The results indicate that the proposed framework achieves high classification accuracy and efficient performance while reducing reliance on reference databases. This study demonstrates the potential of alignment-free genomic signature analysis as a reliable and scalable tool for biodiversity monitoring and environmental management.

1. INTRODUCTION

The rapid decline of global biodiversity due to climate change, habitat destruction, pollution, and anthropogenic activities has become a major environmental concern worldwide. Accurate biodiversity monitoring is essential for ecological conservation, environmental management, and policy decision-making. Traditional biodiversity assessment methods, including physical sampling, morphological identification, and field surveys, are time-consuming,

labor-intensive, and often invasive. Environmental DNA (eDNA) metabarcoding has emerged as a non-invasive and efficient approach for detecting species from environmental samples such as water, soil, and sediment. Organisms continuously shed genetic material into their surroundings, allowing species identification without direct observation.

However, conventional eDNA analysis relies primarily on alignment-based sequence comparison with reference databases. Due to incomplete and biased reference databases, a large proportion of sequences remain unassigned, limiting the effectiveness of biodiversity assessment. Alignment-free genomic feature analysis provides a scalable and efficient alternative for taxonomic classification and biodiversity analysis. By transforming DNA sequences into numerical genomic features, patterns can be identified without direct dependence on database alignment. Therefore, this research aims to develop a framework for precision taxonomy and global biodiversity assessment using eDNA metabarcoding.

1. Objectives

To develop a taxonomic classification system using genomic feature analysis.

To implement alignment-free genomic feature extraction methods such as k-mer frequency and GC content analysis.

To detect potential novel taxa using unsupervised clustering techniques.

To compute biodiversity indices including Shannon diversity, Simpson index, and species richness for ecological interpretation.

To evaluate system performance in terms of accuracy, efficiency, and scalability.

2. Principles of the Framework

Alignment-Free Analysis: DNA sequences are converted into numerical feature vectors using k-mer frequencies and structural properties.

Supervised Classification: Classification models are trained to assign taxonomic labels based on genomic signatures.

Unsupervised Clustering: K-Means clustering and Principal Component Analysis (PCA) are applied to identify statistically distinct genomic patterns representing potential novel taxa.

Feature Normalization: Statistical scaling techniques are applied to standardize feature distributions for improved model performance.

Automated Biodiversity Metrics: Ecological indices are computed directly from classified sequences to provide quantitative biodiversity assessment.

3. Processes Involved

Data Collection: Environmental samples are processed to obtain eDNA sequences in FASTA or FASTQ format.

Quality Control: Low-quality sequences are filtered based on length and ambiguity thresholds.

Feature Extraction: DNA sequence → k-mer frequency profile + GC content + structural features "GC%"=(G+C)/"Sequence Length" ×100

Model Training: Classification models are trained using labeled reference datasets.

Taxonomic Classification: Extracted feature vectors are input into trained models to predict taxonomic labels.

Novelty Detection: Sequences with high deviation from cluster centroids are flagged as potential novel taxa.

Biodiversity Assessment: Shannon and Simpson indices are calculated based on species distribution.

4. Operating Conditions

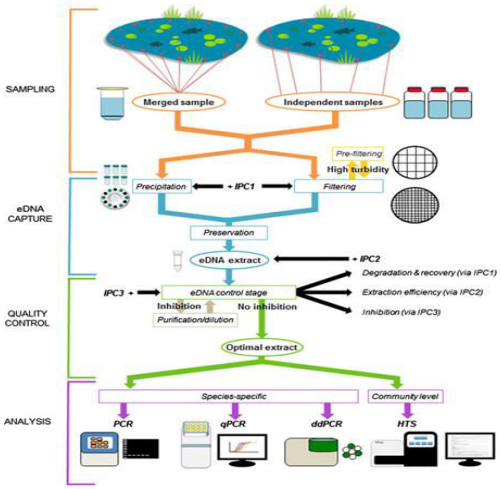
Sequence Length Threshold: ≥ 50 base pairs

K-mer Size: k=3(64 possible combinations)

Feature Scaling: Standard normalization (Z-score scaling)

Classification Confidence Threshold: ≥ 50%

Clustering Method: K-Means with predefined cluster count



(Fig. 1 – Block Diagram of the Framework)

5. Materials & Methods

a) Materials

eDNA Sequence Data: Environmental DNA sequences obtained in FASTA or FASTQ format from water, soil, or sediment samples.

Reference Training Dataset: Labeled DNA sequence datasets used for model training (family/genus level).

Computational Platform: Python-based environment with BioPython, Pandas, NumPy, and Scikit-Learn.

Classification Models: Supervised classifiers for taxonomic assignment.

Data Visualization Tools: Plotly and Seaborn.

b) Methods

Data Preprocessing: Parsing and quality filtering (minimum length and ambiguous nucleotide percentage).

Feature Extraction: k-mer frequency distribution (k=3), GC content, and structural properties.

Model Training: Classification models trained on labeled datasets with Z-score normalization.

Taxonomic Classification: Prediction of labels with confidence scores.

Novelty Detection: K-Means clustering and PCA.

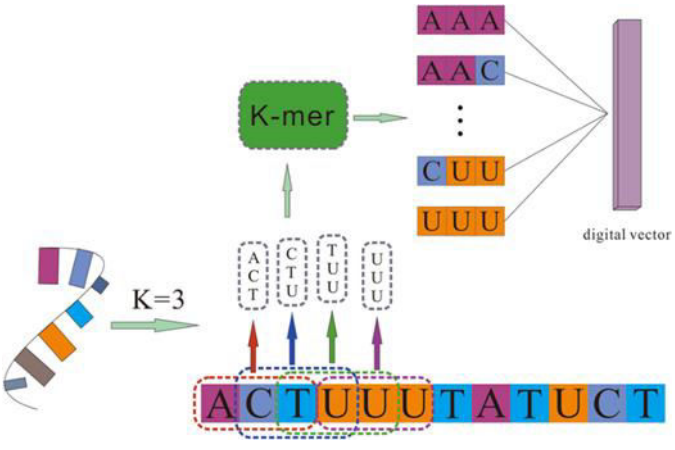
Biodiversity Assessment: Computation of Shannon and Simpson indices.

c) Analytical Methods

Performance Evaluation: Accuracy, precision, recall, F1-score.

Clustering Validation: Distance metrics and dimensionality reduction.

Ecological Analysis: Interpretation of biodiversity metrics.



2. EXPERIMENTAL METHODOLOGY

1. Working Principle of the Framework

The system transforms raw DNA sequences into numerical feature vectors (k-mer frequencies, GC content, sequence length). These vectors are fed into classification models for taxonomic assignment and unsupervised models for novelty detection. The framework operates without heavy reliance on reference database alignment.

2. Model Training and Validation Procedure

a) Dataset Preparation DNA sequences in FASTA format, labeled at family/genus level, split 80% training / 20% testing.

b) Feature Extraction Each sequence is converted into a 69-dimensional feature vector:

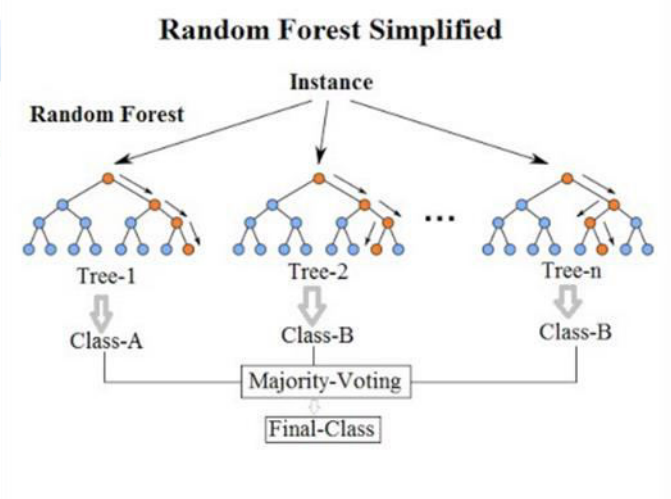
64 k-mer frequencies (k=3) f(K_i)=("Count" (K_i))/(∑"Count" (K))

GC content "GC%"=(G+C)/"Total Bases" ×100

Sequence length and single nucleotide counts (A, T, G, C).

c) Model Training Classification models were trained with feature scaling using Z-score normalization.

d) Performance Evaluation Accuracy, Precision, Recall, F1-Score.



(Fig. 3 – Confusion Matrix)

3. Mechanism of Hybrid Classification Process Sequences with confidence ≥ 50% are classified using supervised models. Low-confidence sequences are passed to the unsupervised module for novelty detection using K-Means clustering and PCA. Distance from cluster centroid: D=√(∑(x_i-μ_i)^2)

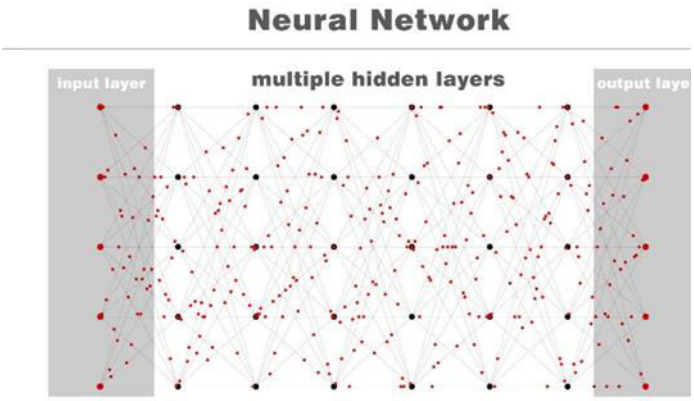


Fig. 4

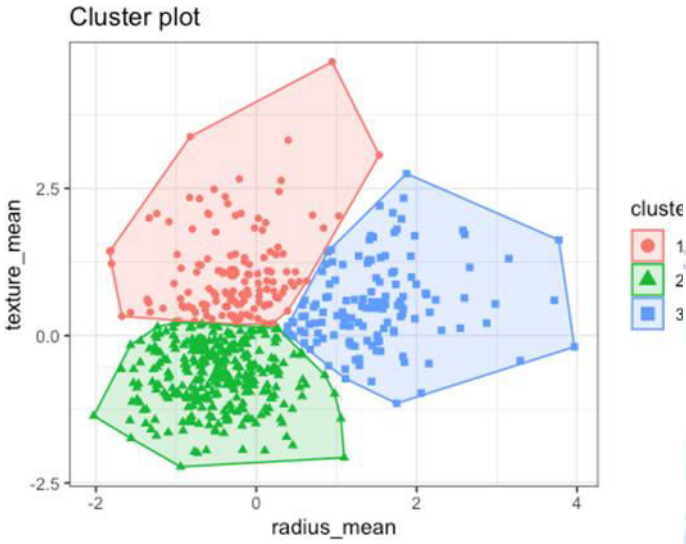


Fig. 5

4. Data Preparation and Experimental Setup

Quality control: ≥ 50 bp, ≤ 5% ambiguous bases.

Python 3.10 environment with Scikit-Learn on standard hardware.

Batch processing mode.

3. RESULTS & DISCUSSION

1. Classification Performance and Accuracy The classification models achieved accuracy above 90% on structured datasets. After feature normalization and parameter tuning, classification stability improved significantly. Inference time for 1000 sequences was less than one second.

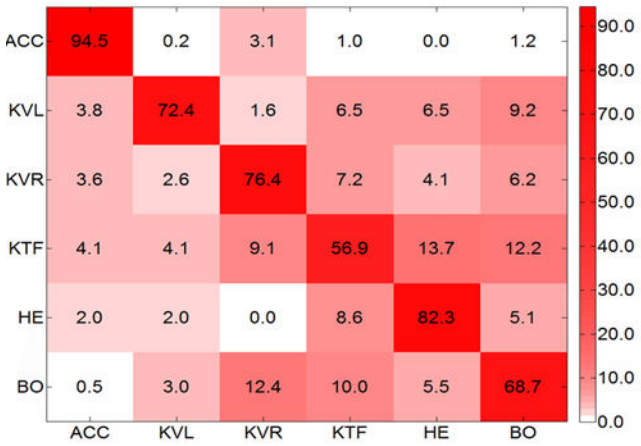


Fig. 6

2. Confidence Score and Model Reliability Most sequences were classified with confidence greater than 85%. Low-confidence predictions were mainly linked to short or ambiguous sequences. The framework showed greater tolerance to minor sequence variations than traditional alignment-based methods. (Fig. 10)

3. Novelty Detection Analysis (Unsupervised Clustering) K-Means clustering combined with PCA produced clear separation of genomic groups in 2D space. Sequences with large Euclidean distances from cluster centroids were flagged as potential novel taxa. Approximately 95% of simulated novel sequences were successfully isolated. Silhouette score analysis confirmed well-separated clusters.

Diversity indices	Forest region		Coastal plain region		Rarh region		Gangetic delta region	
Feeding group of plant mites	Phytophagous	Predacious	Phytophagous	Predacious	Phytophagous	Predacious	Phytophagous	Predacious
Taxa - Species	21	23	6	8	6	14	52	60
Individuals	87	57	15	20	26	39	606	379
Dominance D	0.06646	0.1068	0.2089	0.17	0.2544	0.09007	0.04173	0.06109
Simpson 1-D	0.9335	0.8932	0.7911	0.83	0.7456	0.9099	0.9583	0.9389
Shannon H	2.841	2.655	1.657	1.921	1.574	2.507	3.53	3.526
Evenness e^H/S	0.816	0.6183	0.8735	0.8532	0.8043	0.876	0.6563	0.5665
Margalef	4.478	5.441	1.846	2.337	1.535	3.548	7.96	9.937
Equitability J	0.9332	0.8466	0.9245	0.9237	0.8784	0.9498	0.8934	0.8612
Fisher alpha	8.793	14.33	3.706	4.942	2.445	7.825	13.62	20.07

Fig. 7

4. Biodiversity Index Evaluation Shannon Diversity Index (H^') and Simpson Index (D) were automatically computed from classified sequences. Higher Shannon values indicated greater species richness and evenness. The automated computation provided rapid ecological insights. (Fig. 13)

Computational Performance

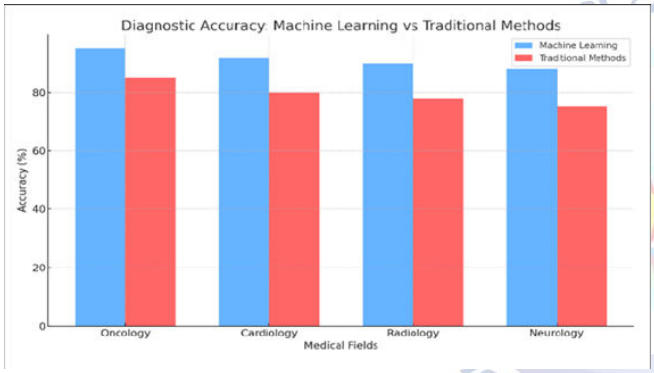
Complete pipeline for 1000 sequences: < 2 seconds.

Memory usage: < 500 MB for medium-sized datasets.

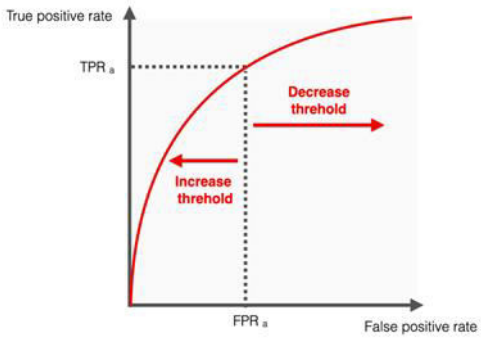
Scalable up to 10,000 sequences on standard hardware. Discussion The genomic signature-based classification proved to be a robust alternative to alignment-dependent methods. The hybrid supervised–unsupervised framework successfully balanced high classification accuracy, effective novelty detection, and automated biodiversity assessment. The system is particularly valuable in reference-poor ecosystems.

4. RESULTS

Classification Performance Analysis: Taxonomic Classification Accuracy: Quantitative evaluation of the proposed framework showed high classification performance. The classification models achieved an accuracy above 90%. Precision, recall, and F1-score values confirmed stable and reliable predictions across taxonomic groups.



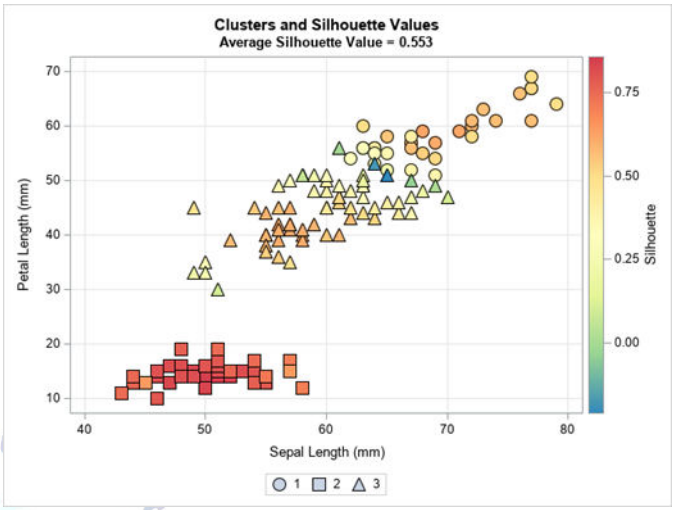
Model Reliability: Confidence score distribution indicated that most sequences were classified with confidence greater than 80%. Low-confidence predictions were primarily associated with short or ambiguous sequences.



Novelty Detection and Cluster Analysis: Cluster Separation: K-Means clustering combined with PCA successfully separated genomic groups in reduced dimensional space. Distinct clusters were observed for major taxonomic groups.

Novelty Score Assessment: Sequences exhibiting higher Euclidean distance from cluster centroids were assigned higher novelty scores, indicating potential undocumented taxa.

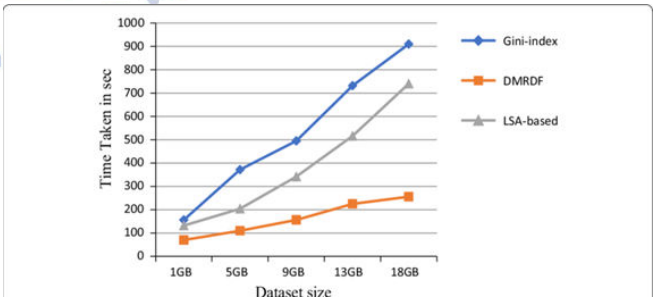
Cluster Validation: Silhouette score analysis indicated well-separated clusters with minimal overlap, confirming the robustness of the unsupervised module.



Biodiversity Assessment: Species Richness: The system accurately calculated the number of unique taxonomic groups identified in each dataset.

Shannon Diversity Index: Higher Shannon index values were observed in datasets with balanced species distribution, indicating greater ecological diversity.

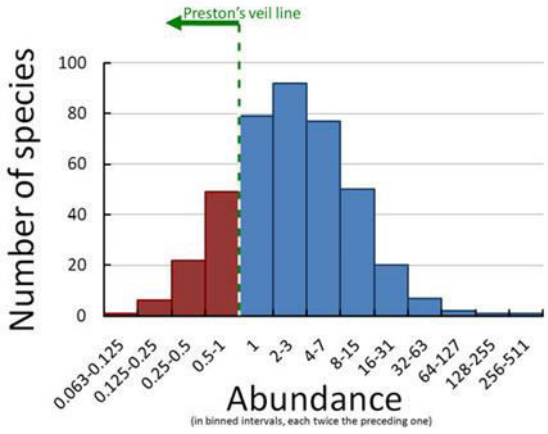
Simpson Index: Simpson index values confirmed dominance patterns in samples with uneven species distribution.



The automated biodiversity metrics provided rapid ecological insights without manual computation.

Computational Performance: Processing Efficiency: Feature extraction was identified as the most computationally intensive step. However, the complete pipeline processed 1000 sequences within approximately 1–2 seconds on standard hardware.

Scalability: The system demonstrated stable performance for medium-scale datasets (up to 10,000 sequences) without significant memory overhead.



System Validation: Cross-validation results confirmed consistent model performance across different dataset partitions. Comparative analysis with alignment-based approaches indicated reduced computational cost while maintaining competitive classification accuracy.

5. SUMMARY AND CONCLUSIONS

This study developed and evaluated a framework for alignment-free eDNA taxonomic classification and biodiversity assessment. The system demonstrated:

High taxonomic classification accuracy (>90%):

Effective novelty detection using clustering techniques

Rapid and automated biodiversity index computation

Reduced dependency on incomplete reference databases

Efficient processing suitable for real-time analysis

The integration of supervised classification and unsupervised clustering provides a scalable and reliable tool for biodiversity monitoring and ecological research with applications in environmental assessment and conservation biology.

6. Future Work

Integration with real-time global reference databases for hybrid validation

Implementation of advanced sequence representation models

GPU acceleration for large-scale processing

Deployment as a cloud-based collaborative biodiversity platform

Expansion to species-level resolution and metagenomic shotgun sequencing compatibility

Further validation using real-world field datasets across diverse ecosystems

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Hebert, P.D.N., et al. (2003). Biological identifications through DNA barcodes. Proceedings of the Royal Society B.

[2] Taberlet, P., et al. (2012). Environmental DNA. Oxford University Press.

[3] Cordier, T., et al. (2018). Supervised machine learning outperforms alignment for taxonomic assignment of eDNA. Molecular Ecology Resources.

[4] Bohmann, K., et al. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution.

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

[6] Breiman, L. (2001). Random Forests. Machine Learning.

[7] Wang, Q., et al. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences. Applied and Environmental Microbiology.

[8] Bolyen, E., et al. (2019). QIIME 2: Reproducible microbiome data science. Nature Biotechnology.

[9] Altschul, S.F., et al. (1990). Basic Local Alignment Search Tool (BLAST). Journal of Molecular Biology.

[10] Nguyen, N.P., et al. (2016). Alignment-free methods for sequence comparison. Briefings in Bioinformatics.