sustainability

MDPI

Review

Unraveling the Microbiome–Environmental Change Nexus to Contribute to a More Sustainable World: A Comprehensive Review of Artificial Intelligence Approaches

Maria Inês Barbosa [1], Gabriel Silva [1,2], Pedro Ribeiro [1], Eduarda Vieira [2], André Perrotta [3], Patrícia Moreira [2] and Pedro Miguel Rodrigues [1,*]

[1] CBQF—Centro de Biotecnologia e Química Fina, Escola Superior de Biotecnologia, Universidade Católica Portuguesa, 4169-005 Porto, Portugal; mibarbosa@ucp.pt (M.I.B.); gsilva@ucp.pt (G.S.); s-pmsbribeiro@ucp.pt (P.R.)
[2] CITAR—Centro de Investigação em Ciência e Tecnologia das Artes, Escola das Artes, Universidade Católica Portuguesa, 4169-005 Porto, Portugal; evieira@ucp.pt (E.V.); prmoreira@ucp.pt (P.M.)
[3] Centre for Informatics and Systems of the University of Coimbra (CISUC), 3030-290 Coimbra, Portugal; avperrotta@dei.uc.pt
* Correspondence: pmrodrigues@ucp.pt

Abstract

This review aims to explore the literature to assess the potential of artificial intelligence (AI) in environmental monitoring for predicting microbiome dynamics. Recognizing the significance of comprehending microorganism diversity, composition, and ecologically sustainable impact, the review emphasizes the importance of studying how microbiomes respond to environmental changes to better grasp ecosystem dynamics. This bibliographic search examines how AI (Machine Learning and Deep Learning) approaches are employed to predict changes in microbial diversity and community composition in response to environmental and climate variables, as well as how shifts in the microbiome can, in turn, influence the environment. Our research identified a final sample of 50 papers that highlighted a prevailing concern for aquatic and terrestrial environments, particularly regarding soil health, productivity, and water contamination, and the use of specific microbial markers for detection rather than shotgun metagenomics. The integration of AI in environmental microbiome monitoring directly supports key sustainability goals through optimized resource management, enhanced bioremediation approaches, and early detection of ecosystem disturbances. This study investigates the challenges associated with interpreting the outputs of these algorithms and emphasizes the need for a deeper understanding of microbial physiology and ecological contexts. The study highlights the advantages and disadvantages of different AI methods for predicting environmental microbiomes through a critical review of relevant research publications. Furthermore, it outlines future directions, including exploring uncharted territories and enhancing model interpretability.

Keywords: microbiome; machine learning; deep learning; environment; forecasting; sustainability goals

Academic Editor: Chenggang Gu

Received: 2 July 2025
Revised: 26 July 2025
Accepted: 7 August 2025
Published: 9 August 2025

Citation: Barbosa, M.I.; Silva, G.; Ribeiro, P.; Vieira, E.; Perrotta, A.; Moreira, P.; Rodrigues, P.M. Unraveling the Microbiome–Environmental Change Nexus to Contribute to a More Sustainable World: A Comprehensive Review of Artificial Intelligence Approaches. *Sustainability* **2025**, *17*, 7209. https://doi.org/10.3390/su17167209

1. Introduction

Microorganisms have a remarkable impact on ecosystems and influence various biological processes. The microbiome, which encompasses microorganisms and their genetic material within a specific environment, has garnered significant attention in recent

years. Unraveling the diversity and composition of microbial communities is crucial for understanding their ecological functions and responses to environmental changes [1,2].

Understanding microbiomes is essential for advancing sustainability because these play a critical role in building strong ecosystems and developing green technologies. By exploring the diversity and functions of microbes, we can unlock solutions to urgent environmental challenges [3]. Microorganisms are the foundation of healthy soil, driving important processes like nutrient cycling. A lively soil microbiome can reduce the need for synthetic fertilizers and enhance crop resilience to drought, pests, and diseases, which in turn can contribute to more stable food production [4]. Furthermore, a healthy soil microbiome improves soil structure and promotes carbon sequestration, which helps mitigate climate change. Beyond agriculture, microbes have unique metabolic capabilities that make them ideal for environmental cleanup through bioremediation. They are also essential to the circular economy by transforming organic waste into valuable resources that provide sustainable alternatives to fossil-fuel-based products [5,6]. Healthy and diverse microbial communities in soils, oceans, air, and even within larger organisms contribute significantly to ecosystem resilience. They play a direct role in global carbon and nitrogen cycles, and research into these functions is crucial for predicting how ecosystems will respond to climate change [7,8].

With the development of New Generation Sequencing (NGS) technologies like shotgun metagenomics and amplicon sequencing, the landscape of microbiome research has been revolutionized [9]. These techniques allow us to comprehensively analyze microbial communities, unraveling their taxonomic composition and functional potential. Capitalizing on these breakthroughs, researchers have started to use the power of Machine learning (ML) and Deep Learning (DL) algorithms to forecast and analyze how environmental factors influence microbial communities [10,11].

ML and DL algorithms have emerged as powerful tools for analyzing complex and high-dimensional biological datasets, such as those encountered. ML focuses on developing algorithms that enable computers to learn patterns from data and make predictions or decisions without being explicitly programmed. DL is a specialized subset of ML that uses multi-layered neural networks to automatically extract complex features from large datasets [10,12]. Supervised algorithms, such as random forests (RFs), support vector machines (SVMs), and neural networks (NNs) (e.g., Classical Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Autoencoders), leverage labeled datasets to predict specific microbial traits or community profiles based on environmental variables [13]. In contrast, unsupervised algorithms like principal component analysis (PCA), clustering (e.g., k-means, self-organized maps) and network-based approaches enable us to uncover hidden patterns and associations within complex microbiome datasets [14]. These techniques promise to unravel the intricate relationships between environmental conditions and microbial dynamics [11].

However, one of the primary challenges in microbiome forecasting lies in interpreting the results generated by these algorithms. While ML algorithms effectively identify associations and predict microbial patterns, understanding the underlying biological mechanisms that drive these changes poses a significant challenge. Deciphering the causal links between environmental conditions and microbial responses necessitates a deeper understanding of microbial physiology, interactions, and the specific ecological context of the studied ecosystem. Overcoming this challenge will be vital to unlock the full potential of microbiome forecasting in environmental monitoring and management [15].

In this comprehensive review, our primary objective is to assess the current development status of state-of-the-art artificial intelligence (AI) algorithms in this domain and to identify the most effective predictive models for environmental microbiome forecasting. We

examine the rapidly evolving intersection of microbiome science and AI for environmental monitoring and forecasting, and critically analyze how a range of AI approaches—from classical ML to DL algorithms—are transforming our ability to predict changes in microbial diversity and community composition in response to environmental fluctuations. An exhaustive examination of peer-reviewed publications up to May 2025 was conducted, prioritizing studies that report quantitative performance metrics such as accuracy, correlation measures, and error rates, with the aim of addressing the following research questions:

- What is the current state of the art in AI-driven prediction of environmental microbiome dynamics?
- How do various ML/DL algorithms compare in their ability to capture complex microbiome–environment interactions?
- To what extent can microbiome changes be accurately predicted using environmental parameters across different ecosystems?
- How might AI microbiome predictions contribute to a more sustainable world?
- What are the key methodological challenges and technical limitations affecting prediction accuracy?
- How effectively do current models capture the bidirectional relationship between microbiome alterations and environmental changes?

By synthesizing the current literature, we evaluate the efficacy of different AI architectures in modeling both how environmental factors influence microbial communities and how microbiome alterations subsequently impact environmental processes, capturing the bidirectional relationship critical to understanding the microbiome–environmental change nexus. We also identify breakthrough methodologies, significant knowledge gaps, and persistent technical challenges in this interdisciplinary field and outline promising future directions for AI-driven microbiome research.

2. Materials and Methods

Document Search

The data for this review was obtained from accessible literature in the Elsevier, Frontiers Media, Oxford University Press, Wiley, ASM Journals, PLOS ONE, Copernicus GmbH, Springer Nature, MDPI, ACS, and Biomedical Central databases. It was based on a document search utilizing Google Scholar, Scopus, and Web of Science as the electronic search engines.

The articles were selected if they focused on the microbiome from the environment, such as soil, water, or air samples. They used AI algorithms to predict the microbiome based on the environmental conditions under which the samples were collected and presented some performance measures for their prediction.

The document search was performed between January 2025 and May 2025. The search was performed with a combination of the following keywords: “microbiome”, “machine learning”, “environment”, “prediction”, “diversity”, “composition”, “abundance”, “metagenomics”, “ecology”, “neural networks”, “artificial intelligence”, “deep learning”, “supervised learning”, “sequencing”, “accuracy”, “performance”.

To narrow the scope, we focused exclusively on studies involving predictive modeling. As a result, correlation-based studies that do not include predictive analysis were excluded, as they are beyond the objectives of this work. Besides that, our selection criteria prioritized research that provided quantitative performance metrics, including accuracy, correlation measures (e.g., R^2), and error rates, and only peer-reviewed, published studies were included. This enabled us to evaluate the efficacy of different AI architectures in modeling both how environmental factors influence microbial communities and how microbiome alterations subsequently impact environmental processes, capturing the bidirectional relationship critical to understanding the microbiome–environmental change nexus.

3. Results

3.1. Environmental Impacts on Microbiome

Climate change induced by human activity has increasingly become a critical concern due to its widespread impacts on ecosystems and biodiversity, with evidence suggesting it contributes to habitat loss, species declines, and in some cases, extinctions. Although extensively established, biodiversity loss is an issue that requires additional research to understand its full effect on ecosystems and how this may affect environmental and even human health [16–21].

Despite awareness of biodiversity loss in diverse ecosystems, environmental research rarely highlights how changes in meteorological parameters and pollutant emissions affect the microbiomes surrounding us. These microbial communities' richness and inherent diversity are critical to maintaining the global ecosystem's health [22]. Since these microorganisms are less noticeable to our eyes, the influence of human activities and their consequences, like air pollution, on these microbiomes has received less attention and research. Changes in the biodiversity activity of the microorganisms that surround us, on the other hand, will surely disrupt the delicate balance of our ecosystems, harming not only the earth's environment [23–25].

Microbiomes have a wide range of effects on their habitats and are often involved in critical processes such as carbon and nutrient cycles. The lack of research on how microbiomes adapt to human-induced changes in environmental circumstances challenges the understanding and relevance of their connection to changes in meteorological parameters and air pollution, particularly in urban areas [26].

The impact of air pollution levels and meteorological conditions on the diversity and dispersion of environmental microbiomes is a new area of research that requires additional study. Regarding climatic conditions, relative humidity [27–29] and air temperature [30,31] have reported the highest influence on outdoor microbial communities. For instance, high humidity levels have been associated with an increase in the communities' pathogenicity [31–33]. As for temperature, a positive correlation was observed between it and the diversity of bacterial communities in southwest Greenland [34,35]. Regarding air pollution, research shows that the relative abundance of total and pathogenic bacteria correlates positively with particle and pollutant concentrations such as carbon monoxide and ozone [31,36].

3.2. High Throughput Sequencing

In microbial ecology, ribosomal RNA gene sequencing has been used to assess microbial diversity in diverse environments, including soil and aquatic ecosystems, hydrothermal vents, and the built environment. Due to the development of high-throughput sequencing (HTS), it is now possible to generate thousands of sequences from hundreds of samples in a single sequencing cycle [37,38].

The increasing availability of HTS technology has made it possible to investigate microbiological diversity in a significantly shorter time than in the past. Natural microbial communities have the potential to provide valuable insights into environmental phenomena and may aid in the prediction of environmental events. Thus, environmental reactions can be evaluated by comparing metagenomes from various environments, such as their participation and response to climate change and environmental stress, their sensitivity to contaminants and biodegradation capabilities, and the health risks they pose [39,40].

Genetic markers specific to each type of microorganism, also known as DNA barcoding, can be employed to identify the species in the environment. DNA identifiers are brief gene sequences unique enough to identify across species. DNA barcoding identifies species by sequencing a short, standardized DNA sequence in a well-defined gene [41,42].

The 16S rRNA gene is highly conserved across all bacterial species and can be used as a marker for distinguishing species [43]. The sequences of 16S rRNA genes from ambient samples have revolutionized our understanding of microbial diversity, especially bacteria, and aid in documenting the microbiomes' diversity. Regarding fungi, the difficulty is more remarkable because many cannot be grown in the laboratory. The strongest candidates for fungi appear to be the internal transcribed spacer (ITS) due to its higher success rate of polymerase chain reaction (PCR) amplification, which makes a substantial difference in its utility as a barcode [44].

Even though they do not provide information on changes in the abundance of individual taxa, higher-level characteristics, such as diversity measures, are frequently used to characterize the microbiome to gain access to a more significant shift or variation in microbial composition. While alpha diversity measures the microbiome diversity of a single sample, beta diversity measures the similarity or dissimilarity between two communities. The Shannon and Simpson Indices are two alpha diversity indices that each represent a distinct component of the variability of a community. Bray–Curtis and Unifrac dissimilarities are extensively used to investigate the relationship between environmental variables and microbial composition across habitats regarding beta diversity [45,46]. Both diversities take into account two aspects of a community: the number of unique organisms in a sample and the range of abundances for each [47].

Contrary to these methods, shotgun metagenomics sequencing enables researchers to collect all genes from all species in a complicated sample [48]. Microbiologists can also use the approach to assess bacterial diversity and identify the number of microorganisms in varied situations. Shotgun metagenomics can be used to research unculturable bacteria that would otherwise be difficult or impossible to study [49,50]. NGS enables researchers to sequence hundreds of organisms in parallel, unlike PCR-based techniques [51,52]. NGS-based metagenomic sequencing may discover shallow abundance members of the microbial community that may be overlooked or are too costly to identify using other approaches since it can combine several samples in a single sequencing run and produce high sequence coverage per sample [53,54].

3.3. Machine and Deep Learning Applications

ML and DL are transforming microbial ecology by enabling researchers to identify, classify, and predict outcomes from complex microbiome data. This field leverages not only traditional ML approaches but also more advanced DL techniques, each with distinct capabilities for analyzing microbial communities.

ML and DL generically employ two main fundamental approaches: supervised and unsupervised learning. Supervised Machine Learning (SML) creates predictive models from labeled data [55]. In microbiome research, SML uses frequency count matrices from NGS instruments (representing microbial species across samples) to establish relationships between microbiome profiles and system traits. These algorithms build generic hypotheses from known patterns to predict future outcomes. To put it another way, SML aims to create a precise model of how class labels are distributed depending on predictive attributes [56–58]. The resulting classifier may then assign class labels to test situations where the predictor feature values are known, but the class label value is unknown. This method may be used to predict numerical continuous outcomes, such as concentration or age, or categorical outputs that can be binary, such as sick and healthy, or multiple, such as disease stages [11,59,60]. Unsupervised Machine Learning (USML) algorithms, on the other hand, analyze unlabeled data to discover inherent structures without predefined response labels. The samples used in these algorithms are not associated with any predefined response label [55]. Instead, the model seeks the potential data structure, grouping

it as best it can [56]. These methods may be particularly valuable in microbiome research as they can uncover complex hidden patterns within large NGS datasets that traditional methods might miss [14]. Within traditional ML, several classical models have proven effective for microbiome analysis. RF trees are a standard tool in microbiome research and have been extensively used to address a variety of challenges [61]. RF constructs multiple decision trees, segmenting samples progressively based on microbial taxonomic abundance. These forests are regulated by two principles: bootstrapping, which involves repeatedly picking random sample subsets with replacement, and node splitting criteria, which use information from a randomly selected feature subset to determine the division of each node in every tree. The ideal split is found using either a node impurity estimate for a classifier, which measures the probability of misclassification of additional samples, or the prediction squared error for a regressor [62]. SVM [63] finds optimal separation boundaries (hyperplanes) that maximize distance between different classes in labeled datasets, excelling with high-dimensional microbiome data where features exceed sample size. ANNs use hierarchical architectures mimicking brain dynamics, with interconnected neurons arranged in layers. Despite being "black boxes" with limited interpretability, ANNs excel at identifying complex datasets, making them an attractive technique for the role of microbes in complex settings [64,65]. Besides these, models such as Extra Trees Classifier (ETC), Logistic Regression (LoR), eXtreme Gradient Boosting (XGB), Decision Trees (DTs), K-Nearest Neighbors (KNN), Logistic Ridge Regression (Ridge), Logistic Lasso Regression (Lasso), Support Vector Classification (SVC), Support Vector Machine Linear (SVML), Support Vector Machine Radial Basis Function (SVMRBF), Ordinary Linear Regression (OLR), Linear Regression with Lasso Regularization (LRLR), Linear Regression With Ridge Regularization (LRRR), Support Vector Regression With Linear Kernel (SVRL), Support Vector Regression with Radial Basis Function (SVRRBF), Random Forest Regression (RFR), Adaboost Regression (ABR), Gradient Boost Regression (GBR), Linear Regression (LR), Quantile Regression Forest (QRF), Extremely Randomized Trees, Elastic Net Regression (ENR), Support Vector Regression (SVR), Aggregated Boosted Tree (ABT), Lasso Regression, Multiple Linear Regression (MLR), Light Gradient Boosting Machine (LGBM), Adaptive Boosting (AdaBoost), and Naive Bayes (NB) [66] have also been applied in this field.

DL algorithms employ specialized multilayered neural networks with complex architectures designed for specific tasks. These networks simulate brain functionality while reducing the need for extensive data preprocessing that traditional ML requires. By automatically extracting features from unstructured data, DL minimizes reliance on human experts since the algorithms can analyze unstructured data such as images and text by processes of automatic feature extraction [64,65]. LSTM neural networks have feedback connections, making them recurrent neural networks. The "Long Short-Term Memory" component suggests that this model may recall or forget information for long or short periods. Microbial communities evolve throughout time, resulting in a series of data [67]. LSTM is inherently suited to dealing with this data, where the temporal sequence is essential. CNNs are a DL tool used to analyze visual data effectively. CNNs are wildly successful for tasks requiring spatial connections, such as image and video processing, since they are intended to automatically learn spatial hierarchies of features from input data adaptively [68]. Autoencoders represent another powerful class of neural networks that learn in an unsupervised manner to encode data into a compressed representation and then reconstruct the original input from this encoding. These networks consist of an encoder that maps the input to a lower-dimensional latent space and a decoder that attempts to recreate the input from this compressed representation [69]. Autoencoders are excellent tools for dimensionality reduction, feature extraction, and anomaly detection, making them particularly valuable for analyzing complex microbiome datasets where iden-

tifying underlying patterns and removing noise are crucial [69]. By learning the intrinsic structure of microbiome data without requiring labeled samples, autoencoders facilitate exploratory analysis and can reveal hidden relationships within microbial communities [69]. In addition, DL architectures such as Recurrent Neural Networks (RNNs), Multi-Layer Perceptrons (MLPs), Residual Neural Networks (ResNets), Deep Neural Networks (DNNs), Deep Count Autoencoder (DCA), and DeepToA [70] have been explored for similar tasks.

The integration of both ML and DL approaches provides complementary tools for understanding microbial communities across diverse ecosystems. While traditional ML offers interpretability and efficiency with structured data, DL excels at extracting complex patterns from unstructured or temporal data without extensive feature engineering. As sequencing technologies advance and data volumes grow, this dual approach becomes increasingly essential for interpreting the role of microorganisms in environmental and health contexts.

3.4. Literature Search Results

The papers used in this review investigated different types of environmental features for microbiome forecasting. Figure 1 shows the number of selected papers published each year, covering the period from 2012 until May 2025.

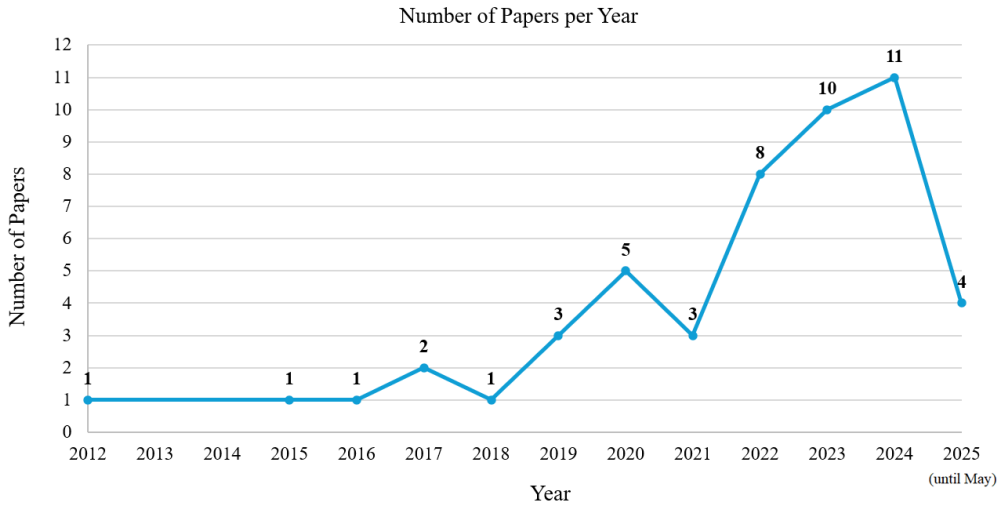

Figure 1. Annual number of publications on the microbiome–environment topic (2012 to May 2025).

Regarding publishers, the distribution of selected papers is as follows: 26 papers from Elsevier (52.0%), 8 papers from Springer Nature (16%), 4 papers from Wiley (8%), 2 papers each from Oxford University Press, BioMed Central, ACS and Frontiers Media (4% each), and 1 paper each from ASM Journals, MDPI, Copernicus GmbH, and PLOS ONE (2% each) (Figure 2).

During this period of time, our literature search found no publications that employed unsupervised methods for predictive analysis of environmental microbiome data relevant to the research questions addressed in this review. Therefore, Tables 1 and 2 present all the studies that match the criteria specified in Section 2, for classical ML and DL algorithms, respectively.

Table 1. State-of-the-art of the present classical ML algorithms.

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Terrestrial (different land uses)	Hermans, S. et al. (2020) [71]	16S rRNA	Predicting soil quality and land use	Composition of soil bacterial communities, bacterial Operational Taxonomic Unit (OTU) tables	3000 samples	RF (hold-out validation)	Accuracy = 85%	No
Terrestrial (agricultural fields)	Chang, H. et al. (2017) [72]	Shotgun Metagenomics	Predicting soil productivity	OTU abundances, environmental, soil and crop productivity data	12 samples	RF (leave-one-out cross-validation)	Accuracy = 79%	No
Terrestrial (pine litter decomposition)	Thompson, J et al. (2019) [73]	16S rRNA	Predicting dissolved organic carbon (OC) concentration	OTU abundances, OC concentration	302 samples	RF (hold-out validation)	$R = 0.676$	No
Terrestrial (agricultural soils across Europe)	Sørensen, M. B. et al. (2025) [74]	ITS	Predicting crop health	Abiotic and biotic data, normalized difference vegetation indexes	885 samples	RF (5-fold cross-validation)	$R^2 = 0.58$, RMSE = 0.14	No
Terrestrial (farmland soils across the USA and Canada)	Wilhelm, R. C. et al. (2022) [58]	eDNA metabarcoding	Predict ecological quality status	Amplicon sequence variant (ASV) abundance profiles, organic matter content, respiration, autoclaved citrate extractable (ACE) protein, active carbon, pH, phosphorus, potassium, minor elements, aggregate stability, available water capacity, surface and subsurface hardness, tillage status	949 samples	RF, SVM (hold-out validation)	$R^2 = 0.80$	Farmland vs. pasture-land soils
Terrestrial (soil, root, and rhizosphere)	Hagen, M. et al. (2024) [75]	16S rRNA	Analysying drought stress impact on microbiome	Relative abundances of bacterial taxa	332 samples	RF (5-fold cross-validation)	Accuracy = 92.3%	Sorghum-Drought vs. Grass-Drought datasets

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Terrestrial (soil samples)	Novielli, P. et al. (2024) [76]	Not mentioned	Analyzing climate change impact on soil health	Environmental, soil microbiota, biochemical recalcitrance and mineral protection factors	623 samples	RF, ETC, LoR, XGB, DT, SVM, KNN (10-fold cross-validation)	Accuracy = 92.3%, AUC = 0.964	No
Terrestrial (soil samples)	Chen, S. et al. (2024) [77]	Not mentioned	Predicting nanomaterials impact on the microbiome	Nanomaterial features (type, size, exposure dose), duration, pH, soil organic matter content, microbial diversity, biomass, enzyme activities	2134 paired observation	RF, XGB (hold-out validation)	R^2 = 0.71	No
Terrestrial (globally distributed soils samples)	Xu, N. et al. (2022) [78]	16S rRNA	Correlating nanoparticles properties with microbiome stability	Nanoparticles and soil characteristics	365 samples	RF (10-fold cross-validation)	R^2 = 0.91	No
Terrestrial (acidic sandy loam)	Chen, B. et al. (2025) [79]	16S rRNA	Predicting heterocyclic compounds impact on the microbiome	Chemical structure, environmental and microbial features	156 samples	XGB (hold-out validation)	R^2 = 0.94, RMSE = 0.008	No
Terrestrial (different land uses)	Ebrahimi, M. et al. (2017) [80]	Not applicable	Predicting Azotobacteria population in soil	pH, electrical conductivity, calcium carbonate equivalent, OC, sand/silt/clay content, hot/cold water extractable OC, light/heavy fraction OC, basal respiration, substrate-induced respiration, bacteria, fungi and actinomycetes counts	50 samples	ANN, Multivariate Linear Regression (MLR) (hold-out validation)	R^2 = 0.76, RMSE = 0.36	No

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Terrestrial (global agroecosystems with Fusarium-susceptible crops)	Sadeghi, S. et al. (2023) [81]	Not applicable	Predicting soil microbial communities based on soil physical and chemical properties under different agricultural management systems and soil depths	Bulk density, sand, silt, clay content, ammonium, nitrate, phosphorus, potassium, electrical conductivity, pH, organic matter, total phospholipid fatty acid, management treatments and soil depths	538 samples	Cubist algorithm (5-fold cross-validation combined with hold-out validation)	$R^2 = 0.96$	No
Terrestrial (six crops across nine countries/ regions)	Yuan, J. et al. (2020) [82]	16S rRNA, ITS	Predicting the occurrence of Fusarium wilt disease in soils	OTUs relative abundances	1549 samples	RF, SVM, LoR (hold-out validation)	Accuracy $> 80\%$	Independent datasets and new field soil samples
Terrestrial (diverse bioregions and land uses)	Xue, P. et al. (2024) [83]	Amplicon	Predicting spatial distributions of dominant bacterial and fungal phyla across Australia, identifying key environmental and the human impacts on microbiome	Relative abundances of dominant phyla, land use, soil type, climate factors (mean annual aridity index, annual precipitation, temperature range, solar radiation, etc), OC, pH, total nitrogen, total phosphorus, total sulphur, electrical conductivity, cation exchange capacity, and clay content	1384 samples	RF, QRF (10-fold cross-validation)	$R^2 = 0.90$	No
Aquatic (contaminated groundwater and water samples)	Smith M. B. et al. (2015) [84]	16S rRNA	Predicting water contamination and geochemical conditions	OTU relative abundances	93 samples	RF (cross-validation)	Accuracy = 82%	No

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Aquatic (bioreactors)	Liu, B. et al. (2022) [85]	16S rRNA	Predicting bioreactor production	OTU relative abundances	54 samples	RF (hold-out validation)	Accuracy = 90%	No
Aquatic (marine benthic)	Cordier, T. et al. (2018) [86]	SSU RNA	Predicting quality status associated with salmon farms	OTU relative abundances, reference biotic indices	144 samples	RF (hold-out validation)	R^2 = 0.89	No
Aquatic (marine benthic)	Frühe, L. et al. (2020) [87]	SSU RNA	Predicting environmental quality of marine aquaculture	OTUs relative abundances	152 samples	RF, SVM (hold-out validation)	R^2 = 0.72	Independent datasets
Aquatic (water column samples)	Janßen, R. et al. (2019) [88]	16S rRNA	Predicting water contamination	Taxon count table	32 samples	RF, ANNs (hold-out validation)	Accuracy = 97.10%	No
Aquatic (stream mesocosms)	Hempel, C. A. et al. (2023) [89]	16S rRNA, ITS, metagenomics, and total RNA sequencing	Predicting environmental stressor levels	Taxa relative abundances	121 samples	KNN, SVML, Ridge, Lasso, RF, SVC, XGB (hold-out validation)	MCC = 0.45	No
Aquatic (groundwater)	Wijaya, J. et al. (2024) [90]	16S rRNA, Metagenomics	Predicting groundwater pollution	Microbial families relative abundances, genes and pathways abundances	35 samples	LoR, SVML, SVMRBF, RF, DT (hold-out validation)	Accuracy = 98%, AUC = 0.99	No
Aquatic (groundwater)	Wijaya, J. et al. (2023) [91]	16S rRNA	Predicting groundwater pollution	OTUs relative abundances, pH, dissolved oxygen, oxidation-reduction potential, electrical conductivity, total petroleum hydrocarbon, temperature, turbidity	42 samples	LoR, SVML, SVMRBF, RF (hold-out validation)	Accuracy = 99%, AUC = 0.99	No

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Aquatic (single long-term agricultural field trial)	Mo, Y. et al. (2024) [92]	16S rRNA, ITS	Identifying key agricultural factors for microbial community	Microbial community data relative abundances, environmental variables (soil pH, total carbon, total nitrogen, soil moisture and soil bulk density), agricultural practices (fertility source, tillage and cover crop application)	96 samples	RF (10-fold cross-validation)	$R^2 > 0.95$, AUC = 0.996	No
Aquatic (three wastewater treatment plants)	Oh, S. et al. (2024) [93]	16S rRNA, metagenomic and metatranscrip-tomic sequencing	Predicting Clostridium perfringens surveillance	Clostridium perfringens abundance, meteorological variables	66 samples	OLR, LRLR, LRRR, SVRL, SVRRBF, RF, ABR, GBR (hold-out validation)	$R^2 = 0.78$	No
Aquatic (sediment water)	Jing, Z. et al. (2025) [94]	16S rRNA	Tracing human activities causing water pollution	OTUs relative abundances, environmental and geographical indices (spatio-temporal, social development, meteorological, physicochemical indicators), microbiological indices (metacommunity type, Shannon diversity, Simpson diversity, ACE diversity metrics)	915 samples	ANN, RF, XGB, LGBM, KNN, SVM (hold-out validation)	$R^2 = 0.924$	No
Aquatic (lake and river)	Kang, J. et al. (2022) [95]	16S rRNA	Predicting the relations between antibiotic features and aquatic bacteria	Physical and chemical properties of antibiotics, microbial diversity indices, relative abundance of bacterial modules, functional pathways	Not mentioned	RF (10-fold cross-validation)	$R^2 = 0.78$	No

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Aquatic (wastewater treatment plants)	Wijaya, J. and Oh, S. (2023) [96]	16S rRNA	Identifying keystone taxa	OTU relative abundances, operational data	38 samples	KNN, DT, LoR, SVML, SVMRBF, RF, LR, SVRL, SVRRBF, RFR (hold-out validation)	Accuracy $\geq$ 91.6%, R^2 = 0.98, MSE = 0.34	No
Aquatic (river catchments)	Zhu, Z. et al. (2024) [97]	16S rRNA, shotgun metagenomics	Predicting river's nitrogen pollution sources	Taxonomic composition and profiles, functional gene annotations, macroscopic characteristics (land use, soil type, elevation, river morphology (length, depth, slope, width)), physicochemical and sediment parameters	Not mentioned	RF (hold-out validation)	Accuracy = 84%, Kappa coefficient = 0.70	Geographically distinct area
Aquatic (wastewater treatment plant)	Wang, L. et al. (2024) [98]	16S rRNA	Identifying the environmental factors that affect microbial communities	OTUs relative abundances, latitude, longitude, climate type, solids retention time, hydraulic retention time, liquor suspended solids, influent biochemical oxygen demand, total nitrogen, total phosphorus, pH, dissolved oxygen, temperature, precipitation	1262 samples	Extremely Randomized Trees (hold-out validation)	Accuracy = 71.43%	Independent dataset
Aquatic (wastewater treatment plant)	Kim, Y. and Oh, S. (2021) [99]	16S rRNA	Predicting operational conditions and identify key microbial taxa	OTU tables, relative abundance of microbial taxa, PCA-transformed coordinates	18 samples	SVML, LoR, SVM, SVMRBF, RF, DT, KNN (hold-out validation)	Accuracy = 93%, AUC = 0.99	No

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Aquatic (coastal marine area)	Larsen, P. E. et al. (2012) [100]	16S rRNA	Predicting microbial community structure	Relative abundance of 24 bacterial orders, in situ and satellite-derived parameters	Not mentioned	ANN (hold-out validation)	Bray-Curtis similarity = 89.7	Hypothetical environmental conditions
Aquatic (seawater)	Glasl, B. et al. (2019) [101]	16S rRNA	Identifying reef microbiomes to use as environmental conditions indicators	OTUs relative abundance, sea surface temperature, chlorophylla, total suspended solids, particulate OC and other water quality parameters	381 samples	RF (hold-out validation)	Accuracy = 92%, Kappa coefficient 88%, R^2 = 0.67, RMSE = 0.5	No
Aquatic (open-ocean habitats)	Lambert, B. S. et al. (2022) [102]	Transcriptomics	Predicting the trophic mode of protists in marine environments	Transcriptomes, gene families, nutrient availability, sea surface temperature, light levels, microbial biomass	>541 samples	RF, XGB (5-fold cross-validation)	Accuracy = 81%, Cohen's k = 0.90	No
Aquatic (river and creek)	Dubinsky, E. A. et al. (2016) [103]	DNA microarray	Identifying and distinguishing fecal contamination sources in water samples	16S rRNA gene fragments	134 samples	RF, SourceTracker (leave-one-out cross-validation)	AUC = 0.97, Sensitivity = 100%, Specificity = 100%	Challenge and field Samples
Aquatic (river)	Wang, C. et al. (2021) [104]	16S rRNA, metagenomics	Predicting the source of water samples	Values of physicochemical indices, abundance data of microbial indices and combination of both	252 samples	RF (hold-out validation)	Kappa Coefficient = 0.8694	No

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Aquatic (wastewater treatment plants from different countries)	Liu, X. et al. (2023) [105]	16S rRNA	Predict microbial compositions	48 environmental data points, relative abundances of bacterial/archaeal taxa, alpha diversity indices (Shannon–Wiener, Pielou's evenness, species richness, Faith's phylogenetic diversity), ASVs and functional groups	777 samples	ANN (hold-out validation)	R^2 = 0.6286	No
Air (indoor environments)	Hampton-Marcell, J. T. et al. (2023) [106]	16S rRNA	Identifying microbial taxa	ASV relative abundance	38 samples	RF (validation not mentioned)	AUC = 0.60	No
Air (different subtracts and materials)	Choi, Y. et al. (2024) [107]	Not mentioned	Differentiating of microbial non-volatile and volatile organic compounds	Non-volatile and volatile organic compounds	261 samples	RF (10-fold cross-validation)	Accuracy = 100%	No
Air (municipal solid waste treatment)	Fang, R. et al. (2023) [108]	16S rRNA	Identifying seasonal exposure biomarkers	ASVs, waste, throat swabs, temperature, humidity, PM2.5, PM10, O33, SO22, NO22, CO, air quality index, waste type, seasonal and spatial factors	71 samples	RF (10-fold cross-validation)	Accuracy = 100%, Precision = 1.00, AUC = 1.00	No
Air (diverse public facilities)	Lee, B. et al. (2025) [109]	Not mentioned	Predicting airborne fungal concentrations	Facility type, floor level, month, air temperature, relative humidity, coarse PM2.5–10, precipitation	137 samples	ENR, KNN, SVR, RF, GB, XGB, DT (hold-out validation)	MAE = 0.42, MSE = 0.28, RMSE = 0.53, R^2 = 0.78	No

Table 1. *Cont.*

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Air (two types of pig houses)	Peng, S. et al. (2023) [110]	16S rRNA	Quantifying the influence of environmental factors	Air pollutants concentrations, OTU relative abundance	48 samples	ABT (cross-validation)	R^2 = 0.969	No
Air (commercial office and shopping mall samples)	Lee, J.Y.Y. et al. (2023) [111]	Not applicable	Estimating concentrations of bioaerosols and particulate matter	Sensors' physicochemical data, ultraviolet light-induced fluorescence (UV-LIF) observations, bioaerosol and PM concentrations	30513 time-series data points	LR, Lasso Regression, RF, XGB (hold-out validation)	Willmott's Index = 0.82	No
Terrestrial, Aquatic (field groundwater samples)	Miao, Y. et al. (2023) [112]	16S rRNA, shotgun metagenomic	Predicting contaminant levels and duration	Microbial taxa relative abundance, 1,4-dioxane and chlorinated solvents concentration, dissolved oxygen, oxidation-reduction potential, pH, total OC, temperature, sampling depth, aquifer material, and injection of electron donor	102 samples	RF, MLR, LGBM, AdaBoost, GBR, SVM, NB, KNN (6-fold cross-validation)	Accuracy = 57%, Kappa coefficient = 0.56, R^2 = 0.81	Independent field datasets
Terrestrial, Aquatic (sediment from coastal salmon aquaculture sites)	Dully, V. et al. (2021) [113]	eDNA metabarcoding	Evaluating eDNA metabarcoding for classifying sediment samples into environmental quality categories	ASVs, Infaunal Quality Index scores	12 samples	RF (leave-one-out cross-validation)	R^2 = 0.91	No

Table 2. State-of-the-art of the present DL algorithms.

Environment	Article	Sequencing Technique	Classification	Input Data	Number of Samples	Model	Metric	Novel Test Context
Terrestrial (simulated environments and real co-culture experiments)	Lee, J. et al. (2020) [114]	Not applicable	Predicting microbial interactions	Fluorescence microscopy	35000 images	CNN, ResNet (5-fold cross-validation)	R^2 = 0.844	No
Terrestrial (italian and philippine rhizosphere)	García-Jiménez, B. (2020) et al. [69]	16S rRNA	Predicting microbial composition from phenotypic features	Temperature, precipitation, plant age, maize line, and variety, microbial abundance profiles	4724 samples	Autoencoder (hold-out validation)	R = 0.7348	Hypothetical climate change scenarios
Terrestrial (five ecosystem types)	Yang, Y. et al. (2022) [115]	ITS	Predicting fungal abundance and diversity	Visible/near-infrared spectra, soil properties, climate, vegetation and terrain data, fungal phyla relative abundances	577 samples	CNN (10-fold cross-validation)	R^2 = 0.73	No
Terrestrial (different soil types)	Wang, Y. and Zou, Q. (2024) [116]	16S rRNA, ITS	Predicting soil-borne fungal diseases	Bacterial and fungal ASV features	6715 samples	DCA + MLP & RF (hold-out and leave-one-out cross-validation)	Accuracy > 90%	No
Aquatic (estuary and a mariculture samples)	Jiang, J. et al. (2022) [117]	16S rRNA	Predicting the relative abundance of *Vibrio* spp.	Temperature, dissolved oxygen, salinity, pH, total nitrogen, total phosphorus and relative abundance of *Vibrio* spp.	150 sets of experimental data	DNN, RF, SVR, ElasticNet, XGB (hold-out validation)	RMSE = 12.16, MAE = 6.67	No
Air (commercial office and shopping mall samples)	Lee, J.Y.Y. et al. (2023) [111]	Not applicable	Estimating concentrations of bioaerosols and particulate matter	Physical and chemical data from sensors, UV-LIF observations, bioaerosol/PM concentrations	30513 time-series data points	LSTM, MLP and RNN (hold-out validation)	Willmott's Index = 0.82	No
All (different "theaters of activity" samples)	Zeng, W. et al. (2022) [118]	Shotgun Metagenomics	Predicting the "theater of activity" of a microbiome	Taxonomic and functional profiles	6048 samples	DeepToA (hold-out validation)	Accuracy = 98.30%	No

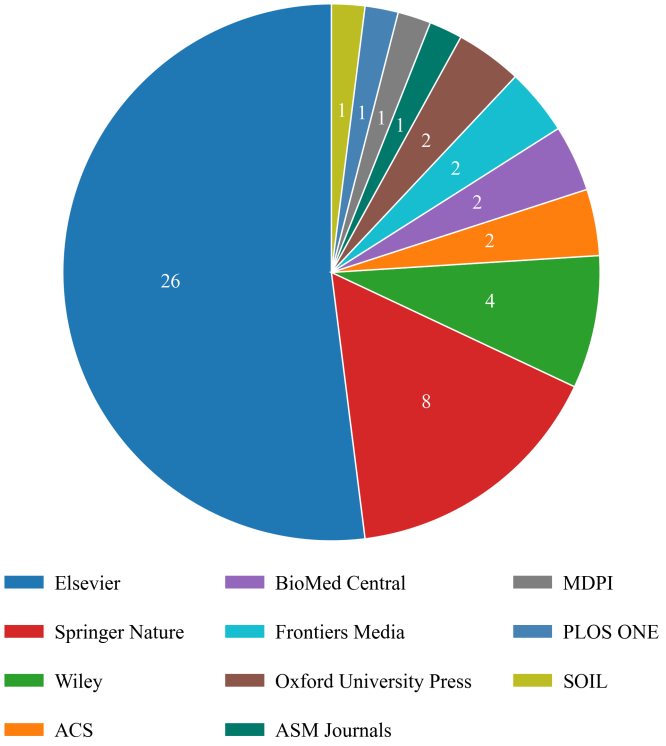

Figure 2. Total number of microbiome–environment papers published by each publisher.

Figure 3 presents a comparison between the performance of ML and DL models independent of the application environment type, while Figure 4 illustrates the comparative performance of these models across individual and combined environment types. Only the distributions of R^2 and accuracy across a range of models were considered, as these were the most commonly reported metrics in the reviewed papers. In cases where values were reported as thresholds (e.g., >90), the threshold value (e.g., 90) was used for the distribution analysis.

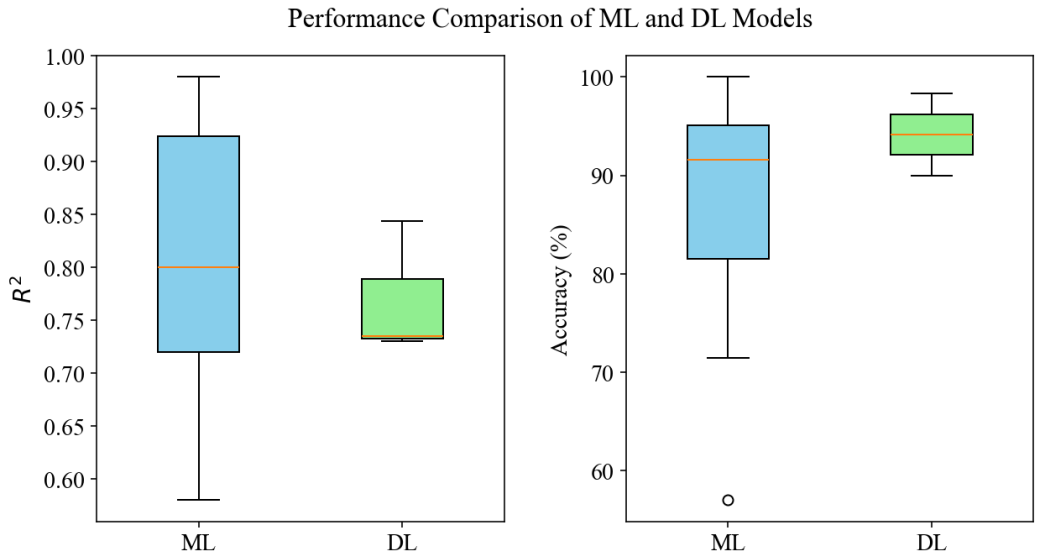

Figure 3. Comparison of ML and DL models in terms of R^2 and Accuracy. The central lines in each box represent the median values. Blue boxes correspond to ML models, and green boxes correspond to DL models.

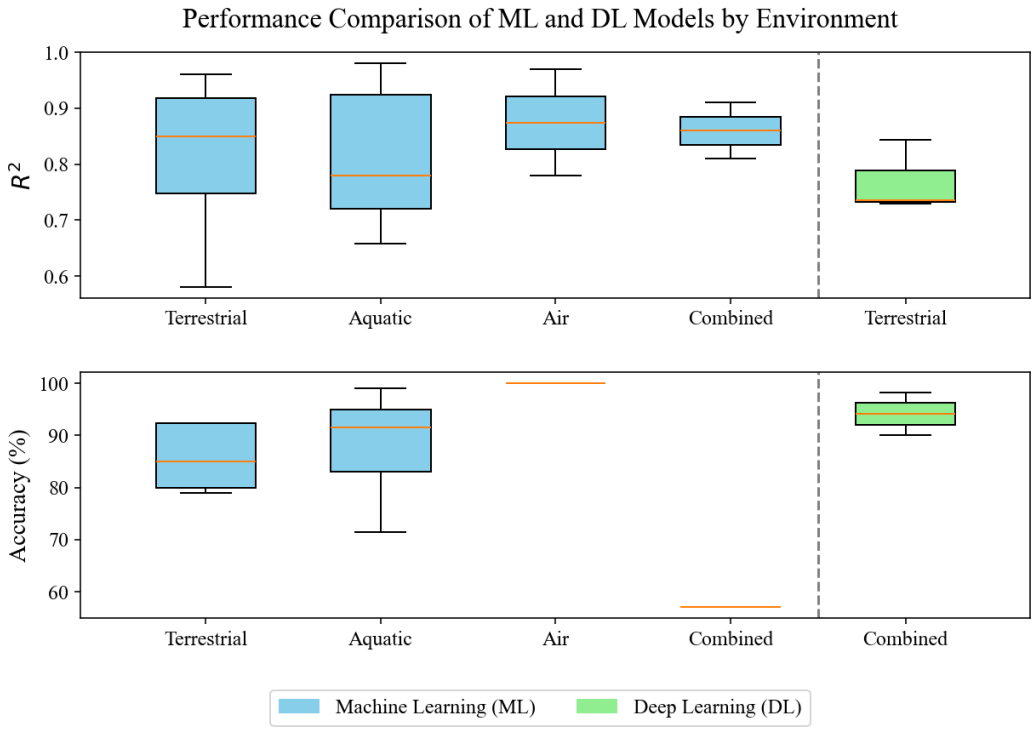

Figure 4. Comparison of ML and DL models by environment in terms of R^2 and accuracy. The central lines in each box represent the median values. The dotted line separates ML from DL results.

Analysis of the Results

For the studies presented in Tables 1 and 2 (summarized in terms of R^2 and accuracy boxplots of Figures 3 and 4) the most commonly used performance metrics are R^2 and accuracy, while the most common environment type is terrestrial, for both ML and DL approaches. When analyzing the studies that reported the highest performance in each environment, it is evident that, for the aquatic environment, the highest performance achieved was 99% accuracy, reported by [91]. This study aimed to train ML models, including the RF, LoR, SVML, SVMRBF models, to predict the status and source of groundwater pollution. Feature importance analysis identified members of the families *Rhodocyclaceae* and *Syntrophaceae* as strong indicators of groundwater polluted with total petroleum hydrocarbons, consistent with their known ecophysiological traits. The usefulness of these microbial indicators was validated through conventional hypothesis testing and phylogenetic analysis. In the study presented in [96], ML models were developed using microbiome data to successfully predict the operational characteristics of three full-scale wastewater systems. For categorical prediction, the models developed included KNN, DT, LoR, SVML, SVMRBF and RF. For regression tasks, the study used LR, SVRL, SVRRBF, and RFR. This study demonstrates that ML can effectively capture complex, non-linear relationships between keystone taxa and environmental conditions—relationships that may be undetectable using conventional methods. Regarding the remaining studies, performance was also satisfactory, with reported accuracy and R^2 values equal to or greater than 71.43% and 0.72, respectively.

For the terrestrial environment, the same performance metrics (accuracy and R^2) were the most commonly used. The studies presented in [75,76] both achieved an accuracy of 92.3%. The first one used an RF model to classify soil drought stress based on the composition of the soil microbiome. The study was able to understand which specific microbes contributed most to the model's predictions, thereby uncovering their roles in drought response. The ML model demonstrated a strong performance, indicating its potential for practical applications in agriculture. These findings highlight the complementary value

of ML and traditional metagenomic analyses in identifying relevant marker taxa and advancing the understanding of soil microbial responses to environmental stressors. In [76], the aim was to understand how various soil characteristics, particularly the soil microbiome, influence the soil's sensitivity to warming, known as Q10. The research trained and evaluated several ML models, which demonstrated strong performance in predicting high or low Q10 values. The study also identified the most influential factors driving these predictions and contributed to addressing the urgent global challenge of climate change, which impacts soil health and, consequently, food security. In terms of R^2, the study [79] reported the highest value (0.94). This research integrated chemical structure properties with soil bacterial diversity and functional data to predict the impact of heterocyclic compounds on soil microbial communities and to support the design of eco-friendly alternatives. An XGB model was used along with Shapley Additive Explanation analysis, through which key chemical structures that help maintain high soil bacterial diversity and functionality were identified. For the same environment, the other studies also demonstrated good overall performance across all metrics. Accuracy values ranged from 78% to 92.3%, while R^2 values ranged from 0.58 to 0.95.

The air environment has been the focus of fewer studies compared to the aquatic and terrestrial environments. Nevertheless, the models generally demonstrated good performance and a similar pattern was observed concerning the most commonly used metrics. For this environment, an accuracy of 100% was obtained in [107,108]. Choi Y. et al. [107] used an RF model to accurately identify indoor fungal contamination by distinguishing microbial volatile organic compounds from other indoor volatile organic compounds. The model was trained on data from various fungal species cultivated on different substrates, along with emissions from common indoor materials. The ML approach effectively identified specific volatile compounds as key indicators of fungal presence and presents a robust method for assessing indoor air quality and detecting fungal growth. Fang R. et al. [108] used ML to analyze the dynamics of inhalable bacteria in municipal solid waste treatment systems and their potential hazards to on-site workers. Also using an RF model, the authors identified microbial biomarkers that distinguish between summer and winter exposure conditions. For R^2, the highest value achieved was 0.969, as reported in [110]. This study investigated the relationship between bacteria and the environment within a piggery, using full-length 16S rRNA gene sequencing. An ABT model was used to identify the influence of air pollutants on bacterial abundance. The results showed that PM2.5 concentration was the most influential factor, followed by ammonia and carbon dioxide levels. These findings highlight the potential health risks of airborne pathogens for pigs and farm workers, as well as the need for improved waste management and environmental controls.

Independent on the environment, the authors of DeepToA's significantly improved metagenomics analysis. DeepToA, in contrast to prior techniques that relied exclusively on 16S community profile data, uses metagenomic data and integrates both taxonomic and functional profiles. This allows for a more in-depth examination of microbiomes and their associated "theatres of activity" (ToAs). It differs from prior techniques, including taxonomic and functional profiles, allowing for a more thorough study. DeepToA explains its predictions based on the initial profiles, emphasizing crucial taxonomic and functional features. A pre-trained embedding matrix is included in the framework for effectively mapping taxonomy information. DeepToA's performance and classification skills, which reached an accuracy of 98.30%, can be enhanced with additional data.

Regarding validation techniques, there are also clear trends and preferences. The most common method is hold-out validation (62%), making it the dominant choice. Following in prevalence are k-fold cross-validation methods (34%), with 10-fold cross-validation

being the most popular (16%). Five-fold cross-validation is also frequently applied (10%), while six-fold cross-validation (2%) and generic mentions of "cross-validation" without a specified number of folds (4%) appear less frequently. Leave-one-out cross-validation (LOOCV) is used four times (8%) and a few cases employ hybrid validation strategies, such as five-fold cross-validation combined with hold-out and hold-out combined with LOOCV. Only one study lacks a specified validation method.

The transformation of data into the final features commonly involved normalization, reported in the majority of studies (64%). This was implemented in various forms, including relative abundance calculations, Hellinger transformation, min-max scaling, total sum scaling (TSS), and centered Log ratio (CLR) transformation. Rarefaction was also frequently used (26%), either alone or in combination with other techniques, to standardize sequencing depth across samples. In addition, many studies applied filtering steps (32%) to remove low-abundance features or to select features based on criteria such as prevalence or interquartile range (IQR). Several studies also employed additional techniques such as log-transformation, cumulative sum scaling (CSS), standardization, or, less commonly, dimensionality reduction. However, not all studies specified the data transformation methods used (12%).

4. Discussion

Several elements of the application of AI algorithms to forecast the ambient microbiome were highlighted in this study. From the start, we can see that this field is very new since the publications shown in Tables 1 and 2 were published between 2012 and 2025. To create vast quantities of data, new sequencing technologies had to mature and become more accessible, only possible in the last years.

The studies demonstrate that two main types of habitats were mainly selected regarding the origin of the analyzed microbiomes: soils and aquatic environments. For soil, the research concentrates on their multiple features to verify shifts in the microbiome, such as predicting their productivity, pollution, or geographical origin. On the other hand, aquatic environment research, which includes oceans, bioreactors, and even fishponds, focuses primarily on employing microbiomes to analyze contaminants and water quality.

Most of the reviewed studies did not explicitly analyze the impact of sequencing depth, marker choice, or similar features on ML and DL predictive performance. As a result, direct comparisons or systematic evaluations of their effects were uncommon. However, some studies acknowledged their indirect influence on data quality and the robustness of the models. For example, factors such as primer selection, target regions and OTU filtering thresholds were noted to shape microbial community profiles and, consequently, the quality of the input data used in ML and DL pipelines. The choice of marker, such as selecting between bacterial, fungal or other communities, was identified as an important factor for model accuracy and prediction stability. Similarly, sequencing depth and the distinction between metagenomic and 16S rRNA data affected the resolution and diversity of microbial features, which in turn influenced model performance. In contrast, many studies emphasized environmental, chemical, or physical variables as key contributors to predictive accuracy. Some also pointed out that combining different types of data, including taxonomic, functional, and environmental information, or using high-dimensional input features, such as the full OTU table or top-ranked OTUs, led to notable improvements in model performance.

Despite this, the reviewed studies demonstrated a wide range of approaches to analyzing feature importance in ML and DL models. Although not all studies explicitly performed feature importance evaluations, many incorporated techniques to identify the most influential input variables. Random Forest was frequently used, often relying on its

built-in importance metrics such as the Gini index, Mean Decrease in Accuracy (MDA), Mean Decrease Gini (MDG), and percentage increase in Mean Squared Error (MSE). Other techniques included Recursive Feature Elimination (RFE), permutation importance, and ensemble methods, such as MUVR, Boruta, and VSURF. A few studies applied more domain-specific tools, such as indicator species analysis, Garson's connection weights for neural networks, or Bayesian network inference through Environmental Interaction Networks. Some studies assessed feature relevance using model-specific metrics like node purity or cross-validation-based ranking. Although the methods for evaluating feature importance varied across model types, there was a consistent effort to extract and report the relative importance of features, reflecting a growing interest in understanding the factors influencing model outcomes.

In terms of algorithm types, we can see that the RF type has been chosen more frequently above all ML models. This type of algorithm has many benefits that may explain why it is used in this type of research, including its flexibility, capacity to handle complicated interactions, and interoperability. These algorithms' ability to handle several kinds of data, such as categorical and numerical data, contributes to their flexibility. This is useful for microbiome data, which may include various information, such as the presence or absence of certain species. Because the microbiome is impacted by multiple associated elements such as soil composition, nutrient availability, temperature, and humidity, these algorithms can capture these complex interactions, enabling the detection of patterns and non-linear correlations. Furthermore, these algorithms provide insight into the significance of the features utilized to create predictions. This may assist in determining which species or environmental traits are most valuable for forecasting changes in the microbiome in the context of the environmental microbiome. This information is also crucial for interpretability since its operation helps us comprehend the decision-making process at each node and can offer an overview of feature significance. This may help researchers learn more about the variables that influence changes in the ambient microbiome.

DL algorithms have the ability to identify and understand complex patterns in environment microbiome data, even when such patterns are not explicitly described. For this particular field, the algorithm more used was the CNN. Compared to classical algorithms, DL algorithms may lead to a better understanding of the interactions and linkages within the microbiome. However, a few drawbacks may be restricting for a still-developing research area. First, these technique models often need substantial computer power and massive data sets for appropriate training, which limits their utility in certain situations. In addition to the quantity of data required, the quality of the data is critical to ensuring excellent results and the model's generalization capability, that is, its ability to perform well on data on which it was not trained. Finally, these approaches are often regarded as "black boxes", which indicates that the model's judgments might be challenging to analyze and explain. This lack of transparency can limit our ability to understand the underlying factors driving model predictions and hide the complex interactions within the ambient microbiome. As a result, even though these models often work well, their complexity can make it hard to understand the biological reasons behind their predictions and may reduce confidence in their results. Because of this, there is growing interest in explainable AI (XAI) methods, such as SHAP (Shapley Additive Explanations) [117,118]. Despite these advances, DL models still face interpretability challenges compared to traditional ML models, which are inherently more transparent. However, it is essential to remember that with the advancement of microbiome technologies and methodologies, ensuring the quantity and quality of data provided to these kinds of models, they can detect more complex relationships with greater accuracy in their predictions, compared with classical ML, even if these are not explainable. As sequencing technologies continue to improve and computa-

tional resources expand, these DL approaches become increasingly powerful for analyzing the intricate patterns within microbial communities. The interpretability gap remains a significant challenge in the field, as researchers must often balance predictive performance against the ability to understand the biological mechanisms driving these predictions. Nevertheless, careful data curation, appropriate validation strategies, and integration with domain knowledge can help mitigate these limitations, allowing researchers to leverage the predictive power of DL while maintaining scientific rigor in microbiome research.

In the reviewed studies, there is a clear predominance of the hold-out method as the primary validation technique. However, cross-validation approaches are also present in a significant portion of the presented literature. The main strength of cross-validation lies in the model selection and hyperparameter tuning phase. It provides a more stable and reliable measure of a model's generalization ability, enabling to confidently choose the best algorithm or configuration. The popularity of the hold-out method relies on its simplicity, computational efficiency, and strong performance on unseen data, especially when dealing with large datasets. This method closely simulates a real-world scenario in which a model is trained and then deployed in production, where it encounters entirely new data [119,120]. The predominance of the hold-out method in the reviewed studies may reflect an emphasis on this final validation step, which is often the most important step before using the model in the real world.

The analysis of the various models' performance is constrained by the fact that they are not evaluated using the same measures, as shown in Tables 1 and 2. However, we can see that the DL models (Table 2) do not exhibit any values lower than any of the classical supervised models (Table 1), whose lowest values were observed in [73,74,112] when considering the same metrics. For the work presented in [111], which employed both classical ML and DL approaches, similar performance was achieved by both methods. Therefore, we can declare that they equal or outperform the classical models in all measures.

Analyses in Tables 1 and 2 also reveal athat studies of microbiomes using AI taken from surfaces or air samples represent a significant gap that has to be addressed in future AI research. Surface samples were only handled by unsupervised algorithms using sequencing techniques that analyze the complete genetic material obtained rather than approaches that analyze certain DNA parts and are specialized for various microbes. Regarding air samples, only 7 studies were published, which is a low number compared to the 17 and 15 studies using aquatic and terrestrial samples, respectively. It is also worth noting that DL algorithms were the only ones that allowed for a thorough investigation of the microbiome since the model published in [118] examined microbiomes with aquatic, terrestrial, and aerial origins.

Finally, our review findings reveal the significant sustainability implications of AI applications in environmental microbiome monitoring. The surveyed literature demonstrates alignment with multiple UN Sustainable Development Goals, with AI-enabled predictions supporting water quality management (SDG 6) [121], climate adaptation strategies (SDG 13) [122], and ecosystem conservation (SDGs 14 and 15) [123,124]. The cost-effectiveness of continuous microbiome surveillance systems represents a promising advancement for proactive environmental management, particularly in resource-limited regions. Despite these advancements, our analysis identified critical gaps that must be addressed to fully realize sustainability benefits. The integration of microbiome predictions with circular bioeconomy applications remains largely theoretical, with empirical studies testing these integrative approaches notably scarce in the literature.

4.1. Critical Analysis for the Selected Papers

Directly comparing the performance of different models is challenging, as they are evaluated using different performance metrics and datasets. In addition, the studies differ

significantly in methodology and outcome measures. These factors limit the feasibility of making direct comparisons or conducting a meaningful meta-analysis. Despite this, the presented studies demonstrated strong performance according to the criteria established by their authors, providing insight into effective applications of ML in microbiome research. These examples show how certain algorithms can perform well in real-world conditions and help guide future research. Unsupervised methods were also not present and could help us understand the structure and functioning of the microbiome since these methods may develop representations of latent features, assisting in the identification of non-linear correlations and hidden information in microbiome data [14,125]. Furthermore, these algorithms do not need labels or explicit data supervision. This might be useful when there is a lack of labeled data or uncertain or subjective classifications. This can help uncover natural groupings, underlying patterns, or hidden structures within the data, such as identifying distinct microbial community types, detecting outliers, or generating hypotheses for further investigation.

Regarding DL, it has only started to be explored in this particular field since 2020 and the number of papers is still limited compared to those using traditional ML. While current DL models show good performance, their complexity and "black-box" nature can limit interpretability and biological insight, which may partly explain why they have not been explored as extensively. However, their ability to automatically learn hierarchical features from raw data, especially in large datasets, as the ones presented in Table 2, makes them a powerful tool for microbiome analysis. DL can potentially reveal complex patterns that traditional ML methods might miss. Nevertheless, challenges related to model transparency and the need for substantial computational resources should be taken into account when applying DL approaches.

4.2. Limitations

This review encountered several methodological constraints inherent to investigating an emerging interdisciplinary field. The intersection of AI with environmental microbiome research represents a relatively recent domain, resulting in a limited corpus of directly relevant publications. Despite implementing an exhaustive search strategy employing multiple keyword combinations across major scientific databases, the current body of literature specifically addressing predictive modeling of microbiome dynamics remains modest, albeit with an anticipated acceleration in publication frequency as the field matures.

Our review methodology necessarily excluded numerous studies that established correlative relationships between environmental parameters and microbiome composition. While these investigations provided valuable insights into microbiome–environment associations, they did not progress to implementing ML or DL algorithms to test predictive capabilities, a critical criterion for inclusion in our analysis. Specifically, we excluded studies that identified environmental drivers of microbial community structure but failed to develop computational frameworks capable of generating testable predictions about microbiome responses to changing environmental conditions. This stringent inclusion criterion, while limiting our sample size, was essential to address our primary research question regarding the predictive performance of various AI approaches in environmental microbiome forecasting.

In general, the reviewed studies revealed persistent challenges in data harmonization, where variability in sampling protocols and incomplete metadata did not make possible cross-study comparisons. Inconsistent protocols can introduce in microbial diversity estimates, directly compromising model generalizability. A more fundamental challenge is that microbial community composition does not always align with metabolic function. For example, Wang et al. [126] found that acetoclastic methanogenesis remained stable

even as microbial populations shifted. The lack of alignment between community composition and function limits the ability of models based only on 16S rRNA data to detect key metabolic processes, potentially limiting their usefulness for environmental forecasting. Furthermore, although multi-omics data such as metabolomics and microbial community profiles offer rich insights into environmental processes, their integration into AI frameworks remains limited. The study by Zhao et al. [127] demonstrates the potential of combining metabolomic and 16S rRNA data to understand rhizosphere dynamics; however, such approaches are rarely adopted in predictive modeling due to data heterogeneity and lack of standardized pipelines.

Another limitation is that some studies rely on small or narrowly focused datasets. Many models were trained on data with limited sample sizes, underrepresentation of key environmental conditions, or restricted geographic and temporal scopes. For instance, some studies explicitly noted that their models could be improved by including more and more diverse samples [58,71,89,94,95]. This lack of data diversity not only reduces the statistical power of the analyses but also introduces a risk of sampling bias, where findings may not accurately reflect the ecological patterns they aim to describe.

Lastly, gaps in interpretability and generalizability remain. Several models were developed using data from specific locations or a limited number of habitats, raising concerns about their applicability to other ecosystems [58,125]. This limitation reduces the potential for developing universally applicable predictive tools. The "black-box" nature of some complex models, such as artificial neural networks, may further obscure the underlying biological mechanisms behind predictions, even when the outputs are accurate [105]. Moreover, many studies relied solely on taxonomic data, while knowing that the integration of additional data layers—such as environmental parameters, functional genomics, or metabolomics—is necessary to build more insightful and accurate models [73,116]. The predominance of 16S rRNA data, although valuable for taxonomic profiling, limits functional inference. The absence of functional gene or metatranscriptomic data restricts the ability of models to predict ecosystem-level processes, particularly those involving complex biogeochemical pathways [128].

5. Conclusions

In conclusion, this study highlights the growing but rapidly evolving topic of using AI systems to predict ambient microbiome traits. Most research is focused on terrestrial and aquatic ecosystems, leaving a substantial gap in studying microbiomes from surfaces or air samples. The study of terrestrial settings focuses mainly on soil qualities, while aquatic studies concentrate on water quality and contaminants. Classical algorithms, particularly RF, were frequently selected for their ability to handle the complexity and variability typical of microbiome datasets, which are impacted by various variables. DL models, although computationally costly and sometimes seen as "black boxes", may be more successful at revealing complex patterns and hidden properties in microbiome data. Performance comparisons between these models remain difficult owing to different assessment metrics, yet DL models match, if not surpass, their classical counterparts. This evidence is clearly shown by comparing Zeng and co-authors' study [118] with Janssen and colleagues' study [88], where a greater prediction accuracy has been achieved with a DL model.

6. Future Directions

With the increasing accessibility of sequencing technologies and the advancement of AI algorithms—particularly unsupervised approaches—this field is expected to experience major breakthroughs in the coming years. These developments will expand our

understanding of environmental microbiomes and uncover deeper, previously unknown patterns in microbial data, especially when leveraging integrated multi-omics approaches that combine metagenomics, metabolomics, and/or transcriptomics [129]. Incorporating functional data such as stable isotope analysis into these models, as shown in studies of methanogenesis [126,128], can further improve predictions of microbial activity despite taxonomic variability.

Improving data quality will be essential to support these advancements and can be achieved through standardized sampling protocols, comprehensive metadata annotation, and cross-study data integration. Simultaneously, focusing on previously unexplored or underrepresented environments, such as air samples and urban surfaces, will broaden the scope of microbiome research. These neglected environments represent promising frontiers for discovery, particularly when studied through innovative approaches like real-time monitoring systems. Moreover, integrating spatial data into microbiome analyses can reveal ecological patterns relevant to climate-sensitive conservation strategies [130]. Collaborations among microbiologists, data scientists, ecologists, and clinicians will be crucial to translating microbiome insights into real-world applications, particularly in environmental monitoring, and help address critical research gaps. For example, recent findings on degradable microplastics and biochar interactions with soil organic matter highlight the role of microbial processes in carbon cycling under anthropogenic stress. Combining stable isotope techniques with multi-omics data from impacted soils can reveal microbial feedback and long-term carbon stability [131].

As computational capabilities continue to evolve, real-time microbiome monitoring and predictive modeling may become increasingly feasible, enabling proactive responses to environmental and microbial changes. Hybrid modeling approaches that combine ecological theory with ML will be particularly valuable for these applications, offering both predictive power and interoperability. Additionally, the adoption of new compilers may improve performance and scalability, facilitating the efficient processing of large microbiome datasets and enabling advanced models to run on edge devices or in the cloud.

Finally, enhancing the interpretability of AI models through the development of explainable AI techniques will improve their transparency, trustworthiness, and acceptance within the scientific community, while future research should simultaneously prioritize developing these explainable methods specifically for microbiome data and expanding studies in developing countries where sustainability challenges are most acute, ensuring these powerful tools can be effectively implemented while addressing the dual imperatives of ecosystem health and human development needs.

By combining these technological advances with rigorous standardization and interdisciplinary collaboration, the field can transform our ability to understand and manage microbial communities in changing environments.

Author Contributions: Conceptualization, M.I.B., G.S. and P.M.R.; methodology, M.I.B., G.S., P.R. and P.M.R.; validation, E.V., A.P., P.M. and P.M.R.; writing—original, M.I.B. and G.S.; writing—review and editing, M.I.B., P.R., E.V., A.P., P.M. and P.M.R.; supervision, P.M. and P.M.R.; funding acquisition, P.M. and P.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by HAC4CG—Heritage, Art, Creation for Climate Change—project. Living in the city: catalyzing spaces for learning, creation, and action towards climate change. NORTE-45-2020-75. SISTEMA DE APOIO À INVESTIGAÇÃO CIENTÍFICA E TECNOLÓGICA—"PROJETOS ESTRUTURADOS DE I&D&I" HORIZONTE EUROPA and FCT for funding through the Strategic Projects CITAR [UIDB/0622/2020 and UIDP/0622/2020] (https://doi.org/10.54499/UIDB/00622/2020 and https://doi.org/10.54499/UIDP/00622/2020) and CBQF [UIDB/50016/2020] (https://doi.org/10.54499/UIDP/50016/2020).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Song, D.; Huo, T.; Zhang, Z.; Cheng, L.; Wang, L.; Ming, K.; Liu, H.; Li, M.; Du, X. Metagenomic Analysis Reveals the Response of Microbial Communities and Their Functions in Lake Sediment to Environmental Factors. *Int. J. Environ. Res. Public Health* **2022**, *19*, 16870. [CrossRef]
2. Carles, L.; Wullschleger, S.; Joss, A.; Eggen, R.I.; Schirmer, K.; Schuwirth, N.; Stamm, C.; Tlili, A. Wastewater microorganisms impact microbial diversity and important ecological functions of stream periphyton. *Water Res.* **2022**, *225*, 119119. [CrossRef]
3. Shade, A.; Peter, H.; Allison, S.D.; Baho, D.L.; Berga, M.; Bürgmann, H.; Huber, D.H.; Langenheder, S.; Lennon, J.T.; Martiny, J.B.H.; et al. Fundamentals of Microbial Community Resistance and Resilience. *Front. Microbiol.* **2012**, *3*, 417. [CrossRef]
4. Philippot, L.; Raaijmakers, J.M.; Lemanceau, P.; van der Putten, W.H. Going back to the roots: The microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* **2013**, *11*, 789–799. [CrossRef]
5. Rittmann, B.E. Biofilms, active substrata, and me. *Water Res.* **2018**, *132*, 135–145. [CrossRef] [PubMed]
6. Urbanek, A.K.; Rymowicz, W.; Mirończuk, A.M. Degradation of plastics and plastic-degrading bacteria in cold marine habitats. *Appl. Microbiol. Biotechnol.* **2018**, *102*, 7669–7678. [CrossRef]
7. Cavicchioli, R.; Ripple, W.J.; Timmis, K.N.; Azam, F.; Bakken, L.R.; Baylis, M.; Behrenfeld, M.J.; Boetius, A.; Boyd, P.W.; Classen, A.T.; et al. Scientists' warning to humanity: Microorganisms and climate change. *Nat. Rev. Microbiol.* **2019**, *17*, 569–586. [CrossRef]
8. Jansson, J.K.; Hofmockel, K.S. Soil microbiomes and climate change. *Nat. Rev. Microbiol.* **2019**, *18*, 35–46. [CrossRef] [PubMed]
9. Bøifot, K.O.; Gohli, J.; Moen, L.V.; Dybwad, M. Performance evaluation of a new custom, multi-component DNA isolation method optimized for use in shotgun metagenomic sequencing-based aerosol microbiome research. *Environ. Microbiome* **2020**, *15*, 1. [CrossRef] [PubMed]
10. McElhinney, J.M.W.R.; Catacutan, M.K.; Mawart, A.; Hasan, A.; Dias, J. Interfacing Machine Learning and Microbial Omics: A Promising Means to Address Environmental Challenges. *Front. Microbiol.* **2022**, *13*, 851450. [CrossRef]
11. Ghannam, R.B.; Techtmann, S.M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1092–1107. [CrossRef]
12. Rodrigues, P.M.; Madeiro, J.P.; Marques, J.A.L. Enhancing Health and Public Health through Machine Learning: Decision Support for Smarter Choices. *Bioengineering* **2023**, *10*, 792. [CrossRef]
13. Lo, C.; Marculescu, R. MetaNN: Accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinform.* **2019**, *20*, 314. [CrossRef]
14. Shi, Y.; Zhang, L.; Peterson, C.B.; Do, K.A.; Jenq, R.R. Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome* **2022**, *10*, 25. [CrossRef]
15. Papoutsoglou, G.; Tarazona, S.; Lopes, M.B.; Klammsteiner, T.; Ibrahimi, E.; Eckenberger, J.; Novielli, P.; Tonda, A.; Simeon, A.; Shigdel, R.; et al. Machine learning approaches in microbiome research: Challenges and best practices. *Front. Microbiol.* **2023**, *14*, 1261889. [CrossRef]
16. Johnson, C.N.; Balmford, A.; Brook, B.W.; Buettel, J.C.; Galetti, M.; Guangchun, L.; Wilmshurst, J.M. Biodiversity losses and conservation responses in the Anthropocene. *Science* **2017**, *356*, 270–275. [CrossRef] [PubMed]
17. Pecl, G.T.; Araújo, M.B.; Bell, J.D.; Blanchard, J.; Bonebrake, T.C.; Chen, I.C.; Clark, T.D.; Colwell, R.K.; Danielsen, F.; Evengård, B.; et al. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science* **2017**, *355*, eaai9214. [CrossRef] [PubMed]
18. Barnosky, A.D.; Matzke, N.; Tomiya, S.; Wogan, G.O.; Swartz, B.; Quental, T.B.; Marshall, C.; McGuire, J.L.; Lindsey, E.L.; Maguire, K.C.; et al. Has the Earth's sixth mass extinction already arrived? *Nature* **2011**, *471*, 51–57. [CrossRef] [PubMed]
19. Ripple, W.J.; Wolf, C.; Newsome, T.M.; Galetti, M.; Alamgir, M.; Crist, E.; Mahmoud, M.I.; Laurance, W.F.; 15,364 Scientist Signatories from 184 Countries. World scientists' warning to humanity: A second notice. *BioScience* **2017**, *67*, 1026–1028. [CrossRef]
20. Bellard, C.; Bertelsmeier, C.; Leadley, P.; Thuiller, W.; Courchamp, F. Impacts of climate change on the future of biodiversity. *Ecol. Lett.* **2012**, *15*, 365–377. [CrossRef]
21. Crist, E.; Mora, C.; Engelman, R. The interaction of human population, food production, and biodiversity protection. *Science* **2017**, *356*, 260–264. [CrossRef]
22. Panthee, B.; Gyawali, S.; Panthee, P.; Techato, K. Environmental and human microbiome for health. *Life* **2022**, *12*, 456. [CrossRef]
23. Staff, A. *Microbes and Climate Change—Science, People & Impacts*; Technical Report; American Society for Microbiology: Washington, DC, USA, 2022. [CrossRef]

24. Staff, A. *FAQ: Microbes and Climate Change*; Technical Report; American Society for Microbiology: Washington, DC, USA, 2017. Available online: https://www.ncbi.nlm.nih.gov/books/NBK513763/ (accessed on 6 August 2025).
25. Flandroy, L.; Poutahidis, T.; Berg, G.; Clarke, G.; Dao, M.C.; Decaestecker, E.; Furman, E.; Haahtela, T.; Massart, S.; Plovier, H.; et al. The impact of human activities and lifestyles on the interlinked microbiota and health of humans and of ecosystems. *Sci. Total Environ.* **2018**, *627*, 1018–1038. [CrossRef] [PubMed]
26. Tiedje, J.M.; Bruns, M.A.; Casadevall, A.; Criddle, C.S.; Eloe-Fadrosh, E.; Karl, D.M.; Nguyen, N.K.; Zhou, J. Microbes and Climate Change: A Research Prospectus for the Future. *mBio* **2022**, *13*, e00800-22. [CrossRef]
27. Gandolfi, I.; Bertolini, V.; Bestetti, G.; Ambrosini, R.; Innocente, E.; Rampazzo, G.; Papacchini, M.; Franzetti, A. Spatio-temporal variability of airborne bacterial communities and their correlation with particulate matter chemical composition across two urban areas. *Appl. Microbiol. Biotechnol.* **2015**, *99*, 4867–4877. [CrossRef]
28. Uetake, J.; Tobo, Y.; Uji, Y.; Hill, T.C.; DeMott, P.J.; Kreidenweis, S.M.; Misumi, R. Seasonal changes of airborne bacterial communities over Tokyo and influence of local meteorology. *Front. Microbiol.* **2019**, *10*, 1572. [CrossRef] [PubMed]
29. Zhen, Q.; Deng, Y.; Wang, Y.; Wang, X.; Zhang, H.; Sun, X.; Ouyang, Z. Meteorological factors had more impact on airborne bacterial communities than air pollutants. *Sci. Total Environ.* **2017**, *601*, 703–712. [CrossRef] [PubMed]
30. Bertolini, V.; Gandolfi, I.; Ambrosini, R.; Bestetti, G.; Innocente, E.; Rampazzo, G.; Franzetti, A. Temporal variability and effect of environmental variables on airborne bacterial communities in an urban area of Northern Italy. *Appl. Microbiol. Biotechnol.* **2013**, *97*, 6561–6570. [CrossRef]
31. Liu, H.; Zhang, X.; Zhang, H.; Yao, X.; Zhou, M.; Wang, J.; He, Z.; Zhang, H.; Lou, L.; Mao, W.; et al. Effect of air pollution on the total bacteria and pathogenic bacteria in different sizes of particulate matter. *Environ. Pollut.* **2018**, *233*, 483–493. [CrossRef]
32. Retter, A.; Karwautz, C.; Griebler, C. Groundwater Microbial Communities in Times of Climate Change. *Curr. Issues Mol. Biol.* **2021**, *41*, 509–538. [CrossRef]
33. Danczak, R.E.; Johnston, M.D.; Kenah, C.; Slattery, M.; Wilkins, M.J. Microbial community cohesion mediates community turnover in unperturbed aquifers. *Msystems* **2018**, *3*, e00066-18. [CrossRef] [PubMed]
34. Šantl-Temkiv, T.; Gosewinkel, U.; Starnawski, P.; Lever, M.; Finster, K. Aeolian dispersal of bacteria in southwest Greenland: Their sources, abundance, diversity and physiological states. *FEMS Microbiol. Ecol.* **2018**, *94*, fiy031. [CrossRef]
35. Li, H.; Yang, Q.; Li, J.; Gao, H.; Li, P.; Zhou, H. The impact of temperature on microbial diversity and AOA activity in the Tengchong Geothermal Field, China. *Sci. Rep.* **2015**, *5*, 17056. [CrossRef]
36. Zhong, S.; Zhang, L.; Jiang, X.; Gao, P. Comparison of chemical composition and airborne bacterial community structure in PM2.5 during haze and non-haze days in the winter in Guilin, China. *Sci. Total Environ.* **2019**, *655*, 202–210. [CrossRef]
37. Sogin, M.L.; Morrison, H.G.; Huber, J.A.; Welch, D.M.; Huse, S.M.; Neal, P.R.; Arrieta, J.M.; Herndl, G.J. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 12115–12120. [CrossRef]
38. Caporaso, J.G.; Lauber, C.L.; Walters, W.A.; Berg-Lyons, D.; Huntley, J.; Fierer, N.; Owens, S.M.; Betley, J.; Fraser, L.; Bauer, M.; et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **2012**, *6*, 1621–1624. [CrossRef]
39. Caracciolo, A.B.; Topp, E.; Grenni, P. Pharmaceuticals in the environment: Biodegradation and effects on natural microbial communities. A review. *J. Pharm. Biomed. Anal.* **2015**, *106*, 25–36. [CrossRef]
40. Caruso, G.; Azzaro, M.; Caroppo, C.; Decembrini, F.; Monticelli, L.S.; Leonardi, M.; Maimone, G.; Zaccone, R.; Ferla, R.L. Microbial community and its potential as descriptor of environmental status. *ICES J. Mar. Sci.* **2016**, *73*, 2174–2177. [CrossRef]
41. Kress, W.J.; García-Robledo, C.; Uriarte, M.; Erickson, D.L. DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.* **2015**, *30*, 25–35. [CrossRef] [PubMed]
42. Valentini, A.; Pompanon, F.; Taberlet, P. DNA barcoding for ecologists. *Trends Ecol. Evol.* **2009**, *24*, 110–117. [CrossRef] [PubMed]
43. Johnson, J.S.; Spakowicz, D.J.; Hong, B.Y.; Petersen, L.M.; Demkowicz, P.; Chen, L.; Leopold, S.R.; Hanson, B.M.; Agresta, H.O.; Gerstein, M.; et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **2019**, *10*, 5029. [CrossRef]
44. Badotti, F.; de Oliveira, F.S.; Garcia, C.F.; Vaz, A.B.M.; Fonseca, P.L.C.; Nahum, L.A.; Oliveira, G.; Góes-Neto, A. Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi). *BMC Microbiol.* **2017**, *17*, 42. [CrossRef] [PubMed]
45. Wagner, B.D.; Grunwald, G.K.; Zerbe, G.O.; Mikulich-Gilbertson, S.K.; Robertson, C.E.; Zemanick, E.T.; Harris, J.K. On the Use of Diversity Measures in Longitudinal Sequencing Studies of Microbial Communities. *Front. Microbiol.* **2018**, *9*, 1037. [CrossRef] [PubMed]
46. Qian, X.B.; Chen, T.; Xu, Y.P.; Chen, L.; Sun, F.X.; Lu, M.P.; Liu, Y.X. A guide to human microbiome research: Study design, sample collection, and bioinformatics analysis. *Chin. Med. J.* **2020**, *133*, 1844–1855. [CrossRef]
47. Walters, K.E.; Martiny, J.B.H. Alpha-, beta-, and gamma-diversity of bacteria varies across habitats. *PLoS ONE* **2020**, *15*, e0233872. [CrossRef]

48. Mande, S.S.; Mohammed, M.H.; Ghosh, T.S. Classification of metagenomic sequences: Methods and challenges. *Briefings Bioinform.* **2012**, *13*, 669–681. [CrossRef]
49. Sharpton, T.J. An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* **2014**, *5*, 209. [CrossRef]
50. Handelsman, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 669–685. [CrossRef]
51. Ross, E.M.; Moate, P.J.; Bath, C.R.; Davidson, S.E.; Sawbridge, T.I.; Guthridge, K.M.; Cocks, B.G.; Hayes, B.J. High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. *BMC Genet.* **2012**, *13*, 53. [CrossRef]
52. Dahui, Q. Next-generation sequencing and its clinical application. *Cancer Biol. Med.* **2019**, *16*, 4–10. [CrossRef] [PubMed]
53. Lynch, M.D.J.; Neufeld, J.D. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **2015**, *13*, 217–229. [CrossRef]
54. Liu, S.; Moon, C.D.; Zheng, N.; Huws, S.; Zhao, S.; Wang, J. Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome* **2022**, *10*, 76. [CrossRef]
55. Haykin, S.O. *Neural Networks and Learning Machines*, 3rd ed.; Pearson: Upper Saddle River, NJ, USA, 2008.
56. Sathya, R.; Abraham, A. Comparison of supervised and unsupervised learning algorithms for pattern classification. *Int. J. Adv. Res. Artif. Intell.* **2013**, *2*, 34–38. [CrossRef]
57. Kotsiantis, S.B.; Zaharakis, I.; Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **2007**, *160*, 3–24.
58. Wilhelm, R.C.; van Es, H.M.; Buckley, D.H. Predicting measures of soil health using the microbiome and supervised machine learning. *Soil Biol. Biochem.* **2022**, *164*, 108472. [CrossRef]
59. Ribeiro, P.; Barbosa, M.I.; Sousa, C.; Rodrigues, P.M. Near-Infrared Spectroscopy Machine-Learning Spectral Analysis Tool for Blueberries (Vaccinium corymbosum) Cultivar Discrimination. *Foods* **2025**, *14*, 1428. [CrossRef]
60. Rodrigues, P.M.; Bispo, B.C.; Garrett, C.; Alves, D.; Teixeira, J.P.; Freitas, D. Lacsogram: A new EEG tool to diagnose Alzheimer's disease. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3384–3395. [CrossRef] [PubMed]
61. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
62. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
63. Soman, K.; Loganathan, R.; Ajay, V. *Machine Learning with SVM and Other Kernel Methods*; PHI Learning Pvt. Ltd.: Delhi, India, 2009.
64. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
65. Fiannaca, A.; La Paglia, L.; La Rosa, M.; Lo Bosco, G.; Renda, G.; Rizzo, R.; Gaglio, S.; Urso, A. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinform.* **2018**, *19*, 61–76. [CrossRef]
66. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning*; Cambridge University Press: Cambridge, UK, 2014.
67. Houdt, G.V.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
68. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
69. García-Jiménez, B.; Muñoz, J.; Cabello, S.; Medina, J.; Wilkinson, M.D. Predicting microbiomes through a deep latent space. *Bioinformatics* **2020**, *37*, 1444–1451. [CrossRef]
70. Prince, S.J. *Understanding Deep Learning*; The MIT Press: Cambridge, MA, USA, 2023.
71. Hermans, S.M.; Buckley, H.L.; Case, B.S.; Curran-Cournane, F.; Taylor, M.; Lear, G. Using soil bacterial communities to predict physico-chemical variables and soil quality. *Microbiome* **2020**, *8*, 79. [CrossRef]
72. Chang, H.X.; Haudenshield, J.S.; Bowen, C.R.; Hartman, G.L. Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. *Front. Microbiol.* **2017**, *8*, 519. [CrossRef] [PubMed]
73. Thompson, J.; Johansen, R.; Dunbar, J.; Munsky, B. Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLoS ONE* **2019**, *14*, e0215502. [CrossRef]
74. Sørensen, M.B.; Faurdal, D.; Schiesaro, G.; Jensen, E.D.; Jensen, M.K.; Clemmensen, L.K.H. Exploring crop health and its associations with fungal soil microbiome composition using machine learning applied to remote sensing data. *Commun. Earth Environ.* **2025**, *6*, 355. [CrossRef]
75. Hagen, M.; Dass, R.; Westhues, C.; Blom, J.; Schultheiss, S.J.; Patz, S. Interpretable machine learning decodes soil microbiome's response to drought stress. *Environ. Microbiome* **2024**, *19*, 35. [CrossRef]
76. Novielli, P.; Magarelli, M.; Romano, D.; de Trizio, L.; Di Bitonto, P.; Monaco, A.; Amoroso, N.; Stellacci, A.M.; Zoani, C.; Bellotti, R.; et al. Climate Change and Soil Health: Explainable Artificial Intelligence Reveals Microbiome Response to Warming. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1564–1578. [CrossRef]
77. Chen, S.; Teng, Y.; Luo, Y.; Kuramae, E.; Ren, W. Threats to the soil microbiome from nanomaterials: A global meta and machine-learning analysis. *Soil Biol. Biochem.* **2024**, *188*, 109248. [CrossRef]

78. Xu, N.; Kang, J.; Ye, Y.; Zhang, Q.; Ke, M.; Wang, Y.; Zhang, Z.; Lu, T.; Peijnenburg, W.; Penuelas, J.; et al. Machine learning predicts ecological risks of nanoparticles to soil microbial communities. *Environ. Pollut.* **2022**, *307*, 119528. [CrossRef]
79. Chen, B.; Liu, M.; Zhang, Z.; Lv, B.; Yu, Y.; Zhang, Q.; Xu, N.; Yang, Z.; Lu, T.; Xia, S.; et al. Data-Driven Approach for Designing Eco-Friendly Heterocyclic Compounds for the Soil Microbiome. *Environ. Sci. Technol.* **2025**, *59*, 1530–1541. [CrossRef]
80. Ebrahimi, M.; Safari Sinegani, A.A.; Sarikhani, M.R.; Mohammadi, S.A. Comparison of artificial neural network and multivariate regression models for prediction of Azotobacteria population in soil under different land uses. *Comput. Electron. Agric.* **2017**, *140*, 409–421. [CrossRef]
81. Sadeghi, S.; Petermann, B.J.; Steffan, J.J.; Brevik, E.C.; Gedeon, C. Predicting microbial responses to changes in soil physical and chemical properties under different land management. *Appl. Soil Ecol.* **2023**, *188*, 104878. [CrossRef]
82. Yuan, J.; Wen, T.; Zhang, H.; Zhao, M.; Penton, C.R.; Thomashow, L.S.; Shen, Q. Predicting disease occurrence with high accuracy based on soil macroecological patterns of Fusarium wilt. *ISME J.* **2020**, *14*, 2936–2950. [CrossRef]
83. Xue, P.; Minasny, B.; Wadoux, A.M.J.; Dobarco, M.R.; McBratney, A.; Bissett, A.; de Caritat, P. Drivers and human impacts on topsoil bacterial and fungal community biogeography across Australia. *Glob. Change Biol.* **2024**, *30*, e17216. [CrossRef]
84. Smith, M.B.; Rocha, A.M.; Smillie, C.S.; Olesen, S.W.; Paradis, C.; Wu, L.; Campbell, J.H.; Fortney, J.L.; Mehlhorn, T.L.; Lowe, K.A.; et al. Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. *mBio* **2015**, *6*, 00326-15. [CrossRef] [PubMed]
85. Liu, B.; Sträuber, H.; Saraiva, J.; Harms, H.; Silva, S.G.; Kasmanas, J.C.; Kleinsteuber, S.; da Rocha, U.N. Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture. *Microbiome* **2022**, *10*, 48. [CrossRef] [PubMed]
86. Cordier, T.; Forster, D.; Dufresne, Y.; Martins, C.I.M.; Stoeck, T.; Pawlowski, J. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* **2018**, *18*, 1381–1391. [CrossRef]
87. Frühe, L.; Cordier, T.; Dully, V.; Breiner, H.W.; Lentendu, G.; Pawlowski, J.; Martins, C.; Wilding, T.A.; Stoeck, T. Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Mol. Ecol.* **2020**, *30*, 2988–3006. [CrossRef]
88. Janßen, R.; Zabel, J.; von Lukas, U.; Labrenz, M. An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. *Mar. Pollut. Bull.* **2019**, *149*, 110530. [CrossRef]
89. Hempel, C.A.; Buchner, D.; Mack, L.; Brasseur, M.V.; Tulpan, D.; Leese, F.; Steinke, D. Predicting environmental stressor levels with machine learning: A comparison between amplicon sequencing, metagenomics, and total RNA sequencing based on taxonomically assigned data. *Front. Microbiol.* **2023**, *14*, 1217750. [CrossRef]
90. Wijaya, J.; Park, J.; Yang, Y.; Siddiqui, S.I.; Oh, S. A metagenome-derived artificial intelligence modeling framework advances the predictive diagnosis and interpretation of petroleum-polluted groundwater. *J. Hazard. Mater.* **2024**, *472*, 134513. [CrossRef]
91. Wijaya, J.; Byeon, H.; Jung, W.; Park, J.; Oh, S. Machine learning modeling using microbiome data reveal microbial indicator for oil-contaminated groundwater. *J. Water Process Eng.* **2023**, *53*, 103610. [CrossRef]
92. Mo, Y.; Bier, R.; Li, X.; Daniels, M.; Smith, A.; Yu, L.; Kan, J. Agricultural practices influence soil microbiome assembly and interactions at different depths identified by machine learning. *Commun. Biol.* **2024**, *7*, 1349. [CrossRef] [PubMed]
93. Oh, S.; Byeon, H.; Wijaya, J. Machine learning surveillance of foodborne infectious diseases using wastewater microbiome, crowdsourced, and environmental data. *Water Res.* **2024**, *265*, 122282. [CrossRef] [PubMed]
94. Jing, Z.; Zhang, Y.; Liu, X.; Li, Q.; Hao, Y.; Li, Y.; Gao, H. Identifying human activities causing water pollution based on microbial community sequencing and source classifier machine learning. *Environ. Int.* **2025**, *195*, 109240. [CrossRef]
95. Kang, J.; Zhang, Z.; Chen, Y.; Zhou, Z.; Zhang, J.; Xu, N.; Zhang, Q.; Lu, T.; Peijnenburg, W.; Qian, H. Machine learning predicts the impact of antibiotic properties on the composition and functioning of bacterial community in aquatic habitats. *Sci. Total Environ.* **2022**, *828*, 154412. [CrossRef]
96. Wijaya, J.; Oh, S. Machine learning reveals the complex ecological interplay of microbiome in a full-scale membrane bioreactor wastewater treatment plant. *Environ. Res.* **2023**, *222*, 115366. [CrossRef]
97. Zhu, Z.; Ding, J.; Du, R.; Zhang, Z.; Guo, J.; Li, X.; Jiang, L.; Chen, G.; Bu, Q.; Tang, N.; et al. Systematic tracking of nitrogen sources in complex river catchments: Machine learning approach based on microbial metagenomics. *Water Res.* **2024**, *253*, 121255. [CrossRef]
98. Wang, L.; Lu, W.; Song, Y.; Liu, S.; Fu, Y.V. Using machine learning to identify environmental factors that collectively determine microbial community structure of activated sludge. *Environ. Res.* **2024**, *260*, 119635. [CrossRef]
99. Kim, Y.; Oh, S. Machine-learning insights into nitrate-reducing communities in a full-scale municipal wastewater treatment plant. *J. Environ. Manag.* **2021**, *300*, 113795. [CrossRef] [PubMed]
100. Larsen, P.E.; Field, D.; Gilbert, J.A. Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods* **2012**, *9*, 621–625. [CrossRef]
101. Glasl, B.; Bourne, D.G.; Frade, P.R.; Thomas, T.; Schaffelke, B.; Webster, N.S. Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome* **2019**, *7*, 94. [CrossRef]

102. Lambert, B.S.; Groussman, R.D.; Schatz, M.J.; Coesel, S.N.; Durham, B.P.; Alverson, A.J.; White, A.E.; Armbrust, E.V. The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2100916119. [CrossRef] [PubMed]
103. Dubinsky, E.A.; Butkus, S.R.; Andersen, G.L. Microbial source tracking in impaired watersheds using PhyloChip and machine-learning classification. *Water Res.* **2016**, *105*, 56–64. [CrossRef]
104. Wang, C.; Mao, G.; Liao, K.; Ben, W.; Qiao, M.; Bai, Y.; Qu, J. Machine learning approach identifies water sample source based on microbial abundance. *Water Res.* **2021**, *199*, 117185. [CrossRef] [PubMed]
105. Liu, X.; Nie, Y.; Wu, X.L. Predicting microbial community compositions in wastewater treatment plants using artificial neural networks. *Microbiome* **2023**, *11*, 93. [CrossRef]
106. Hampton-Marcell, J.T.; Ghosh, A.; Gukeh, M.J.; Megaridis, C.M. A new approach of microbiome monitoring in the built environment: feasibility analysis of condensation capture. *Microbiome* **2023**, *11*, 129. [CrossRef]
107. Choi, Y.; Kang, B.; Kim, D. Effective detection of indoor fungal contamination through the identification of volatile organic compounds using mass spectrometry and machine learning. *Environ. Pollut.* **2024**, *363*, 125195. [CrossRef]
108. Fang, R.; Chen, T.; Han, Z.; Ji, W.; Bai, Y.; Zheng, Z.; Su, Y.; Jin, L.; Xie, B.; Wu, D. From air to airway: Dynamics and risk of inhalable bacteria in municipal solid waste treatment systems. *J. Hazard. Mater.* **2023**, *460*, 132407. [CrossRef]
109. Lee, B.G.; Jeong, K.H.; Kim, H.E.; Yeo, M.K. Machine learning models for predicting indoor airborne fungal concentrations in public facilities utilizing environmental variables. *Environ. Pollut.* **2025**, *368*, 125684. [CrossRef]
110. Peng, S.; Luo, M.; Long, D.; Liu, Z.; Tan, Q.; Huang, P.; Shen, J.; Pu, S. Full-length 16S rRNA gene sequencing and machine learning reveal the bacterial composition of inhalable particles from two different breeding stages in a piggery. *Ecotoxicol. Environ. Saf.* **2023**, *253*, 114712. [CrossRef] [PubMed]
111. Lee, J.Y.; Miao, Y.; Chau, R.L.; Hernandez, M.; Lee, P.K. Artificial intelligence-based prediction of indoor bioaerosol concentrations from indoor air quality sensor data. *Environ. Int.* **2023**, *174*, 107900. [CrossRef]
112. Miao, Y.; Zhou, T.; Zheng, X.; Mahendra, S. Investigating Biodegradation of 1,4-Dioxane by Groundwater and Soil Microbiomes: Insights into Microbial Ecology and Process Prediction. *ACS ES T Water* **2023**, *4*, 1046–1060. [CrossRef]
113. Dully, V.; Balliet, H.; Frühe, L.; Däumer, M.; Thielen, A.; Gallie, S.; Berrill, I.; Stoeck, T. Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture—An inter-laboratory study. *Ecol. Indic.* **2021**, *121*, 107049. [CrossRef]
114. Lee, J.Y.; Sadler, N.C.; Egbert, R.G.; Anderton, C.R.; Hofmockel, K.S.; Jansson, J.K.; Song, H.S. Deep learning predicts microbial interactions from self-organized spatiotemporal patterns. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1259–1269. [CrossRef]
115. Yang, Y.; Shen, Z.; Bissett, A.; Viscarra Rossel, R.A. Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions. *Soil* **2022**, *8*, 223–235. [CrossRef]
116. Wang, Y.; Zou, Q. Deep learning meta-analysis for predicting plant soil-borne fungal disease occurrence from soil microbiome data. *Appl. Soil Ecol.* **2024**, *202*, 105532. [CrossRef]
117. Jiang, J.; Zhou, H.; Zhang, T.; Yao, C.; Du, D.; Zhao, L.; Cai, W.; Che, L.; Cao, Z.; Wu, X.E. Machine learning to predict dynamic changes of pathogenic *Vibrio* spp. abundance on microplastics in marine environment. *Environ. Pollut.* **2022**, *305*, 119257. [CrossRef]
118. Zeng, W.; Gautam, A.; Huson, D.H. DeepToA: An ensemble deep-learning approach to predicting the theater of activity of a microbiome. *Bioinformatics* **2022**, *38*, 4670–4676. [CrossRef]
119. Goodfellow, I.A.; Courville.; Bengio, Y. *Deep Learning*; Adaptive Computation and Machine Learning; The MIT Press: Cambridge, MA, USA, 2016.
120. Soper, D.S. Greed Is Good: Rapid Hyperparameter Optimization and Model Selection Using Greedy k-Fold Cross Validation. *Electronics* **2021**, *10*, 1973. [CrossRef]
121. Ortigara, A.; Kay, M.; Uhlenbrook, S. A Review of the SDG 6 Synthesis Report 2018 from an Education, Training, and Research Perspective. *Water* **2018**, *10*, 1353. [CrossRef]
122. Campbell, B.M.; Hansen, J.; Rioux, J.; Stirling, C.M.; Twomlow, S.; (Lini) Wollenberg, E. Urgent action to combat climate change and its impacts (SDG 13): Transforming agriculture and food systems. *Curr. Opin. Environ. Sustain.* **2018**, *34*, 13–20. [CrossRef]
123. Recuero Virto, L. A preliminary assessment of the indicators for Sustainable Development Goal (SDG) 14 "Conserve and sustainably use the oceans, seas and marine resources for sustainable development". *Mar. Policy* **2018**, *98*, 47–57. [CrossRef]
124. Ishtiaque, A.; Masrur, A.; Rabby, Y.W.; Jerin, T.; Dewan, A. Remote Sensing-Based Research for Monitoring Progress towards SDG 15 in Bangladesh: A Review. *Remote Sens.* **2020**, *12*, 691. [CrossRef]
125. Jiang, Y.; Luo, J.; Huang, D.; Liu, Y.; dan Li, D. Machine Learning Advances in Microbiology: A Review of Methods and Applications. *Front. Microbiol.* **2022**, *13*, 925454. [CrossRef] [PubMed]
126. Wang, S.; Sun, A.; Yang, S.; Ni, R.; Lin, X.; Shu, W.; Price, G.; Song, L. Dominance of acetoclastic methanogenesis in municipal solid waste (MSW) decomposition despite high variability in microbial community composition: Insights form natural stable carbon isotope and metagenomic analyses. *Energy Environ. Sustain.* **2025**, *1*, 100018. [CrossRef]

127. Zhao, M.; Zou, G.; Li, Y.; Pan, B.; Wang, X.; Zhang, J.; Xu, L.; Li, C.; Chen, Y. Biodegradable microplastics coupled with biochar enhance Cd chelation and reduce Cd accumulation in Chinese cabbage. *Biochar* **2025**, *7*, 31. [CrossRef]
128. Chu, Y.; Zhang, X.; Tang, X.; Jiang, L.; He, R. Uncovering anaerobic oxidation of methane and active microorganisms in landfills by using stable isotope probing. *Environ. Res.* **2025**, *271*, 121139. [CrossRef] [PubMed]
129. Pang, Q.; Zhao, G.; Wang, D.; Zhu, X.; Xie, L.; Zuo, D.; Wang, L.; Tian, L.; Peng, F.; Xu, B.; et al. Water periods impact the structure and metabolic potential of the nitrogen-cycling microbial communities in rivers of arid and semi-arid regions. *Water Res.* **2024**, *267*, 122472. [CrossRef] [PubMed]
130. Teron, G.; Bordoloi, R.; Paul, A.; Singha, L.B.; Tripathi, O.P. Effect of altitude on soil physico-chemical properties and microbial biomass carbon in the Eaglenest Wildlife Sanctuary of Arunachal Pradesh. *Geol. Ecol. Landscapes* **2024**, 1–19. [CrossRef]
131. Lin, J.; Cheng, Q.; Kumar, A.; Zhang, W.; Yu, Z.; Hui, D.; Zhang, C.; Shan, S. Effect of degradable microplastics, biochar and their coexistence on soil organic matter decomposition: A critical review. *TrAC Trends Anal. Chem.* **2025**, *183*, 118082. [CrossRef]