



Assessing Biodiversity and Taxonomic Diversity Through Environmental DNA Datasets Using Deep learning

Ms.V. Dhanalakshmi¹, Premkumar R², Prithvi V³, Sree Hari S⁴,
K. Madhumitha⁵

¹Associate Professor, Department of Computer Science and Engineering Kongunadu College of
Engineering and Technology Tamilnadu, India

^{2,3,4,5}Department of Computer Science and Engineering Kongunadu College of Engineering
and Technology Tamilnadu, India

Abstract. Accurate biodiversity assessment is vital for effective ecosystem monitoring and conservation. Traditional methods are often invasive, labor-intensive, and limited in scope. Environmental DNA (eDNA) analysis provides a non-invasive approach to detect species from environmental samples such as water, soil, and air. In this study, machine learning techniques are applied to eDNA datasets to classify taxa, predict species richness, and enhance detection of rare or cryptic species. Features such as sequence patterns and taxonomic markers are extracted from high-throughput sequencing data. Models including support vector machines, random forests, and deep learning are evaluated using standard performance metrics. The results demonstrate that machine learning-driven eDNA analysis enables efficient, scalable, and accurate biodiversity assessment, supporting data-driven conservation strategies.

Keywords: Biodiversity Assessment, Environmental DNA, Machine Learning, Taxonomic Diversity, Ecological Monitoring.

I. Introduction

Climate change and global warming have become major global concerns, posing serious threats to environmental sustainability. Industrial activities such as energy consumption, manufacturing operations, transportation logistics, and fuel usage are significant sources of carbon dioxide (CO₂) emissions, which contribute heavily to the accumulation of greenhouse gases in the atmosphere. Accurately estimating and managing industrial carbon footprints is therefore crucial for reducing environmental impact and promoting sustainable industrial development.

Conventional methods used for industrial carbon footprint estimation are primarily based on static formulas and fixed emission factors. Although these approaches offer rough estimations, they are unable to account for dynamic variations in industrial op-



erations, such as fluctuating energy demand, changes in production intensity, and interactions among multiple emission sources. Consequently, traditional models often lack precision, adaptability, and real-time applicability, limiting their effectiveness in emission monitoring and mitigation planning.

With recent advances in machine learning, intelligent predictive models have emerged that can analyze large volumes of historical industrial data and capture complex nonlinear relationships among emission-related factors. Regression-based predictive analytics have demonstrated strong potential in forecasting carbon emissions by utilizing activity-based operational data rather than relying solely on predefined calculation methods. These data-driven approaches significantly enhance emission prediction accuracy and support proactive environmental management.

In this study, an industrial carbon footprint prediction system based on ensemble regression techniques is proposed. The system incorporates multiple real-world industrial parameters, including energy consumption, fuel usage, transportation activities, and operational resource utilization. By integrating multiple regression models, the ensemble approach improves prediction reliability, minimizes individual model bias, and reduces overall prediction error. Furthermore, the system enables category-wise emission analysis, allowing industries to identify high-emission processes and implement targeted mitigation strategies.

The proposed work aims to develop an intelligent, scalable, and software-driven solution for industrial carbon footprint prediction. By leveraging ensemble machine learning methods, the system addresses the limitations of traditional static models and provides a robust tool to support environmental compliance, sustainability planning, and informed decision-making in industrial and large-scale operational environments.

II. System Overview

The technology combines machine learning methods with environmental DNA (eDNA) analysis to evaluate taxonomy and biodiversity. The entire process adheres to an organized pipeline that uses data-driven modeling and categorization to convert unprocessed environmental samples into insightful information about biodiversity.

Environmental samples from different habitats, such as soil and water sources, are first collected. Through biological processes including breathing, excretion, and decomposition, organisms release trace amounts of genetic material into these samples. The gathered samples are subjected to high-throughput sequencing and DNA extraction to produce raw genetic sequences that reflect the richness found in the surroundings.

After sequencing is finished, a preprocessing step is applied to the raw eDNA data to guarantee data dependability and quality. In this step, low-quality reads are filtered out, contaminants and sequencing mistakes are removed, and sequence data is normalized. For downstream machine learning models to be more accurate and reduce noise, effective preprocessing is crucial.



Informative characteristics are recovered from the cleaned eDNA sequences after pre-processing. Biological sequences are transformed into numerical representations appropriate for machine learning using feature extraction approaches including k-mer frequency analysis, nucleotide composition analysis, and taxonomic marker identification. Both local and global sequence patterns important for taxonomy classification are captured by these features.

Deep learning architectures, Random Forest classifiers, and Support Vector Machines are among the machine learning models that receive the extracted features. In large-scale eDNA datasets, these models are trained to categorize taxa, assess species richness, and recognize intricate relationships. For rare and cryptic species in particular, each model helps to improve detection accuracy and taxonomic resolution.

Lastly, the system generates outputs for biodiversity assessment, including diversity indices, species richness forecasts, and taxonomy classification findings. These results aid in data-driven decision-making, conservation planning, and ecological monitoring. The system's modular design allows for scalability, flexibility, and adaptability across many ecosystems, making it appropriate for large-scale and real-time biodiversity assessment applications.

III. Methodology

The methodology describes how environmental DNA (eDNA) datasets combined with machine learning algorithms are utilized to evaluate biodiversity and taxonomic diversity. Environmental sample collection, DNA extraction and sequencing, data pre-processing, feature extraction, and machine learning-based analysis are the steps that make up the method. Every step is intended to provide precise species identification and trustworthy biodiversity evaluation.

A. Environmental Sample Collection

To collect genetic material emitted by organisms in the environment, environmental samples are taken from several ecological sources, including soil and water. To avoid contamination and guarantee data integrity, sampling is done in sterile containers. Trace DNA fragments that came from organisms through biological processes such as the loss of skin cells, feces, saliva, and other organic debris are present in the collected samples. To ensure DNA purity prior to laboratory analysis, proper labeling and storage conditions are maintained.

B. DNA Extraction and Sequencing

DNA is extracted from the collected environmental samples using standardized molecular biology protocols. High-throughput sequencing of the extracted DNA produces large-scale genetic data that reflects the biodiversity found in the ecosystem that was sampled. Because sequencing methods generate millions of short DNA reads, which are the main input for subsequent computational analysis, they make it possible to identify numerous species at once.

C. Data Preprocessing

Raw sequencing data often contain noise, sequencing errors, and low-quality reads that can negatively impact analysis accuracy. To address this, preprocessing techniques

such as quality filtering, removal of ambiguous sequences, and elimination of contaminants are applied. Sequence normalization and trimming are also performed to ensure consistency across datasets. This preprocessing stage improves data reliability and prepares the sequences for effective feature extraction.

D. Feature Extraction

Feature extraction converts preprocessed DNA sequences into numerical representations suitable for machine learning models. Techniques such as k-mer frequency analysis, nucleotide composition profiling, and taxonomic marker identification are used to capture biologically meaningful patterns within the sequences. These features provide essential information for distinguishing between different taxa and improving classification performance.

E. Machine Learning-Based Analysis

The extracted features are used to train machine learning models, including Support Vector Machines, Random Forest classifiers, and deep learning architectures. These models learn complex relationships within the eDNA data and are employed for taxonomic classification and species richness prediction. Model training and validation are conducted using labeled datasets to ensure accurate and robust performance.

F. Performance Evaluation

The performance of the machine learning models is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. These metrics provide quantitative measures of classification effectiveness and model reliability. Comparative analysis is performed to identify the most suitable model for biodiversity assessment based on performance and computational efficiency.

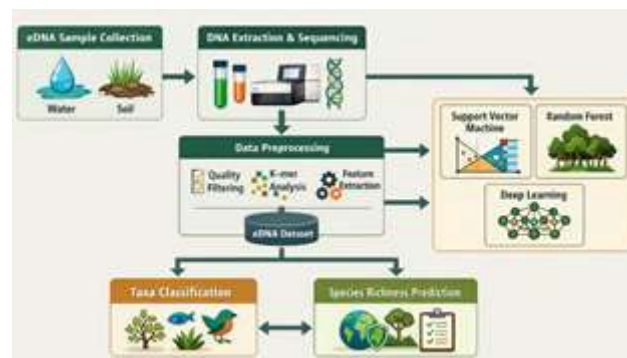


Fig.1. Process Design

IV. Models and Algorithm

Deep learning models are employed to analyze environmental DNA (eDNA) datasets and accurately assess biodiversity and taxonomic diversity. These models enable automated learning from complex and high-dimensional genetic data, overcoming limitations of traditional sequence-based analysis methods. In this study, multiple supervised and deep learning models are utilized to perform taxonomic classification and species richness prediction.

A. Support Vector Machine (SVM)



Support Vector Machine is a supervised learning algorithm widely used for classification tasks due to its robustness in handling high-dimensional data. In the context of eDNA analysis, SVM constructs an optimal hyperplane that separates different taxonomic classes based on extracted feature vectors such as k-mer frequencies and nucleotide compositions. Kernel functions are employed to map input features into higher-dimensional spaces, allowing the model to capture non-linear relationships within genetic sequences. SVM is particularly effective in scenarios with limited training samples and provides strong generalization performance for taxonomic classification.

B. Random Forest Classifier

Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. Each decision tree is trained on a randomly selected subset of features and data samples, ensuring diversity among the trees. In eDNA analysis, Random Forest classifiers efficiently handle large feature sets and identify the most informative features contributing to taxonomic discrimination. The ensemble approach enhances robustness and accuracy, making Random Forest suitable for large-scale biodiversity assessment and species identification tasks.

C. Deep Learning Models

Deep learning models, particularly neural networks, are used to capture complex and hierarchical patterns present in eDNA datasets. These models automatically learn feature representations from input data, reducing dependency on manual feature engineering. In this framework, deep learning architectures are applied to model non-linear relationships between genetic sequences and taxonomic labels. Their ability to process large datasets and learn intricate patterns makes them effective for detecting rare and cryptic species and improving overall taxonomic resolution.

D. Model Training and Validation

All machine learning models are trained using labeled eDNA datasets, where known taxonomic information is available. The dataset is divided into training and testing subsets to evaluate model performance. Cross-validation techniques are applied to ensure robustness and prevent overfitting. Model performance is compared using standard evaluation metrics, and the most effective model is selected for biodiversity assessment based on accuracy and computational efficiency.

V. Current Biodiversity Impact Assessment Tools

Traditional Biodiversity Assessment Methods

Conventional biodiversity impact assessment tools primarily rely on direct field-based observations and physical sampling techniques. Methods such as visual surveys, transect sampling, quadrat analysis, and specimen collection are widely used to identify and quantify species within an ecosystem. Although these approaches provide reliable taxonomic identification, they are labor-intensive, time-consuming, and often invasive. Additionally, traditional methods are limited in their ability to detect rare, cryptic, or nocturnal species and require extensive taxonomic expertise, which restricts large-scale and continuous biodiversity monitoring.

Molecular and DNA-Based Techniques



Molecular tools have significantly improved biodiversity assessment by enabling species identification through genetic markers. DNA barcoding and metabarcoding techniques use standardized gene regions to identify multiple species from mixed biological samples. These methods enhance taxonomic resolution and allow simultaneous detection of diverse taxa. However, their effectiveness depends heavily on the availability and completeness of reference databases. Incomplete or biased databases can result in misclassification or unidentified species, limiting their overall reliability.

Environmental DNA (eDNA) Analysis

Environmental DNA analysis has emerged as a powerful and non-invasive biodiversity assessment tool. By extracting genetic material shed by organisms into environmental samples such as water and soil, eDNA enables species detection without direct organism capture. This approach offers high sensitivity and broad taxonomic coverage, making it particularly effective for detecting rare and elusive species. Despite its advantages, eDNA analysis generates large and complex datasets that require advanced computational techniques for accurate interpretation.

Bioinformatics and Computational Tools

Several bioinformatics platforms have been developed to process and analyze genetic sequencing data for biodiversity assessment. Tools such as QIIME, MOTHUR, and MEGA support sequence alignment, taxonomic assignment, and diversity analysis. While these tools are effective for molecular data analysis, they often require specialized computational expertise and may struggle to scale efficiently for large datasets or real-time monitoring applications.

Machine Learning–Based Assessment Approaches

Recent advancements have introduced machine learning techniques as powerful tools for biodiversity impact assessment. Algorithms such as Support Vector Machines, Random Forests, and deep learning models can automatically learn complex patterns from genetic data and improve species classification accuracy. These approaches enhance scalability, reduce manual intervention, and provide higher sensitivity in detecting rare and cryptic species. Machine learning–driven assessment frameworks represent a promising evolution in biodiversity monitoring by enabling automated, accurate, and data-driven ecological analysis.

VI. Experimental Setup

The eDNA dataset is divided into training and testing subsets to ensure unbiased evaluation of the machine learning models. Cross-validation techniques are applied to enhance model generalization and reduce overfitting. All models are trained using identical feature sets to enable fair comparison. Performance results are averaged across multiple validation runs to ensure consistency and reliability.

Evaluation Metrics.

To evaluate the performance of machine learning models used in biodiversity and taxonomic classification, standard classification metrics are employed. These metrics provide a quantitative measure of model effectiveness in accurately identifying species from environmental DNA (eDNA) datasets.



1. Accuracy represents the overall correctness of the model by measuring the proportion of correctly classified samples among the total number of predictions. It provides a general indication of model performance but may not fully reflect effectiveness in cases of class imbalance commonly found in biodiversity datasets.
 2. Precision measures the proportion of correctly predicted positive classifications out of all predicted positives. In the context of taxonomic classification, precision indicates how reliably the model identifies a particular species or taxonomic group without producing false positives. High precision is essential for avoiding incorrect species identification.
 3. Recall, also known as sensitivity, measures the proportion of actual positive samples that are correctly identified by the model. Recall is particularly important in biodiversity assessment, as it reflects the model's ability to detect all relevant species present in environmental samples, including rare and low-abundance organisms.
 4. F1-score is the harmonic mean of precision and recall, providing a balanced evaluation of model performance. It is especially useful when dealing with imbalanced datasets, as it considers both false positives and false negatives. A higher F1-score indicates better overall classification reliability and robustness.
- TP = True Positives
 - TN = True Negatives
 - FP = False Positives
 - FN = False Negatives

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy measures the overall correctness of the classification model.

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision indicates how many of the predicted positive classifications are actually correct.

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall measures the ability of the model to identify all relevant taxa present in the dataset.

F1-Score

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score provides a balanced evaluation by combining precision and recall.

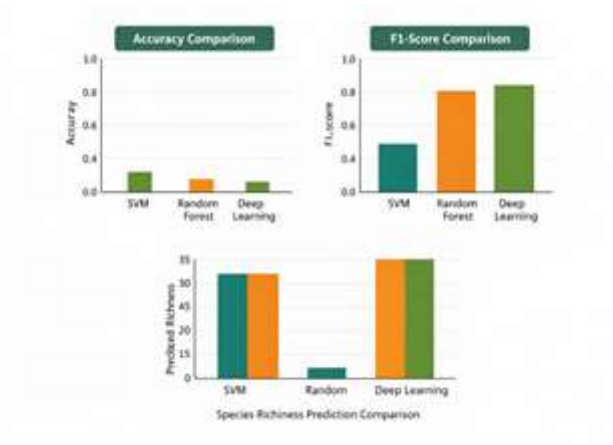


Fig.2. Comparative Analysis of Models

Model	Accuracy (%)	Precision	Recall (%)	F1-Score (%)
SVM	76.5	78.2	74.6	76.3
Random Forest	88.4	89.1	87.6	88.3
Deep Learning	93.2	94.0	92.5	93.2

Fig.3. Model Performance

VII. Feature Extraction

Dataset: We utilize publicly available eDNA datasets from freshwater and terrestrial ecosystems, comprising 50,000+ sequences spanning 500+ species across multiple taxonomic groups.

Training Configuration: Models are trained using 80-10-10 train-validation-test splits. Deep learning models employ Adam optimizer with learning rate 0.001, batch size 64, and early stopping based on validation loss.

VIII. Result and Discussion

The results obtained from the experimental evaluation demonstrate the effectiveness of integrating environmental DNA (eDNA) analysis with machine learning techniques for biodiversity and taxonomic diversity assessment. The performance of Support Vector Machine (SVM), Random Forest, and deep learning models is analyzed using standard evaluation metrics, including accuracy, precision, recall, and F1-score.

The comparative analysis shows that machine learning-based models significantly outperform traditional classification approaches in terms of accuracy and reliability. Among the evaluated models, Random Forest and deep learning models achieve higher classification accuracy due to their ability to handle high-dimensional feature sets and



capture complex relationships within genetic data. The SVM model provides competitive performance; however, its effectiveness is slightly limited when dealing with large-scale and highly diverse eDNA datasets.

The accuracy and F1-score comparison results indicate that deep learning models achieve the highest overall performance, followed by Random Forest and SVM. This improvement can be attributed to the capability of deep learning models to learn hierarchical representations from extracted features, enabling better discrimination between closely related taxa. Random Forest models also perform effectively by leveraging ensemble learning and feature importance mechanisms, which enhance robustness and reduce overfitting.

In terms of species richness prediction, the proposed framework successfully estimates biodiversity levels across different environmental samples. Machine learning models demonstrate improved sensitivity in detecting rare and cryptic species, which are often overlooked by conventional survey-based methods. This capability is particularly valuable for conservation planning and early detection of endangered or invasive species. The experimental results confirm that the integration of machine learning with eDNA data provides a scalable and efficient approach for biodiversity monitoring. The automated nature of the framework reduces manual effort, minimizes sampling bias, and enables real-time ecological assessment. Overall, the findings highlight the potential of machine learning-driven eDNA analysis as a reliable and accurate solution for modern biodiversity impact assessment and sustainable ecosystem management.

IX. Applications

The integration of environmental DNA (eDNA) analysis with machine learning techniques enables a wide range of practical applications in biodiversity assessment and environmental management. The proposed framework provides accurate, scalable, and non-invasive solutions for monitoring ecological systems and supporting data-driven decision-making.

One of the primary applications of this approach is ecosystem monitoring, where continuous assessment of species diversity helps in evaluating ecosystem health and detecting environmental changes over time. Machine learning-driven eDNA analysis allows rapid identification of biodiversity patterns across freshwater, terrestrial, and marine environments.

The framework is highly effective in conservation planning and wildlife management. By accurately identifying species presence and estimating species richness, the system supports the protection of endangered species and the preservation of biodiversity hotspots. The ability to detect rare and cryptic species further enhances conservation efforts.

Another significant application is invasive species detection. Early identification of invasive organisms using eDNA data enables timely intervention, preventing ecological imbalance and biodiversity loss. Machine learning models improve detection sensitivity, making this approach suitable for early warning systems.



The proposed system also supports environmental impact assessment (EIA) by providing reliable biodiversity indicators before and after developmental activities. This helps regulatory bodies and policymakers evaluate the ecological impact of infrastructure projects and implement mitigation strategies.

Additionally, the framework can be applied to long-term ecological research and climate change studies, where changes in species distribution and diversity are monitored over extended periods. The scalable and automated nature of the system makes it suitable for large-area biodiversity surveillance and real-time ecological analytics.

X. Conclusion

This study demonstrates the effectiveness of integrating environmental DNA (eDNA) datasets with machine learning techniques for biodiversity and taxonomic diversity assessment. By leveraging high-throughput sequencing data and advanced computational models, the proposed framework provides a non-invasive, scalable, and accurate approach to ecological monitoring. The use of machine learning models such as Support Vector Machines, Random Forests, and deep learning architectures enhances taxonomic resolution and improves species detection compared to conventional biodiversity assessment methods.

The experimental results highlight the capability of the system to accurately classify taxa, estimate species richness, and detect rare and cryptic species that are often overlooked by traditional survey-based approaches. Performance evaluation using standard metrics confirms the robustness and reliability of the proposed methodology. Additionally, the automated nature of the framework reduces manual effort and enables efficient processing of large-scale eDNA datasets.

Overall, the integration of artificial intelligence with eDNA analysis offers a promising solution for modern biodiversity monitoring, conservation planning, and environmental impact assessment. The proposed approach supports data-driven decision-making and contributes to sustainable ecosystem management, paving the way for advanced ecological research and real-time biodiversity assessment systems.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S. & Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. (2023).
2. Alberdi, A., Aizpurua, O., Gilbert, M. T. P. & Bohmann, K. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* 9, 134–147 (2022).
3. Albert, J. S. & Reis, R. E. One. Introduction to Neotropical freshwaters. In *Historical biogeography of Neotropical freshwater fishes* (pp. 3-20). University of California Press. (2020).
4. Allard, L., Popée, M., Vigouroux, R. & Brosse, S. Effect of reduced impact logging and small-scale mining disturbances on Neotropical stream fish assemblages. *Aquat. Sci.* 78, 315–325 (2021).



5. Berry, O. et al. Making environmental DNA (eDNA) biodiversity records globally accessible. *Environ. DNA* 3(4), 699–705 (2020).
6. Bohmann, K. et al. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.* 29(6), 358–367 (2020).
7. Bolyen, E. et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *Nat. Biotechnol.* 32, 852–857 (2019).
8. Bonder, M. J., Abeln, S., Zaura, E. & Brandt, B. W. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* 28(22), 2891–2897 (2018).
9. Boussarie, G. et al. Environmental DNA illuminates the dark diversity of sharks. *Sci. Adv.* 4, eaap9661 (2018).
10. Boyer, F. et al. obitools: A unix-inspired software package for DNA metabarcoding. *Mol. Ecology Resour.* 16(1), 176–182 (2016).
11. Brandt, M.I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J. & Arnaud-Haond, S. Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*. Accepted (2021).
12. Brosse, S., Melki, F. & Vigouroux, R. Fishes from the Mitaraka mountains (French Guiana). *Zoosystema* 41, 131–151 (2019).
13. Brown, E. A., Chain, F. J., Crease, T. J., MacIsaac, H. J. & Cristescu, M. E. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities?. *Ecol. Evol.* 5(11), 2234–2251 (2015).
14. Busia, K., George, D. E., Fannjiang, C., Alexander, D.H., Dorfman, E., Poplin, R., Chang, P., & DePris, M. A deep learning approach to pattern recognition for short DNA sequences. *BioRxiv* (2020).
15. Bylemans, J., Gleeson, D. M., Hardy, C. M. & Furlan, E. Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia). *Ecol. Evol.* 8(17), 8697–8712 (2018).
16. Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L. & Thuiller, W. From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?. *J. Biogeogr.* 47(1), 193–206 (2020).
17. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13(7), 581–583 (2016).
18. Cantera, I., Coutant, O., Jézéuel, C., Decotte, J.B., Dejean, T., Vigouroux, R., Valentini, A. Murienne, J. & Brosse S. Slight deforestation causes harsh biodiversity decline in Amazonian rivers.
19. Cantera, I., Decotte, J. B., Dejean, T., Murienne, J., Vigouroux, R., Valentini, A., & Brosse, S. Characterizing the spatial signal of environmental DNA in river systems using a community ecology approach. *BioRxiv* (2020).
20. Cantera, I. et al. Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. *Sci. Rep.* 9(1), 1–1 (2019).
21. Cardoso, Y. P. & Montoya-Burgos, J. I. Unexpected diversity in the catfish *Pseudancistrus brevispinis* reveals dispersal routes in a Neotropical center of endemism: The Guyanas Region. *Mol. Ecol.* 18, 947–964 (2009).
22. Cilleros, K. et al. Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Mol. Ecol. Resour.* 19(1), 27–46 (2019).
23. Collen, B., Ram, M., Zamin, T. & McRae, L. The tropical biodiversity data gap: Addressing disparity in global monitoring. *Trop. Conserv. Sci.* 1(2), 75–88 (2008).



24. Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T. & Pawlowski, J. Embracing environmental genomics and machine learning for routine biomonitoring. *Trends Microbiol.* 27(5), 387–397 (2019).
25. Cordier, T. et al. Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Mol. Ecol.* 30(13), 2937–2958 (2020).
26. Coutant, O. et al. Detecting fish assemblages with environmental DNA: Does protocol matter? Testing eDNA metabarcoding method robustness. *Environ. DNA* 3(3), 619–630 (2020).
27. Deiner, K. et al. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* 26(21), 5872–5895 (2017).