

Research on an Environmental Category Inference Model for Microbiome Data Based on Neural Networks

Fang'annan Zhu

College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

f_a_n_zhu@hust.edu.cn

Abstract. This paper proposes a neural network–based computational model that infers the environmental category of a sample from microbiome data by analyzing the composition and abundance features of microbial communities. We adopt a two-layer fully connected neural network architecture and perform data standardization, feature engineering, and model training on microbial abundance data. On a multi-class environmental dataset, the model achieves strong performance with a micro-averaged AUC of 0.9675 and a macro-averaged AUC of 0.9701. The experimental results indicate that the model effectively captures distinctive community patterns across environments, providing a quantitative analytical tool for environmental microbiology, niche analysis, and environmental monitoring. Biological interpretation of the model outputs further reveals potential associations between specific microbial genera and environmental categories, laying the groundwork for deeper investigation into the mechanisms of environmental adaptation in microbial communities.

Keywords: microbiome, neural networks, environmental inference, machine learning, community analysis

1. Introduction

Microbial communities, as one of the most abundant biological resources on Earth, are widely distributed across diverse environments, and their composition and structure are closely linked to the conditions of their habitats [1]. Distinct physicochemical properties of environments such as freshwater wetlands, host digestive systems, and biofilms shape unique microbial community characteristics, while the metabolic activities of microbes, in turn, influence environmental material cycling and ecological functions [2]. Therefore, inferring the environmental origin of samples from microbiome data is not only a key topic in microbial ecology but also of great significance for environmental monitoring, ecological conservation, and human health.

Traditional environmental microbiology has relied heavily on cultivation techniques and physiological–biochemical analyses, which are limited in their ability to comprehensively reveal the complex associations between microbial communities and environments. The advent of high-throughput sequencing has generated massive amounts of microbiome data, making it possible to explore microbe–environment relationships through computational approaches. In recent years,

machine learning methods have been widely applied in microbiome data analysis—for example, random forest–based community classification and support vector machine–based environmental factor association analysis. However, these methods still face limitations when dealing with the high dimensionality and nonlinearity of microbiome data [3].

Neural networks, as powerful nonlinear modeling tools, can automatically learn higher-order features from data and have achieved groundbreaking results in fields such as image recognition and natural language processing. Yet, applying neural networks to environmental inference from microbiome data remains in its early stages, with challenges including difficulties in data standardization, high feature dimensionality, and limited biological interpretability.

This study aims to develop a data-driven neural network model to address the core problem of “how to infer the environmental category of a sample from microbiome data.” By implementing rigorous data preprocessing, optimizing the network architecture, and conducting systematic model evaluation, we achieve accurate identification of multiple environmental categories. Furthermore, biological interpretation of the model outputs reveals intrinsic associations between microbial communities and their environments. The results not only provide a new computational approach for microbiome analysis but also offer fresh perspectives for understanding the mechanisms of environmental adaptation in microbial communities.

2. Related work

The study of the environmental specificity of microbial communities has long been a central topic in microbial ecology. Early research compared microbial compositions across different environmental samples and found significant associations between specific microbial groups and environmental factors. For example, in soil environments, the abundance of Actinobacteria is positively correlated with organic matter content, whereas in aquatic environments, the distribution of Cyanobacteria is influenced by light and temperature [2]. These findings provide a biological foundation for environmental inference based on microbial composition.

With the advancement of high-throughput sequencing technologies, multivariate statistical methods have become mainstream in community analysis. Approaches such as principal component analysis (PCA) and non-metric multidimensional scaling (NMDS) are widely used to visualize patterns of microbial community differentiation across environments. However, these methods cannot directly achieve quantitative prediction of environmental categories [3]. The introduction of machine learning has driven predictive analysis of microbiome data. For instance, Random Forest can identify environment-associated indicator species through importance scoring, achieving an accuracy of 85% in distinguishing between soil and aquatic environments [4]. In the medical domain, Yang et al. employed univariate and multivariate Cox regression to identify microbial features associated with the overall survival of endometrial cancer patients, constructing a prognostic model based on nine microbial genera. This demonstrated the potential of microbiome features in disease prognosis prediction [5].

Horizontal gene transfer (HGT), as an important mechanism of microbial environmental adaptation, mediates the spread of functional genes and may influence the environmental specificity of microbial communities. Studies have shown that HGT occurs more frequently in low spatial entropy environments (such as biofilms), leading to greater similarity in functional traits among microbial communities. This provides a functional-level explanation for environmental inference models [1]. At the same time, quorum sensing (QS) systems regulate microbial community behaviors through signaling molecules, and differences in QS networks across environments may further reinforce the environmental specificity of microbial communities [2].

In recent years, deep learning has been increasingly applied in microbiome research. Convolutional neural networks (CNNs) have been used to process 16S rRNA gene sequence data by extracting sequence features to predict environmental types, while recurrent neural networks (RNNs) have shown advantages in analyzing temporal microbiome data. However, most of these studies have focused on binary classification tasks within single environmental types, lacking systematic validation for multi-class environmental inference.

Building on existing research, this study addresses the challenge of multi-class environmental inference by developing a fully connected neural network model tailored to microbial abundance data. Through rigorous feature engineering and model optimization, we achieve improved predictive accuracy. Moreover, by integrating microbial ecological theory into the interpretation of model results, this work compensates for the shortcomings of previous studies in terms of multi-class adaptability and biological interpretability.

3. Materials and methods

3.1. Data sources

The microbiome data used in this study were obtained from publicly available environmental sample databases, covering 11 environmental categories: biofilm, bog, cecum, coral reef, fecal, nasal cavity, pharynx, rumen, saliva, sediment, and tongue dorsum. The dataset was divided into a training set (SourceCM) and a test set (QueryCM), containing 1,101 and 221 samples respectively. The microbial composition of each sample was represented by the absolute abundance of OTUs (operational taxonomic units). The environmental label data (SourceLabel and QueryLabel) recorded detailed environmental classification information for each sample.

3.2. Data preprocessing

Data preprocessing aimed to transform the raw microbial abundance data into a feature matrix suitable for model training. The steps included: (1) Matrix transposition: converting the sample–OTU matrix into an OTU–sample matrix to facilitate subsequent aggregation by taxonomic levels. (2) Zero-value removal: eliminating OTUs with zero abundance across all samples to reduce redundant features. (3) Taxonomic aggregation: splitting OTUs according to taxonomic information (kingdom–phylum–class–order–family–genus) and selecting the genus level as the analytical unit. The abundances of OTUs belonging to the same genus were summed to reduce feature dimensionality. (4) Feature alignment: since the OTU composition may differ between training and test sets, alignment was achieved by retaining the features present in the training set and filling missing features in the test set with zeros, ensuring consistency in input dimensions. (5) Relative abundance calculation: converting genus-level abundances into relative abundances (the proportion of a genus within the total abundance of a sample) to eliminate the effect of sequencing depth differences. (6) Standardization: applying StandardScaler to standardize the relative abundance data (mean = 0, standard deviation = 1), thereby avoiding the influence of scale differences among features during model training.

3.3. Label processing

The environmental label data were organized hierarchically. To simplify the classification task, the terminal level of the hierarchy was extracted as the target label. The labels were converted from text

to integer codes using LabelEncoder, generating a mapping between environmental categories and encoded values, ultimately forming an 11-class multi-classification task.

3.4. Model construction

This study adopted a two-layer fully connected neural network architecture, with the following structure: (1) Input layer: the number of nodes equaled the feature dimension (i.e., the number of genera), receiving the standardized microbial abundance data. (2) Hidden layer: 100 nodes with ReLU activation to enhance the model's nonlinear fitting capability. (3) Output layer: the number of nodes equaled the number of environmental categories (11), using the Softmax activation function to output the prediction probabilities for each category.

The model training parameters were set as follows: (1) Loss function: cross-entropy loss (CrossEntropyLoss), suitable for multi-class classification tasks. (2) Optimizer: Adam optimizer with an initial learning rate of 0.001. (3) Learning rate scheduler: StepLR strategy, reducing the learning rate by half every 5 epochs. (4) Batch size: 32. (5) Training epochs: 25.

3.5. Model evaluation metrics

The performance of the model was evaluated using ROC curves and AUC (area under the curve) values, including: (1) Per-class AUC: computing the ROC curve and AUC for each environmental category to reflect the model's recognition ability for specific classes. (2) Micro-averaged AUC: treating all categories as a whole, calculating global false positive rate (FPR) and true positive rate (TPR). This metric is particularly suitable for imbalanced multi-class tasks. (3) Macro-averaged AUC: calculating the arithmetic mean of the AUC values across all categories to reflect the model's overall classification performance.

4. Experimental results

4.1. Model training process

During the 25 training epochs, the model's loss gradually decreased and eventually stabilized. In the initial epochs (1–5), the loss dropped rapidly from 1.4323 to 0.0886, indicating that the model quickly learned the basic associations between microbial abundances and environmental categories. In the subsequent epochs (6–25), the loss decreased slowly to 0.0282, suggesting that the model continued to optimize detailed features and ultimately reached a well-converged state.

4.2. Model performance evaluation

4.2.1. ROC curves and AUC values

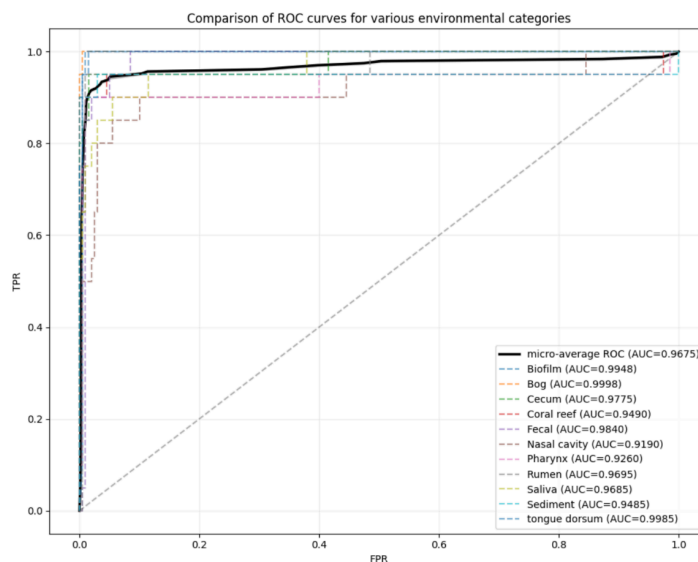


Figure 1. Comparison of ROC curves for various environmental categories

ROC curve analysis on the test set (Figure 1) shows that the micro-averaged ROC curve (black solid line) is close to the upper-left corner, with an AUC of 0.9675, demonstrating excellent overall classification performance. Among the category-specific ROC curves (colored dashed lines), the highest AUC was observed for bog (0.9998), followed by tongue dorsum (0.9985) and biofilm (0.9948), indicating that the model is highly effective in recognizing these environments. The relatively lower values were obtained for nasal cavity (0.9190) and coral reef (0.9490), though these still represent strong performance levels.

4.2.2. Multi-class AUC statistics

The micro-averaged AUC reached 0.9675, while the macro-averaged AUC was 0.9701. Both values are close to 1, confirming the stability and accuracy of the model in the multi-class environmental inference task. The slightly higher macro-averaged AUC indicates that the model performed well on minority classes and was not significantly affected by differences in sample sizes across categories.

4.3. Analysis of environmental category recognition difficulty

Differences in AUC values across environmental categories reflect varying levels of recognition difficulty. For bog environments, the AUC was the highest, suggesting that these natural wetland samples possess distinctive anaerobic microbial features—such as methanogenic archaea—that were effectively captured by the model, making them the easiest to classify [2]. For nasal cavity samples, classification was more challenging, possibly because their microbial communities overlap with those of other host-associated environments, such as the pharynx and saliva [4]. The coral reef category showed relatively lower AUC values, which may be attributed to the complexity of coral reef ecosystems. Coral reefs consist of multiple microhabitats (e.g., coral surfaces, surrounding

water, sediments), leading to large inter-sample variability in microbial composition and increasing the difficulty for the model to learn discriminative features [1].

5. Discussion

5.1. Biological interpretation of model performance

The neural network model constructed in this study demonstrated superior performance in predicting environmental categories from microbiome data, largely because it captured the intrinsic associations between microbial communities and their habitats. From the perspective of microbial ecology, physicochemical factors such as temperature, pH, and nutrient availability act as filters, selecting microbial taxa with specific adaptive traits and thereby shaping distinctive community profiles [2]. For example, the anaerobic and organic matter-rich conditions of bog environments promote the enrichment of methanogens and sulfate-reducing bacteria, whose abundance patterns became key features for the model to identify bog samples.

Horizontal gene transfer (HGT), as an important mechanism of microbial adaptation, may further strengthen the environmental specificity of microbial communities. Previous studies have shown that in environments with low spatial heterogeneity, such as biofilms, HGT occurs at higher frequencies, enabling rapid dissemination of functional genes within the community and resulting in more homogeneous functional characteristics [1]. The high recognition accuracy for the Biofilm category (AUC 0.9948) observed in this study may be linked to this mechanism—tight microbial aggregation within biofilms facilitates genetic exchange, making their community features easier to distinguish from those of other environments.

Quorum sensing (QS)-mediated microbial communication may also contribute to model performance. QS signaling networks vary across environments and regulate collective microbial behaviors, including biofilm formation and antibiotic production [3]. For instance, the high abundance of DSF signals in sediment environments may shape unique community structures, enabling the model to identify Sediment samples (AUC 0.9485) based on the abundance patterns of associated taxa [2].

5.2. Limitations and future improvements

Despite the promising results, several limitations remain: (1) Subjectivity in feature selection: Aggregating features at the genus level may obscure fine-scale differences at the species level. Future work could integrate species-level annotations or functional gene data to enhance model resolution. (2) Sample distribution bias: Some environmental categories, such as coral reef, had relatively small sample sizes, which may have affected the model's ability to learn their representative features. Approaches such as data augmentation or active learning could help mitigate this issue. (3) Limited interpretability: The “black-box” nature of neural networks restricts detailed understanding of microbiome–environment relationships. Incorporating attention mechanisms could highlight taxa that contribute most strongly to predictions, thereby improving interpretability.

5.3. Application prospects

The model proposed in this study provides a new tool for environmental microbiology, with potential applications including: (1) Environmental monitoring: Rapid inference of habitat types for unknown samples based on microbial community data, supporting ecological surveys and environmental assessments. (2) Microbial resource discovery: Identification of signature taxa in

specific environments, providing leads for the discovery of industrially relevant enzymes, antibiotics, or other bioresources. (3) Ecological restoration: Designing targeted strategies to optimize microbial community structures, based on model-revealed microbiome–environment associations, to improve degraded ecosystems. (4) Medical applications: Extending this framework to human microbiome research, where environment-like microbial signatures could contribute to disease diagnosis and prognosis prediction.

6. Conclusion

This study developed a neural network–based model for predicting environmental categories from microbiome data. Through rigorous preprocessing and model optimization, the framework achieved high classification accuracy across 11 environmental categories (micro-averaged AUC 0.9675; macro-averaged AUC 0.9701). The results demonstrate that the model effectively captured distinctive microbial patterns across environments, with performance differences interpretable through ecological concepts such as HGT and QS systems. The findings not only provide a new computational approach for microbiome analysis but also offer empirical support for understanding microbial adaptation to diverse environments. Looking forward, integrating multi-omics data and enhancing model interpretability are expected to further increase the biological relevance and application potential of this framework.

References

- [1] Wenzhi Xue; Juken Hong; Runmeng Zhao; Huaxiong Yao; Yi Zhang; Zhuojun Dai; Teng Wang. Spatial entropy drives the maintenance and dissemination of transferable plasmids [J]. *Molecular Systems Biology*, 2025, 21(7): 856-869.
- [2] Ying Jin; Wenkang Chen; Jie Hu; Jinfeng Wang; Hongqiang Ren. Constructions of quorum sensing signaling network for activated sludge microbial community [J]. *ISME Communications*, 2024, 4(1): ycae018.
- [3] Kalinga Pavan T Silva; James Q Boedicker. A neural network model predicts community-level signaling states in a diverse microbial community [J]. *PLoS Computational Biology*, 2019, 15(6): e1007166.
- [4] Hong Juken; Xue Wenzhi; Wang Teng. Emergence of alternative stable states in microbial communities undergoing horizontal gene transfer [J]. *eLife*, 2025, 13: RP99593.
- [5] Yuting Yang; Yuchen Meng; Ziyang Xu; Qin Zhang; Miaomiao Li; Fanbing Kong; Suping Zhang; Xinling Li; Yihua Zhu. Leveraging microbiome signatures to predict tumor immune microenvironment and prognosis of patients with endometrial carcinoma [J]. *Discover Oncology*, 2025, 16(1): 299.