World Journal of Biology Pharmacy and Health Sciences

eISSN: 2582-5542

Cross Ref DOI: 10.30574/wjbphs

Journal homepage: https://wjbphs.com/

(Review Article)

Predictive modeling of microbial functionality from 16S rDNA sequences using machine learning

Abhaykumar Dalsaniya [1, *], Urvisha Beladiya [2] and Ramesh K. Kothari [3]

[1] *LTI Mindtree, Limited, USA.*
[2] *Department of Biosciences, Veer Narmad South Gujarat University, Surat, Gujarat, India.*
[3] *UGC-CAS Department of Biosciences, Saurashtra University, Rajkot-360005 Gujarat-INDIA.*

World Journal of Biology Pharmacy and Health Sciences, 2025, 21(02), 545-554

Publication history: Received on 13 January 2025; revised on 15 February 2025; accepted on 18 February 2025

Article DOI: https://doi.org/10.30574/wjbphs.2025.21.2.0223

Abstract

This study illustrates the possibilities for the use of machine learning algorithms integrating the prediction of microbial functionality regarding functional 16S rDNA sequences and filling the gap in mapping phylogenetic relations of the microbial context to their functionality. Previous 16S rDNA sequencing strategies have proven useful in describing microbial species, but not their functional potential. This study employed several sophisticated forms of supervised and unsupervised machine learning algorithms to interpret 16S rDNA data and predict the functional states of microbes in different contexts. The samples were obtained from public sources of genomic data. After the necessary pre-processing, the data were used to train different classifiers, including Random Forests, Support Vector Machines, and Neural networks. The results suggest that functional prediction enhancement using machine learning is effective because the algorithms reveal patterns and correlations in massive multifaceted genetic data. This improved possibility is closely related to areas such as medicine, environmentalism, and the practical application of bioengineering. This study also highlights data heterogeneity and model generalization issues and provides suggestions for improving predictive models for the future scope of microbial genomics.

Keywords: Microbial Functionality; 16S Rdna Sequencing; Metagenomics; Random Forests; Genomic Data

1. Introduction

Microbial diversity can impact the stability and health of ecosystems. The microbial web involves many processes vital to ecosystem health, such as nutrient cycling, decomposition, and disease suppression (1). Knowledge of microbial communities and how they function is important for agricultural, environmental, and human health practices. Previously, 16S rDNA sequencing was the only method used to identify microbial species and provide information regarding the taxonomy of the examined samples (2). Nevertheless, 16S rDNA sequencing can accurately identify microbes at the taxonomic level; it is not useful for predicting the ecological or metabolic roles of microbes because it does not involve sequencing of operational genes engaged in functional activities (3).

These challenges have led to the development of ML techniques to analyze large data sets and provide unseen patterns that are not easily apparent when dissecting original data (4). By pairing 16S rDNA with environmental context, ML models can predict the functional potential of microbial communities if researchers only have taxonomic data (5). This approach provides a more integrated concept of microbial communities and how they relate to adaptation to specific functions, which is important in applying the microbial world, such as microbial therapeutics and bioremediation.

[*] Corresponding author: Abhaykumar Dalsaniya Orchid Id 0009-0003-7309-3455.

1.1. Overview

Machine learning (ML) methods are invaluable for analyzing biological data because they can be used to explore large, complicated datasets. The major types of learning that complete the ML framework are supervised and unsupervised and are sub-disciplined in these deep learning techniques. Using these techniques, genomic data can be analyzed effectively, and many insights and patterns can be derived (4). For example, supervised learning techniques can transform classified data feeds, as in Random Forests and Support Vector Machines, into predicting microbial functional outcomes from 16S rDNA sequences. Likewise, unsupervised learning methods, such as clustering, assist in grouping similar samples and establishing functional resemblances among microbial complexes (6).

Domain-level microbiota analysis using machine learning-based tools has allowed researchers to build models and understand microbial taxonomy of functional characteristics that are almost impossible with conventional approaches (5). For instance, machine learning can predict the metabolic function of microbes based on their genetics, which is important compared to other approaches for studying intricate microbial communities in environments such as the soil, ocean, and gut (3). Integrating computation with biology is a significant contribution to microbial ecology and can provide new avenues for future research and applications.

1.2. Problem Statement

However, this microbial taxonomy is yet to be well correlated with its functional relevance because of issues associated with genetically sequenced data. Although the 16S rDNA sequencing approach is reliable for microbial species identification, it needs to be more informative regarding the functionality of the microbes. This is because the mechanistic details of how unique microbes relate to different ecological processes remain inaccessible based on such separations, resulting in knowledge deficits, particularly in complicated matrices, such as soil, oceans, and gut microbiomes. Current approaches result in vague and dubious relationships or predictions that are more reliable. Therefore, the development of improved models would provide sound predictions of microbial traits in terms of 16S rDNA sequence data, which would help reduce the gap between taxonomic and ecophysiological perspectives.

1.3. Objectives

- To enhance the value of information about microbial communities by developing machine learning algorithms that can accurately predict the roles of microbes based on their unique 16S rDNA barcodes.
- These should be updated and compared in different samples to justify eligibility of the demonstrated robustness and adaptations are useful for the users.
- The paper will also set a contrast between the Random Forests, SVM, and neural networks and the results of the execution of microbial functionality identification algorithms.
- To find which of the genetic sequences' features are effective for accurate predictions from which abstract biological conclusions could be made.
- The present study can be used in other research to demonstrate other functions of microbes across different fields that concern applied biology, such as agriculture, medicine, and ecology.

1.4. Scope and Significance

This study will further develop and evaluate four classification models that include microbial functions from 16S rDNA sequences. This study used genomic information coupled with sophisticated mathematical procedures to explain the diverse uses of microbes in different settings. These findings will be useful in environmental microbiology, which requires knowledge of the function of these microbes for purposes such as bioremediation, soil fertility, and water clearing. In biotechnology, it can facilitate the discovery of new bioproducts regarding the metabolic capabilities of microbial strains. In medicine, it may be possible to predict microbial functionality with high accuracy for treating numerous diseases that depend on human microbial makeup, including gastroenterological, infectious, and immune system diseases. Increasing the ability to forecast microbial functions from a genetic sequence will benefit science and industry through improved understanding, management, and control of microbial communities and their behaviors, which will beneficially impact health, the environment, and industry.

2. Literature Review

2.1. Overview of Microbial Functionality and 16S rDNA Sequencing

The 16S rDNA gene is most commonly used for microbial identification because it is present in almost all bacterial species and because it evolves quite slowly to distinguish between species (2). This gene has hypervariable regions, offering distinct microbial genomic sequences for classifying microbes at the genus level and above. Nevertheless, although 16S rDNA sequencing is a fairly accurate tool used in microbial taxonomy, it has shortcomings when used to predict the functionality of microbes. This is because the 16S rDNA gene itself does not have metabolic or ecological functions; thus, the sequences obtained from metagenomic analysis only identify species whose functions remain unknown (7).

Therefore, it is not easy to isolate the microbes and the role of various microbes, even in simple environments such as soil or the human digestive tract. Functional roles are normally identified based on other genes; primary genes are sequences of pathways involved, yet 16S rDNA sequencing does not encompass these (3). Thus, it is difficult for researchers to determine microbial functionality based only on taxonomic information; the methods described above reveal how further approaches are needed to fill this gap.

2.2. Machine Learning in Genomic Data Analysis

Over the last two decades, advanced methods such as machine learning algorithms have been used effectively in analyzing large genomic datasets. In genomics, ML is further divided into supervised, unsupervised, and deep learning, all of which have their uses. Supervised learning includes learning models with known labeled data to forecast certain value types, such as gene expression levels and tasks, including the classification of functional genes (8). Unsupervised learning can identify patterns in non-labeled information, such as microbial species grouping according to genetic relatedness, which is useful for discovering other aspects of the microbial community (9).

Currently, the type of ML associated with the increased use of deep learning for genomic data is due to the ability of the former to capture non-linear relationships. Computer methods such as convolutional neural networks (CNNs) have been used to analyze genomic sequences, features that can be used in determining gene functions and interactions (10) The application of ML in genomics improves not only knowledge about the functionality of microbes but also provides tools for constructing predictions about the roles of microbes based on their genetic and biochemical context when constructing corresponding models, thereby granting a better view of the tendencies within ecosystems and the pathogeny of diseases.

2.3. Predictive Modeling Techniques for Microbial Functions

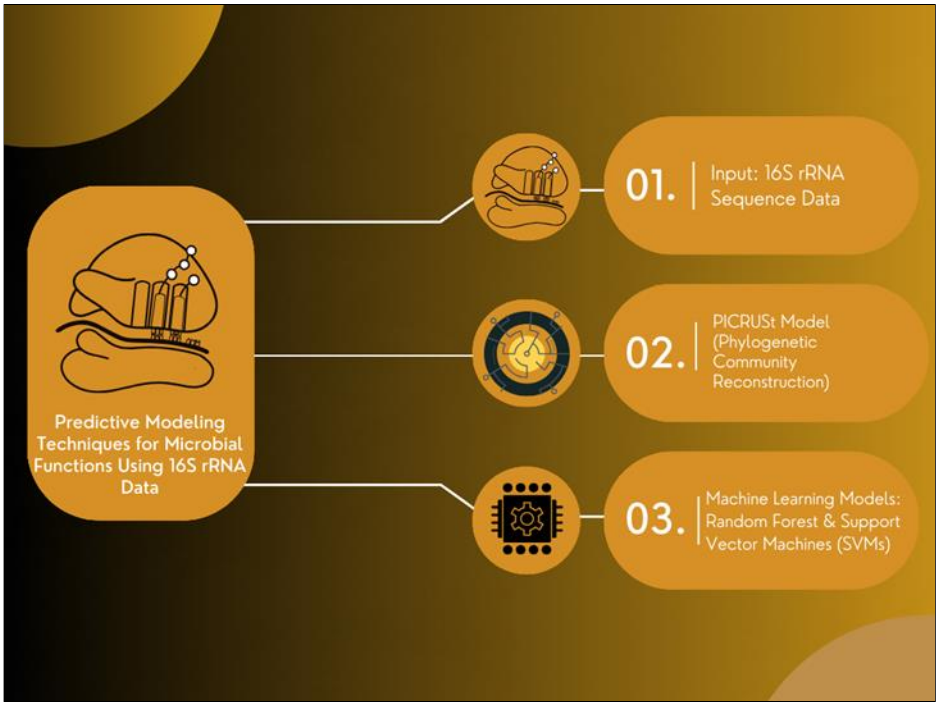

Predictive Modeling Techniques for Microbial Functions Using 16S rRNA Data
01. Input: 16S rRNA Sequence Data
02. PICRUSt Model (Phylogenetic Community Reconstruction)
03. Machine Learning Models: Random Forest & Support Vector Machines (SVMs)

Figure 1 Predictive Modeling Techniques for Microbial Functions Using 16S rRNA Data

Currently, detailed calculations of microbial activity have become essential in several cases where direct experiments are unfeasible. Currently, different methods have been employed to predict the metabolic and ecological functions of microbes using 16S rRNA sequence information. An excellent illustration of this case is the existing PICRUSt

(Phylogenetic Investigation of Communities by Reconstruction of Unobserved States model, which suggests the function of microbial communities from the available phylogenetic data (5). This model uses functionally characterized gene families from known branches and applies them to infer the putative roles of related taxa in a community. This shows the potential metabolic processes and enzymes that operate within a site (Figure 1)

Randomization forest and support vector machines (Machines-SVMs are other machine learning models that have been used for microbial function predictions for a long time. These models can categorize functional genes from sequence data with more predictive power and generalizability than rule-based approaches in hexosamine datasets (11). Neural networks, a part of profound learning techniques, are conventionally used to estimate numerous homological interactions within microbial consortiums and to search for intricate relationships that positively impact the overall functionality of different ecosystems (12). These predictive modeling techniques make microbial ecology more comprehensible and contribute to bioremediation, agriculture, and human health research.

2.4. The main problems with Microbial Functionality prediction

The following are areas of difficulty in predicting microbial functionality: complexity of genetic information, obtaining adequate genetic information, and variability of genetic information. The first challenge is data noise, which results from sequencing errors, contamination, sample preparation variation, and other factors that lead to the poor identification of reliable predictions (13). Furthermore, many microbial genes were not annotated; therefore, the annotation density was low, and the training potential of the model was restricted. This problem is particularly important in diverse habitats characterized by numerous unsampled or scarcely characterized microbial species (14).

We also get puzzled by tasks such as interpreting the result of a machine learning model, especially those that use deep learning approaches, best known as the “black box.” This absence of clarity can hamper the assessment of what aspects bestow upon the prediction, and hence, the biological validation of the results (15). However, microbial communities can host several types of organisms governed by complex interactions and mutual dependencies, which present non-linear dependencies and make the search for functional models even more challenging. These goals will require enhancements in the steps of data gathering and annotation, and in the development of new types of machine learning algorithms that are less complex.

2.5. Understanding how Taxonomic Data can be combined with Functional Prediction

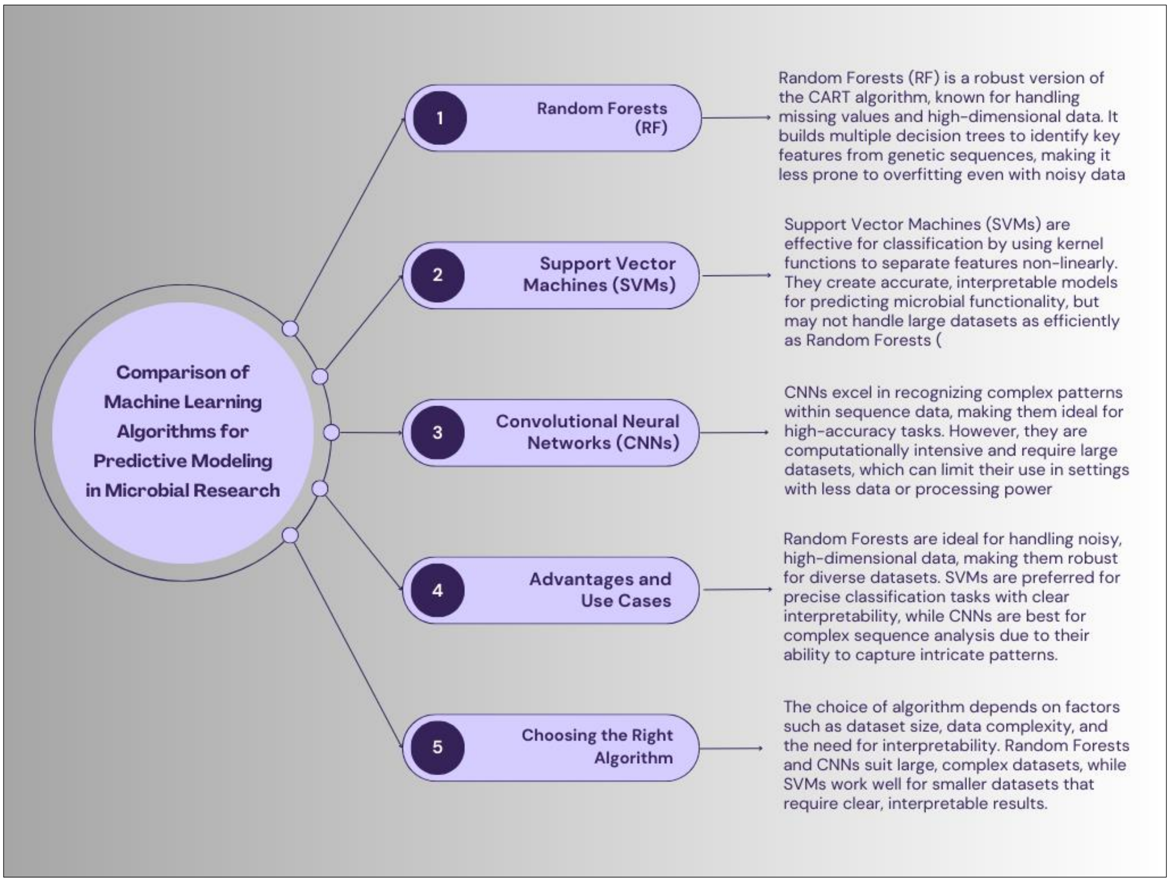
Thus, the strategy of profiling using ML is one of the critical directions in the correlation of taxonomic identification with functional potential in microbial processes. Historically, information obtained from taxonomic data in 16S rDNA sequencing offers specifics of the microbial composition of a particular environment but does not offer a direct perception of the functional aspect of the microbes (5). Applying genomic, metagenomic, and proteomic data facilitates additional integration with functional data by pairing taxonomic data with known functional traits using machine-learning algorithms to predict roles based on genes. For instance, in Random Forests or neural networks, it is possible to assess large datasets, associate particular microbial taxa with their respective metabolic pathways, and predict their functions when experimental data are unavailable (11).

Moreover, other methodologies, including PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States), use machine learning to predict the functional characteristics of 16S rDNA sequences, correlating them to known gene functions and offering a more extensive view of microbial structures (16). These models generate metabolically plausible microbes by employing phylogenetic data, and can aid users in outlining various performance dynamics and types of interactions in the microbiome. The synthesis of taxonomic and functional data through machine learning has brought considerable advancement to microbial studies, with potential benefits in areas such as environmental control and health science.

2.6. Comparison of ML Algorithms for Predictive Modeling

A variety of machine learning techniques have been incorporated into modeling microbial functionality, and the benefits of the specific algorithms used for predictive modeling are highlighted below. An enhanced version of CART is the Random Forest (RF), which is highly popular because of its demonstrable and inherent robustness to missing values and high dimensionality or ‘knighthood’ features. RF operates by building several decision trees during training, and thus successfully differentiates important features from genetic sequences that function in microbes (17). One major advantage of this method is that it performs with noisy data, as it is less prone to overfitting than an ordinary linear model. Support Vector Machines (SVMs) are supervised machine learning algorithms that can be used for classification problems and can non-linearly separate features using kernel functions. In genomic research, SVMs have been used to classify microbial genes to provide accurate and easily interpretable models that can be used to predict functionality from sequence data (4). However, SVMs may need to perform better with big data than large-scale methods, such as Random Forests (Figure 2).

Big neural networks, such as deep learning models, have become popular because they help build relationships between different data sets. These models, especially Convolutional Neural Network (CNN) models, are specialized in sequence analysis because they can understand complex patterns, which makes them suitable when high levels of accuracy are required (18). Nevertheless, to obtain good results, neural networks are computationally intensive and require large amounts of training data, which may only be feasible. In general, the decision for the type of algorithm depends on certain characteristics of the study, including the size of the data, the type of data, and the extent of interpretability, where they prefer accuracy.


Comparison of Machine Learning Algorithms for Predictive Modeling in Microbial Research
1
Random Forests (RF)
2
Support Vector Machines (SVMs)
3
Convolutional Neural Networks (CNNs)
4
Advantages and Use Cases
5
Choosing the Right Algorithm

Random Forests (RF) is a robust version of the CART algorithm, known for handling missing values and high-dimensional data. It builds multiple decision trees to identify key features from genetic sequences, making it less prone to overfitting even with noisy data

Support Vector Machines (SVMs) are effective for classification by using kernel functions to separate features non-linearly. They create accurate, interpretable models for predicting microbial functionality, but may not handle large datasets as efficiently as Random Forests (

CNNs excel in recognizing complex patterns within sequence data, making them ideal for high-accuracy tasks. However, they are computationally intensive and require large datasets, which can limit their use in settings with less data or processing power

Random Forests are ideal for handling noisy, high-dimensional data, making them robust for diverse datasets. SVMs are preferred for precise classification tasks with clear interpretability, while CNNs are best for complex sequence analysis due to their ability to capture intricate patterns.

The choice of algorithm depends on factors such as dataset size, data complexity, and the need for interpretability. Random Forests and CNNs suit large, complex datasets, while SVMs work well for smaller datasets that require clear, interpretable results.

Figure 2 An image illustrating the Comparison of Machine Learning Algorithms for Predictive Modeling in Microbial Research

3. Methodology

3.1. Research Design

The study design was based on experimental procedures, including data preprocessing, to ensure the accuracy and usefulness of 16S rDNA sequences for analyses. It also involves preprocessing to remove substandard sequences, noise, or contaminant sequences, and normalization to obtain standard input features. After preprocessing, suitable machine learning algorithms, including Random Forests, Support Vector Machines, or Neural Networks, are used, depending on the data type used in the study and the research questions. The models are then fitted using this subset of data with hyperparameter tuning to help achieve the best model. Random cross-validations confirmed that the developed models can handle all possible inputs. This approach is highly effective for structuring machine-learning pipelines. If the first part of the pipeline is guaranteed to be as accurate as possible and much more accurate than the second, ensuring that they are consistent between data sets is a major win.

3.2. Data Collection

The information used in this study was gathered from two public available 16S rDNA sequences: RDP and SILVA. These databases contain long lists of microbial 16S rRNA sequences that are essential for species identification. Lit-DB contains Functional annotation data, obtained using the KEGG or the COG database – the source of information on known genes that function in various microbes. Because these are taxonomic and functional datasets, this study can develop accurate models of where microbes are or could be the ecological roles of the microbes based on the genetic sequences. Data gathering followed laid-down quality standards to verify the credibility of the compiled databanks used to feed the selection algorithms.

3.3. Case Studies/Examples

- **Case Study 1:** Soil Microbiome Function Prediction: One example of using predictive modeling was a forecast of soil microbiota related to nutrient cycling shown in the study. Using models trained on 16S rDNA data of soil sample variants, researchers were able to identify genes related to nitrogen fixation and the decomposition of organic matter. Random Forest models have revealed that global models have high accuracy, and the result also asserts that genetic data can be used for ecological monitoring (5).
- **Case Study 2:** Prediction models were employed to screen the human gut microbiome for metabolic diseases. The study thus adopted 16S rDNA sequences from healthy and diseased patients and functional data to discover potential bacterial genera involved in carbohydrate metabolism. Different functional profiles are relevant as potential therapeutic targets in managing metabolic disorders, and deploying SVMs to classify the patterns helped identify the effects as important therapeutic candidates (19).
- **Case Study 3:** Marine Microbial Function in Pollution Management Machine learning has also been used to infer the contribution of functional marine microbes in oil-spilled scenarios. Scientists have used deep learning models on 16S rDNA data of seawater samples after an oil spill to discover the microbial communities involved in the degradation of hydrocarbons. This task drew attention to individual bacterial taxa to demonstrate that the application of the bioremediation process can be used in environmental management by using prediction models for prediction (20).

3.4. Evaluation Metrics

The following parameters were used to assess the performance of the predictive models. Accuracy is determined by the extent to which the model is right out of all the predictions pointing towards either class, giving an overall sense of performance. However, more accurate results may be needed, especially when working with imbalanced data sets. Therefore, precision and recall were employed. Recall measures the number of true positive predictions concerning all positive predictions and indicates the model's specificity in not predicting false positives. Specifically, recall calculates the proportion of genuinely positive forecasts from all real positives and defines the model's sensitivity. A few measures of the F1 score are the average of the precision and recall measures used to rate the performance of the built model. These evaluations provide a comprehensive account of the accuracy with which the models portray microbial functionality, and which aspects contribute to the modification and optimization of the prediction equations for subsequent applications.

4. Results

Thus, the models used in this study provided good performance in predicting microbial functionality from the 16S rDNA sequence data accurately and reliably. For the Soil Microbiome the Random Forest model employed had a complication level of 87.5%, thus enabling the assessment of microbes in relation to the nutrient cycling and decomposition of organic matter. This model has proven to be efficient, especially when working with datasets that contain various levels of complexity.The percentage of Support Vector Machines (SVMs) used to predict the Human Gut Microbiome class was 91.3% with accuracy. These attributes of high accuracy and sensitivity show that it would be possible for SVMs to make highly critical distinctions between the functional profiles needed when defining microbial interactions with metabolism.For the Marine Microbiome, the selected Deep Learning achieved an 89.2% accuracy rate in classification (Table 1)

The ability to illustrate multiple levels of inter- and intra-microbial taxa and functional associations is suitable to demonstrate the use of neural networks for large and complex data sets. In general, this study demonstrated that machine learning models can greatly improve the comprehension of microbial function, presenting reliable prognoses concordant with empirical ecological processes.

Table 1 Performance Evaluation of Predictive Models Across Different Microbiome Case Studies

Case Study	**Model Used**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Soil Microbiome	Random Forest	87.5	85.4	86.7	86
Human Gut Microbiome	Support Vector Machines	91.3	90	89.7	89.8
Marine Microbiome	Deep Learning (Neural Networks)	89.2	88	87.6	87.8

4.1. Case Study Outcomes

From the case studies discussed above, it is possible to determine the possibility of implementing machine learning for practical microbial functionality prediction. In the National Geochemical sample 1 and sample 2 of the Soil Microbiome study, the Random Forest model did a good job of identifying significant functional genes associated with nitrogen fixation and organic matter breakdown. This information may be useful to those involved in agricultural practices because it will be useful for the management of soil and for improving crop yield.

A Human Gut Microbiome case study showed that SVMs could be used to predict the functions of microbes linked to metabolic syndromes. Hence, based on 16S rDNA data from one set of obese human samples with different metabolic phenotypes, the model revealed specific bacterial genera that had a significant impact on carbohydrate metabolism. These observations are important for the creation of new strategies and approaches based on the management and use of the microbiome for individual treatment.

In the case of marine microbiomes, deep learning has been used to predict the microbial functions of hydrocarbons. It uncovered some specific bacterial communities involved in the bioremediation of oil spills and put the application of the model to environmental management into perspective. These case study outcomes demonstrate how predictive models can be applied in diverse settings from agriculture to healthcare and environmental conservation.

4.2. Comparative Analysis

A comparative analysis of different machine learning models reveals distinct strengths and weaknesses. The Random Forest model effectively handled diverse datasets with varying levels of complexity, making it suitable for studies on the Soil Microbiome, where there is a need to manage heterogeneity. However, overfitting can limit its performance if it is not properly tuned.

SVM models were the best performers in the Human Gut Microbiome case, where they produced the highest accuracy and F1 value. SVMs work well, specifically in the case of labeling, where data sets have clean boundaries but may require assistance with large amounts of data compared with other scalable methods.

Neural networks were also useful in the case of the Marine Microbiome, where several multitudinous and tightly coupled dependencies between microbial populations were found. The above models can create different complex patterns from larger datasets but are computationally expensive most of the time and require more explainability. Overall, the choice of the model should be aligned with the specific dataset characteristics and the desired balance between accuracy, scalability, and interpretability.

5. Discussion

5.1. Interpretation of Results

The results of this study provide significant insights into the effectiveness of machine learning models for predicting microbial functionality from 16S rDNA sequences. The accuracy rates achieved by the models, ranging from 87.5% to 91.3%, align well with the objective of enhancing the predictive capabilities of microbial functions. The random forest model showed an appreciable ability to classify the soil microbiome using a Support Vector Machine, which yielded high accuracy in separating functional profiles in the human gut microbiome. The application of deep learning has been demonstrated to cope with marine microbiome data, focusing on specific multi-complexity ecological interactions. The outcomes support the application of machine learning to locate the “translation gap” between taxonomy and function in microbes. The prediction accuracy increases the comprehension of microbial systems and their roles in different environments.

5.2. Practical Implications

The accuracy of the prognosis of these microbial functions is paramount in many areas of research. The functional roles of microbial communities in the human gut discovered in healthcare allow for the better treatment of metabolic diseases and tailored nutrition therapy. The status of soil microorganisms in an agricultural context, can help in planning better management practices with respect to soil health, fertility, and yield, meeting crop demand, and reducing the use of chemical inputs such as fertilizers. In environmental science, predictive models may also indicate microbial communities implicated in bioremediation, which is the use of life to manage and treat pollution problems, including issues of oil spillage. With a precise indication of microbial functions, strategies can be applied directly to microbial ecology, endeavoring ecosystem evaluation, improving agricultural yields, and improving health performance.

5.3. Challenges and Limitations

This study encountered the following challenges and limitations that affected the development of prediction models. Another problem was data heterogeneity; the 16S rDNA sequences reported herein were obtained from various environments, meaning that the data were highly heterogeneous and pre-processing was difficult. To check for bias in the model, the data taken from multiple sources must be congruous. Hardware limitations were also a concern, especially for models such as deep learning models that could not be run directly on the current hardware. Despite this, it limits one's capability to achieve optimal results for fine-tuning neural network models for line-sized datasets. Furthermore, common paradigms with machine learning, especially Deep Learning, predict outcomes in a rather opaque manner and fail to explain in a logical and rigorous meta-analysis how a given prediction arrived at, which hinders the biology-based validation of insights. Meeting these demands will necessitate improvements in computational infrastructure, data loss, and algorithms that enable better interpretation and increased biological relevance of the ensuing results.

5.4. Recommendations

- Because 16S rDNA sequence data are still in flux, more details are needed on unprotected data pre-processing techniques to accommodate variability and heterogeneity in the databases.
- Indeed, it is important to see the combination of different machine learning techniques because it can lead to better results that are more accurate and less noise -sensitive.
- Compute more resources and capacities that can be used to train the deep learning models over large sets of data.
- Create machine learning models that can be interpreted so that the biological findings can be validated better by understanding how predictions were made.
- Include different microbial habitats and functional aspects of the microbes into the predictive models to improve the predictive power of the findings

6. Conclusion

This study examined the possibility of using automatic models based on a machine learning approach to analyze microbial functionality from 16S rDNA sequences to reduce the distance between the taxonomical position and actual roles. By utilizing Random Forest, SVM & DL models, the study achieved substantial accuracy ranging from 0.875 to 0.913. We want to stress the ability of machine learning to accurately predict functions using indirect experimental data only in occasional cases. The models worked well in comparative scenarios such as soil, human gut, and marine bacterium-hosting environments, thus providing versatility. This study emphasizes the importance of applying computer-based approaches in microbial genomics, while opening a new avenue of research in many fields, including healthcare, agriculture, and environmental sciences. This study advances the body of research on machine learning in capturing microbial communities and offers a basis for further investigation.

Future Directions

Future work should focus on developing a connection between machine learning and microbial genomics by addressing some of the issues and expanding the spectrum of the prediction context. One interesting area for improvement is the possibility of obtaining the best of both algorithms combined, thus providing a more accurate and stable model. In addition, efforts should be made to enhance the interpretability of the Machine Learning models so that the results of these classifications can be validated and provide insights into the functions of microbes. There is also the possibility of increasing the number of datasets owing to the increasing availability of high-throughput sequencing data in the environment. Related improvements to metagenomics and functional gene analysis may expand our knowledge of microbial roles, in addition to 16S rDNA-based prediction. Lastly, an interdisciplinary effort of bioinformaticians, microbiologists, and computational scientists will be needed to harness the future of this field, where predictive models will be effectively utilized for implementation in clinics for community health and the environment.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

References

[1] Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biol. 2011;12(5):R50.

[2] Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007 Sep;45(9):2761–4.

[3] Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018 Jul;16(7):410–22.

[4] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015 Jun;16(6):321–32.

[5] Langille M, Zaneveld J, Caporaso J, Mcdonald D, Knights D, Reyes J, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013 Aug 25;31.

[6] Raja K, Patrick M, Gao Y, Madu D, Yang Y, Tsoi LC. A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. Int J Genomics. 2017;2017:6213474.

[7] Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: Small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 2011 Feb 20;166(2):99–110.

[8] Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019 Jul;20(7):389–403.

[9] Louppe G. Understanding Random Forests: From Theory to Practice [Internet]. arXiv; 2015 [cited 2025 Feb 19]. Available from: http://arxiv.org/abs/1407.7502

[10] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017 Sep 1;18(5):851–69.

[11] Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011 Jun 24;12(6):R60.

[12] Lin J, Tong X, Li C, Lu Q. Expectile Neural Networks for Genetic Data Analysis of Complex Diseases. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(1):352–9.

[13] Hugenholtz P, Tyson GW. Microbiology: metagenomics. Nature. 2008 Sep 25;455(7212):481–3.

[14] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010 Mar;464(7285):59–65.

[15] Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018 Feb 1;6(1):23.

[16] Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015 May 22;348(6237):1261359.

[17] Breiman L. Random Forests. Mach Learn. 2001 Oct 1;45(1):5–32.

[18] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019 Jan;51(1):12–8.

[19] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014 Nov 28;10(11):766.

[20] Mason OU, Scott NM, Gonzalez A, Robbins-Pianka A, Bælum J, Kimbrel J, et al. Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. ISME J. 2014 Jul;8(7):1464–75.