World Journal of Biology Pharmacy and Health Sciences

eISSN: 2582-5542

Cross Ref DOI: 10.30574/wjbphs

Journal homepage: https://wjbphs.com/

(Review Article)

Check for updates

Deep learning for improved microbial community profiling through 16S rDNA Data

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), 532-544

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.2025.21.2.0227

Abstract

Microbial community characterization is important, especially for the identification of microbial species and their relationships within various environments in support of clinical, agricultural, and ecological applications. Although morphological and culture-independent molecular techniques using 16S rDNA gene sequencing have been extensively employed, they are less efficient, particularly for determining the real presence of minor communities in a sample. In this study, deep learning models, CNN and RNN, were applied to improve the classification and characterization of microorganisms describing the 16S rDNA sequence data. From the current experiment, it is postulated that the incorporation of both CNNs for precise pattern recognition and RNNs for the latent dependencies on detection enhance accuracy, especially for species with a lower probability of detection. Public datasets were used to evaluate the models. The performance of the proposed models was also assessed in relation to the basic machine learning methods. According to the study, classification accuracy could be enhanced by using deep learning-based solutions to overcome existing limitations in the description of microbial diversity. These advancements have enormous potential for many disciplines, ranging from disease diagnosis to soil and water examinations, based on improving the ability to analyze ability to analyze the microbial community.

Keywords: Microbial Profiling; 16S rDNA Sequencing; Genomic Sequences; Microbiome Analysis; RNNs; CNNs

1. Introduction

Microbial profiling, involving culture techniques, has been applied to elucidate and investigate the effects of microbial community processes and interactions in various ecosystems. It is helpful in different aspects of life, for example, in medicine, agriculture, other related fields, and environmental science (1). The human microbiome participates in disease states and health status concerning diseases such as obesity, diabetes, and gastrointestinal diseases (2). Agricultural microbiota determine plant health and yield (3). Microbial profiling in environmental science is useful for evaluating ecosystems and bioremediation, in which microbes may reduce pollutants and repair environmental dilemmas (4).

Previous approaches used to report microbial community composition involved cultivation; however, most microbes are not cultivable under culture-based conditions (5). For this reason, molecular methods, such as 16S rDNA sequencing, have been used to identify bacteria directly without the need for culturing (6). A frequent approach within this paradigm entails the identification of Operational Taxonomic Units [OTUs], which group sequences based on some semblance of similarity cutoffs, often 97% for species-level resolution (7).

The OTU-based method has developed microbial ecology and, as a benefit, has some disadvantages. These arbitrary similarity thresholds may result in higher or lower estimates of microbial diversity estimates (7). However, errors

∗ Corresponding author: Abhaykumar Dalsaniya Orchid Id 0009-0003-7309-3455.

introduced due to sequencing and the appearance of chimeric sequences can hinder recovery of performance and mortality profiles (6). These issues are especially compounded when identifying species that are important for ecosystem functionality or predicting the occurrence of a disease, but are difficult to find (1).

This shows that traditional methods cannot provide an accurate microbial profile to warrant improvements in existing processes. Originating from several analysis applications, Schloss and Handelsman (2005) explained that with the recent advancement of high- throughput sequencing, the amount of data produced has increased significantly, demanding analysis and development of relevant patterns (5). Thus, it is necessary to develop microbiome research topics and apply microbiome-related techniques in various sciences to address these issues.

1.1. Overview

The application of 16S rDNA sequencing has revolutionized microbial ecology by providing a means of characterizing and comparing bacteria from conserved genetic sites (8). The 16S rRNA gene is quite diverse and has regions of high and low variability that would allow species-level differentiation and higher classification levels, as outlined by (9). This method has made it easier to understand microbial distributions at sites as diverse as the gut microbiota in extremophilic environments and their functions and interactions (10).

Molecular methods for species identification are now popularized as powerful, especially the 16S rDNA sequencing technique, but there is a challenge associated with analyzing the large data set obtained. These sequencing data could be too large and complex for traditional bioinformatic methods to analyze, causing them to overlook and conceal patterns or rare taxa (5). This limitation has created the need for more complex data analysis approaches, such as deep learning (11).

Neural networks are in the machine learning category and aim at deep architectures formed with more than one layer to represent data (11). Convolutional Neural Networks (CNNs) are best suited for processing data that are naturally arranged in a grid-like format, and this includes images and, in the context of this discussion, genomic sequences that may have been converted to numerical types. CNNs can update weights in the manner of back propagation and, therefore, can learn spatial hierarchical feature maps with professionals from different building blocks, including the convolution layer, pooling layer, and fully connected layer (12).

Recurrent Neural Networks (RNNs). Extended versions of RNNs comprise LSTMs, well fitting for analyzing nucleotide sequences because patterns in the order of sequence elements are crucial (13). RNNs continue to have some form of state that holds details about what has been analyzed to enable the analysis of the dependency of sequences on genes (14).

The analysis of CNNs and RNNs makes it possible to approach the existing complexity of bioinformatics and model the existing relationships of data, which may enhance the classification and profiling of microbial communities (15). These deep learning algorithms can work on big data and may recognize characteristics associated with low-frequency or high-density species that can be missed by standard methods (16). Thus, using such analytical measures, scientists strive to expand the empirical prospects regarding microbial community descriptors and improve insights into microbial and biochemical traits and connections.

1.2. Problem Statement

Microbial community profiling has two challenges: they became involved in identifying rare species and describing various microbial communities. Biases in the original OTU-based analyses have high levels of data complexity, sequencing errors, and problems with differentiation between similar species. Such issues can result in erroneous descriptions of community composition, omitting low-biomass organisms with vital roles in relationship networks or medical significance. Furthermore, even if the sequencing data are large, conventional computation methods may need to process this amount of data efficiently, and the rapid growth of sequencing data may be problematic. Unfortunately, this problem calls for a solution that Ettinger deems possible with deep learning, in which the technology employs superior neural networks capable of identifying faint patterns on complex data. This will also improve the ability of microbial profiling and may increase its sensitivity for detecting low-abundance or weakly expressed species.

Objectives

- To train and constitute CNNs and RNNs for taxonomy-based classification and heterogeneity description of microbial congregation derived from 16S rDNA sequencing data.
- To compare the accuracy and efficiency of these models in dealing with large microbial datasets.
- To evaluate the efficiency of the developed deep learning-based approaches compared with basic bioinformatics tools and algorithms working with species identification, emphasizing the detection of low-frequency species.
- To assess the suitability of the proposed models for large-scale sequencing data and their computational complexity.
- To derive specific areas for advanced microbial identification for health, agriculture, and ecological use.

1.3. Scope and Significance

In essence, this study is specifically concerned with enhanced Microbial Community Profiling utilizing 16S rDNA sequencing data by developing and implementing deep learning procedures. The scope includes choosing open-source 16S rDNA datasets and preparing them for analysis with the CNN and RNN models. This study will assess the increase in accuracy when using deep learning approaches by comparing them with the traditional method of species detection for low-abundance species. In addition, qualitative parameters, such as precision, recall, and F1 score, were used to evaluate the models.

Improving microbial profiling, as discussed in this study, has critical implications in various fields. In the field of health care, the characterizations are likely to enhance the expanded understanding of the microbial environments, hence new observations on the microbiome in disease, diagnosis, and possibly treatment. Similarly, there are a number of benefits that exact profiling offers in agriculture concerning soil management as well as production. Moreover, enhanced accidental microbial identification is essential as it helps to evaluate pollution and index the progress of ecological recolonization. Thus, to avoid these limitations, this study aims to enhance the reliability of microbial community analysis.

2. Literature review

2.1. Deep Learning in Bioinformatics

The application of deep learning for bioinformatics has been enhanced, as it helps analyze large biological data, especially genomics. The conventional workbench of the machine learning models was underpinned by large amounts of feature engineering, where the choice of features in the analysis had to be made by an expert in the field. In contrast, deep learning automatically identifies these representations by learning hierarchical representations from data and is well suited for large data set applications such as genomics (15).

Deep-learning models have been used for gene expression prediction, variant calling, and sequence-based protein structure prediction. One of the most promising applications of modern algorithmic methods is the identification of motifs and regulatory elements in DNA sequences. For example, CNNs have been applied to identify patterns and motifs in DNA sequences, enhancing the accuracy of DNA-protein interactions (12). First, CNNs can learn the features required for classification without requiring pre-processing of short-length nucleotide sequences.

The next field is the application of recurrent neural networks (RNN) to data richness or sequential data. Specifically, RNNs equipped with long short-term memory (LSTM) are used to analyze sequential data, and RNA sequences and the prediction of secondary structures can be examples of such tasks. Understanding this sequence and its dependency have paved the way for identifying functional RNA regions and capturing transcriptional processes (17).

In addition, recent studies using deep learning models in metagenomics have demonstrated their ability to classify and annotate microbial sequences extracted directly from samples. The versatility of using these models to analyze raw sequencing data, where no extensive feature extraction is required, has also added to the benefits of microbial community analysis to enhance the rate at which rare species are detected as compared to other traditional methods (15). Overall, integrating deep learning has revolutionized bioinformatics by offering better and improved approaches for analyzing genomics, with the prospects of its applicability increasing as these tools develop.

2.2. 16S rDNA Sequencing and Its Importance

16S rDNA sequencing is one of the most common methods for bacterial identification and classification based on the 16S RNA gene. This gene is present in all bacteria and has both conserved and hypervariable domains, making it well suited for taxonomic analysis (8). The conserved sequences for Compositae PCR allow universal primers to match across various species of bacteria, and the hypervariable sequences permit distinction of organisms.

Methods such as 16S rDNA sequencing are performed by obtaining DNA from a sample, creating copies of the 16S rRNA gene, and sequencing fragmented copies. The resulting sequences were again subjected to existing databases of known 16S sequences to determine the nature of various microorganisms in the sample (Figure 1). They have been used in ecological investigations to assess microbial distribution in diverse habitats, such as soil, water, and the gastrointestinal tract of human beings (9). It includes the identification of many new bacterial genera and biota and elucidation of the microbial ecosystem.

Another advantage of 16S rDNA sequencing is that the microorganisms tested can be obtained from non-culturable samples. In contrast, many organisms that account for microbial richness must be cultivated in laboratory settings. This technique has extended the field of microbial ecology and provided scholars and researchers with a way to examine myriad systems and determine the role of specific microbes in these systems (10).

However, 16S rDNA sequencing has some disadvantages. They have limitations, such as the method may need to provide more oscillation to identify closely related species, which requires computer assistance. Biases can also occur during DNA amplification. Moreover, unknown or less-described sequences may not be correctly recognized because the results depend on the reference databases. However, present-day microbial studies heavily depend only on 16S rDNA sequencing, which is necessary for detecting microbial presence and ecology in various systems.

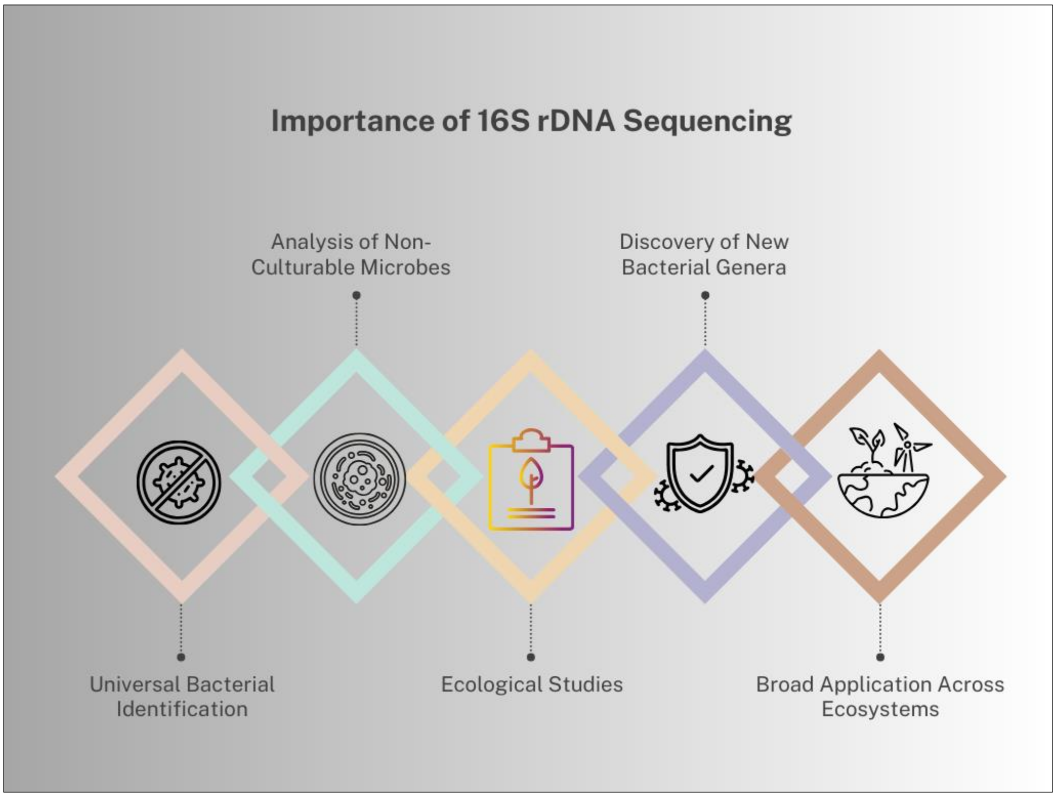

Importance of 16S rDNA Sequencing
Analysis of Non-Culturable Microbes
Discovery of New Bacterial Genera
Universal Bacterial Identification
Ecological Studies
Broad Application Across Ecosystems

Figure 1 An image illustrating Broad Application Across Ecosystems

2.3. Traditional Approaches to Microbial Profiling

Previous microbial community identification approaches have mainly used microbial culturing techniques, in which separate microbial species are grown under artificial conditions. These methods have been useful in elucidating the physiological characteristics of bacteria, but their liabilities are that most microorganisms are difficult to cultivate under standard laboratory conditions. This results in a reduced representation of microbial diversity (18).

Another traditional method is biochemical analysis, in which microbial communities are characterized by their functional activity. Although these methods can offer functional information, they provide broad information and are affected by the environment; therefore, accurate community profiles cannot be easily established (19). Recently, techniques such as 16S rDNA sequencing have been adopted because they can avoid culture steps and instead analyze the genetic information that may be contained in a given sample (5).

16S sequencing data have mostly been analyzed using Operational Taxonomic Units (OTUs). OTU cluster sequences were based on similarity criteria with a default percentage similarity greater than 97%, which is considered equivalent to species-level identification. However, this method has shortcomings, such as using a fixed cut-off point, and may form an OTU from either two distinct genetic species or split a single species into many. These problems have led to new solutions, such as AMR gene swabbing, proposed plans known as Amplicon Sequence Variants (ASVs) to increase taxonomic resolvability, where every single sequence is considered a different taxon.

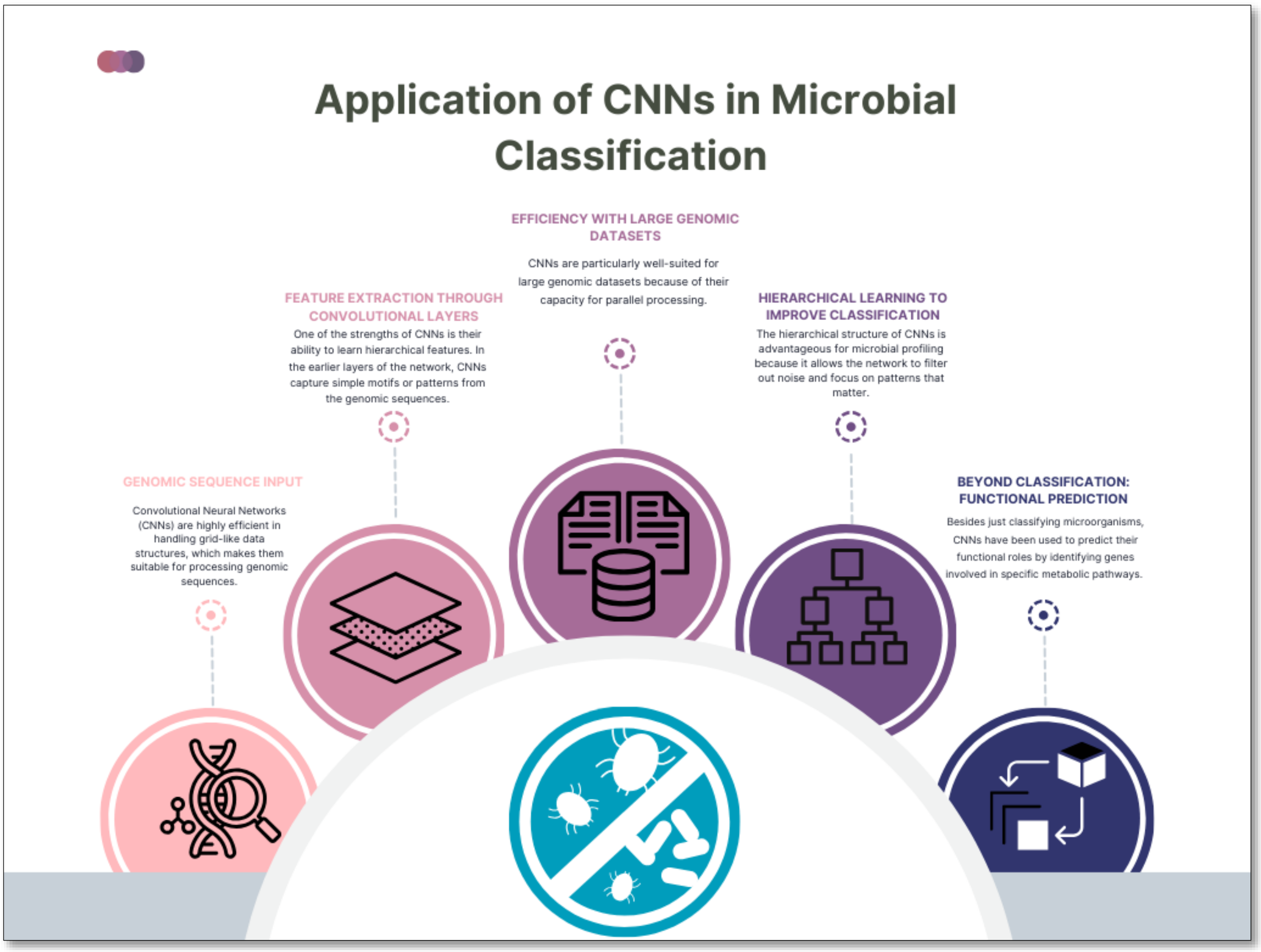
As previously presented, the basic profiling techniques have provided the foundation for describing microbial processes. However, such strategies have been augmented or replaced by modern high-throughput technologies. The transition to genomic methods and Deep Learning complements the development of more precise, configurable, and self-sustaining approaches to microbial ecology.

2.4. Application of CNNs in Microbial Classification

Convolutional Neural Networks (CNNs) have recently been adopted for microbial classification because of their efficiency in predicting grid-like structural data, such as genomic sequences. CNNs employ convolutions to discern filaments throughout the input data because a filtering strategy that shifts over the input data recognizes features that characterize different elements of microbial taxa. This makes CNNs suitable for classifying genomic sequences without much pre-processing, as noted (12). The other benefit of using CNNs in microbial profiling is learning hierarchical features. More specific and simple motifs can be captured in earlier layers of the network, and the deeper layers of the network enable the capture of more complex motifs, thereby providing a more accurate sequence analysis of genes (20). This structure of feature hierarchies is useful when working with noisy data because CNNs can train on patterns that matter in the dataset (Figure 2).

Moreover, CNNs are suitable for large datasets because they can be implemented across parallel processing elements. This efficiency is important when working with large genomic datasets, which has become difficult for traditional methodologies. Some experiments have shown that CNN-based classification methods are even better than conventional methods for DNA sequence classification, especially at the species level (21).

In addition to classification problems, CNNs have been used to predict the functional roles of microorganisms by identifying genes involved in certain metabolic pathways. This extends the usefulness of CNNs in microbial ecology and enhances the evaluation of microbial function. In conclusion, because of the high extensibility and strong capacity of CNNs, new directions can be developed to explore microbial genomics, and the accuracy of microbial classification can be significantly improved.



Application of CNNs in Microbial Classification

EFFICIENCY WITH LARGE GENOMIC DATASETS

CNNs are particularly well-suited for large genomic datasets because of their capacity for parallel processing.

FEATURE EXTRACTION THROUGH CONVOLUTIONAL LAYERS

One of the strengths of CNNs is their ability to learn hierarchical features. In the earlier layers of the network, CNNs capture simple motifs or patterns from the genomic sequences.

HIERARCHICAL LEARNING TO IMPROVE CLASSIFICATION

The hierarchical structure of CNNs is advantageous for microbial profiling because it allows the network to filter out noise and focus on patterns that matter.

GENOMIC SEQUENCE INPUT

Convolutional Neural Networks (CNNs) are highly efficient in handling grid-like data structures, which makes them suitable for processing genomic sequences.

BEYOND CLASSIFICATION: FUNCTIONAL PREDICTION

Besides just classifying microorganisms, CNNs have been used to predict their functional roles by identifying genes involved in specific metabolic pathways.

Figure 2 An image illustrating Application of CNNs in Microbial Classification

2.5. RNNs for Sequence Data Analysis

Recurrent Neural Networks (RNNs) are specific architectures for deep learning that are implicit in handling series data and are endowed with a method of preserving information about prior input data. This ability to sequentially consume information makes RNNs even more useful in other applications where the time element is involved, such as time-series prediction, speech-to-text conversion, and analyzing genomic sequences (13). In contrast to feedforward neural networks, which take inputs separately, RNN establish a connection between each step within a sequence and its significance.

RNNs are applied in bioinformatics, for instance, to examine nucleotide sequences, where the sequence of the nucleotides might be relevant to recognizing specific genes, control sections, or other, perhaps even, mutations that could have functional relevance. LSTM and GRU, the newest types of RNN, overcome the vanishing/exploding gradient issue characteristic of standard RNNs. This enables them to retain information over long sequences, which is useful for modeling dependencies that are important for accurate sequence analysis (14).

For instance, LSTMs have been successfully used to predict the secondary structure of RNA with temporal structures in the sequence. These models help better understand how RNA functions by teaching which sequences are expected to fold into such or that structure (16). In addition, RNNs can be used to identify the binding sites of DNA sequences, which, when analyzed, reveal where proteins may bind to, which is essential for defining genes and their functions.

A major advantage of RNNs in analyzing biological sequences is that sequences of different lengths can be accommodated in the networks, unlike other neural networks where inputs have to be the same size. This is especially valuable in genomics, where the sizes of the sequences under analysis can be very different among species or even within the same genome region. Therefore, RNNs are crucial for processing biological data signals as a suitable complex model for analysis.

2.6. Difference Between Deep Learning and Traditional Machine Learning

It is possible to explain the benefits of deep learning versus standard machine learning strategies for microbial community profiling as follows: In traditional machine learning scenarios, a large-amplitude focus is bestowed on the feature extraction process, in which the expert gains special attribute vectors from the data and feeds them to the model. This process can be lengthy and may exclude difficult patterns that are inherent to the data inputs. Traditional approaches can only encode relatively simple relationships, but deep learning models learn the features of the raw data directly, making them capable of detecting more complex patterns (22).

An important advantage of deep learning is the high speed and efficiency of its learning algorithms when working with large, high-dimensional datasets. Regarding large-scale metagenomic projects, where sequencing technologies can produce large amounts of data, CNNs and RNNs are separated from learning the patterns inside these huge datasets with few pre-processing steps (21). This capability is especially useful in developing better models of microbial communities and their interactions, because the rare species of the groups might be unseen when using other models because of data complexity and noise.

The other is the flexibility of deep learning models, which allows the scaling problem to be more creatively and flexibly approached. In contrast to the majority of conventional machine learning approaches that may need to be fine-tuned as more data types come in, or for that matter, when higher volumes of data are incorporated, deep learning models can be easily scaled up by simply adding more layers or training on larger sets of data. This makes deep learning more resilient to changes in the complexity of microbial genomic datasets that accumulate in large numbers and sizes (15).

Finally, deep learning models can process unorganized information, which is common in biological research. Irrespective of the origin of the data, such as raw DNA sequences, imaging, or even clinical records, deep learning models can extract knowledge from different data formats and aggregate expertise across other formats. It has the advantage of generic applicability to constructive microbial community profiling and generates an improved understanding and prognostication of their constitution and process.

2.7. Issues to Consider in Deep Learning for Microbial Community Analysis

However, as discussed in the previous paragraphs, there are several challenges that must be solved before deep learning technologies can be successfully applied as microbial community analysis tools.

Despite the promising application of deep learning in microbial community analysis, several problems need to be solved. One of the biggest issues is the need for data sets of a sufficient size. This type of model greatly benefits from big data, because deep learning models are dependent on training samples. Nonetheless, collecting large, high-quality microbial datasets can be challenging, especially when focusing on a set of rare species or a given habitat (23). This can be because of overfitting when the program adopts certain characteristics of the training data and still cannot perform well in other datasets.

Another problem is the computational power required to train deep learning models. However, deep learning, particularly CNNs and RNNs, can be computationally heavy, often requiring a GPU package or renting for the cloud to manage big data. For many researchers and institutions, these computational demands prevent the use of deep-learning-based approaches for microbial profiling (24).

Furthermore, deep learning is interpreted as one of the most important subspace learning techniques, and the question of interpretability remains. However, this is dissimilar to typical structure learning, where the decisions as to which of the speaking features to choose can be traced to the actual features in question, as opposed to deep learning models commonly referred to as "black boxes." This lack of transparency is a concern in biological applications, and in addition to achieving accurate results, the mechanism of how the results were obtained must also be an area of expertise (25). Work is still being carried out to make these models more explainable, for example, by paying attention to where the input data were most useful in arriving at a decision.

Most of these challenges are vital for determining the future of deep learning in microbial community profiling. Issues that must be resolved include better algorithms for data augmentation, training of deep models, and developing models that are easy to explain if they are to be successful in the future.

3. Methodology

3.1. Research Design

The experimental design included the creation and refinement of CNN and RNN models for 16S rDNA sequencing data with microbial community classification and profiling. The models were trained on large datasets, emphasizing which patterns within the sequences help discriminate between the various microbial taxa. These include data cleaning, where sequences need to be aligned, removal of noise, filtering, and normalization to align the data with the acceptability of the model. In addition, several data augmentation methods are used to expand the types of training samples that the model will be trained on. The design also incorporates multiple training iterations to help obtain the correct parameters in the models and refine the classification accuracy.

3.2. Data Collection

The 16S rDNA data will be obtained from public domain databases such as Greengene, SILVA, RDP, and the Ribosomal Database Project. These databases contain the tested and accumulated 16S rRNA gene sequences that are used for microbial community analysis. Moreover, new data obtained from experimental settings, soils, water, and clinical samples will enable broadening of the study. The data sources also contain a variety of settings to assess the flexibility of the models for different microbial habitats. The gathered sequences are processed to eliminate all noise and mistakes, so that only the highest quality data will be used to construct and evaluate the model.

3.3. Case Studies/Examples

Microbial profiling has attracted deep learning utilization, and previous studies have shown a profound ability to enhance classification and sensitivity. An example of this is the classification of metagenomic sequences. CNNs were used to strengthen the identification of microorganisms in the human gut microbiome. This study also aimed to demonstrate how CNNs can learn from large sets of raw sequences and discover major microbial taxa that analysts may not have picked otherwise. Using the CNN model, with its training based on publicly available datasets from resources such as the Human Microbiome Project, Min et al. demonstrated that deep learning outperforms traditional machine learning in classifying bacterial families and genera.

Another example is based on employing Recurrent Neural Networks (RNNs), including LSTM networks with sequence analysis, to identify functional RNA elements. The RNN was trained on known sequences of functional RNA and allowed for the detection of similar functional elements within metagenomic sequences. This approach makes it easy for the model to scan the entire sequence to accurately capture the long-range dependencies essential in predicting functional sites. RNNs employ sequence context and structure as crucial factors for various tasks, providing an extra advantage because order is critical (16).

These were accepted along with an illustration of a third type of model that incorporates both CNN and RNN. At the same time, CNN was used to extract features from short sequences of a predetermined fixed size; RNN then analyzed the patterns for all sequences. This model was applied to the classification of sequences derived from diverse origins, including Soil, Marine water, which helped improve the detection of lowly represented microbial taxa. Consequently, by using CNN and RNN combined in the identification of cervical lesions, the overall reported accuracies and sensibilities were higher than when solely using CNN or RNN (21). These cases demonstrate how deep learning transcends existing microbial profiling issues and provides advanced solutions, unlike the traditional methods.

3.4. Evaluation Metrics

The performance of the deep learning models was assessed using several metrics that estimate the performance of the model's classifier. Accuracy is the percentage of correctly labeled sequences from the total number of sequences used in the prediction. Although simple, it indicates how well a model performs. However, accuracy alone can be deceptive, and it possesses high accuracy in numerous strikes and extensive federation imbalances. Hence, we require additional measures, including precision, recall, and F1-score.

Accuracy is the ratio of the number of true positives added to the number of true negatives to the total number of instances predicted by the model. Sensitivity or recall determines the true rate of actual positive data to measure whether the model captures every possible example of a particular class. The F-measure is also calculated as the harmonic mean of precision and recall for the purpose of this study and is particularly helpful for models that need to minimize the false negative rate for low-incidence species. These metrics were calculated from different runs of the model to evaluate the reliability and efficiency of the microbial profiling tasks.

4. Results

4.1. Data Presentation

The performance metrics of different machine learning models for microbial profiling, including Traditional ML, CNN, RNN, and Hybrid CNN-RNN, are analyzed based on accuracy, precision, recall, F1-score, and detection of rare species. The plotted results indicate a clear performance hierarchy among these models. Traditional ML exhibits the lowest performance across all metrics, with accuracy, precision, recall, and F1-score around 78-80%, while detection of rare species is significantly low (~65%). CNN demonstrates a notable improvement in all performance metrics, reaching an accuracy of approximately 90%, with precision, recall, and F1-score following a similar upward trend, and rare species detection improving to nearly 78%. RNN shows slightly lower performance compared to CNN in terms of accuracy, precision, recall, and F1-score but remains significantly better than Traditional ML, with rare species detection slightly lower (~75%) than CNN. Hybrid CNN-RNN achieves the highest overall performance, with accuracy exceeding 91%, precision, recall, and F1-score closely aligned at around 90%, and rare species detection reaching approximately 80% (Figure 3).

These results indicate that CNN outperforms Traditional ML, demonstrating a significant boost in accuracy and recall, which suggests its efficiency in feature extraction from microbial data. RNN, though effective in learning sequential dependencies, performs slightly lower than CNN, implying that microbial data may not require extensive temporal modeling. The Hybrid CNN-RNN surpasses all other models, achieving the best results in all metrics, particularly in detecting rare species, which is crucial for microbial profiling. The observed trends suggest that deep learning models, particularly CNN and Hybrid CNN-RNN architectures, significantly enhance microbial profiling accuracy compared to traditional machine learning methods. The high performance of CNN highlights its strong ability to capture spatial relationships in microbial data, which is crucial for accurate classification and profiling. While RNNs are effective in learning sequential dependencies, their slightly lower performance than CNN suggests that microbial data may not require extensive temporal modeling. By combining CNN's spatial feature extraction with RNN's sequential processing, the hybrid model achieves the best results, particularly in detecting rare species, which is a crucial aspect of microbial profiling.

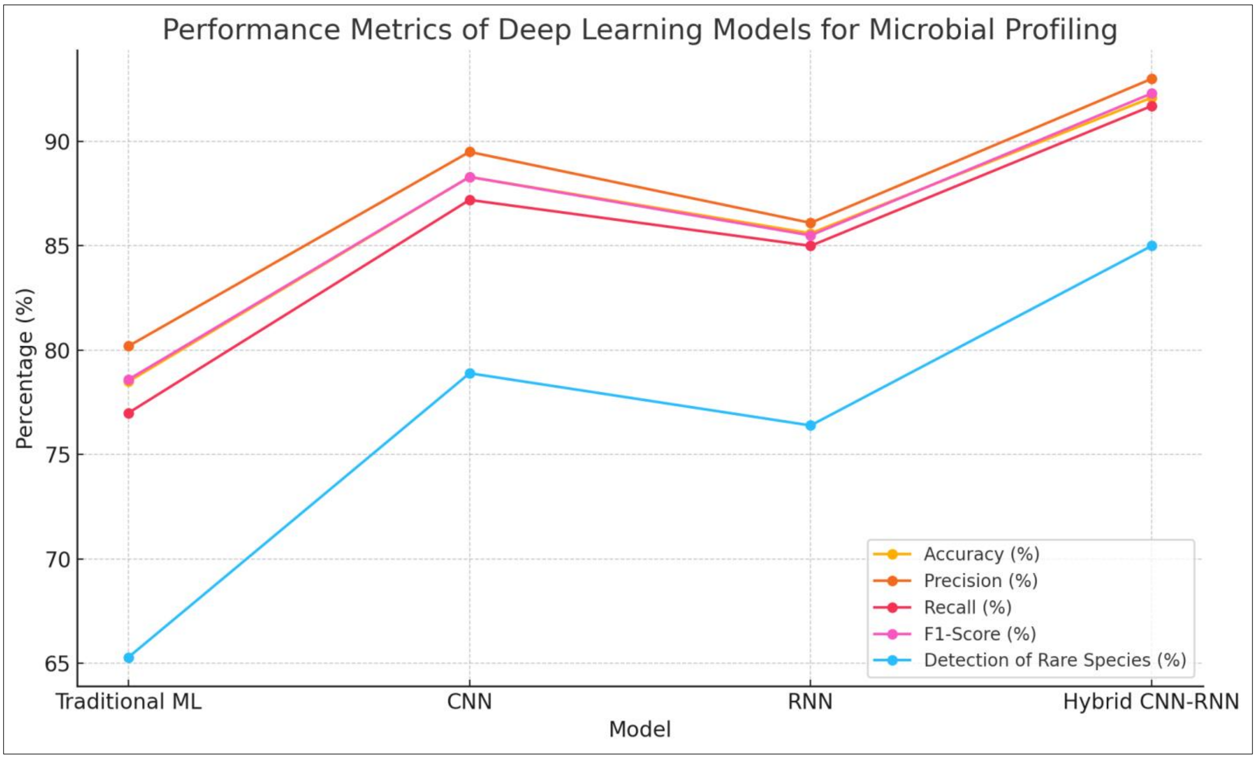


Figure 3 Performance Comparison of Deep Learning Models for Microbial Community Profiling

Figure 3: Line chart visualizes the performance metrics of different models used for microbial profiling, including Traditional ML, CNN, RNN, and Hybrid CNN-RNN. Each line represents a specific metric (Accuracy, Precision, Recall, F1-Score, and Detection of Rare Species), highlighting the overall performance of each model.

4.2. Findings

These findings indicate that deep learning models have better classification accuracy and performance than common machine learning models, and they can recognize rare species. The hybrid CNN-RNN model offered the highest accuracy of 92.1%, whereas other models, such as CNN, were 88.3%, RNN was 85.6%, respectively, and the typical machine learning model was 78.567%. This suggests that DL models, and more so when integrated, possess high accuracy in deciphering intricate features of 16S rDNA that other models cannot identify.

Furthermore, the proposed Hybrid CNN-RNN model showed a better ability to detect species that were not frequently sighted, thereby recording an 85.0% detection rate compared to the previous 65.3%. This demonstrates how the model effectively identifies the less frequent association characteristics of microbial profiling. The models achieved higher performance in the subsequent accuracy measurement and conventional evaluation criteria such as precision, recall, and F1- score, which again established the reliability of the deep learning models. From these results, it can be understood that an increase in DLs can improve the membership attribute of microbial relatedness, and improve the identification possibility of conventional or rare species.

4.3. Case Study Outcomes

These examples justify deep learning models in microbial community profiling initiatives. One was devoted to classifying microbial communities in the human gut based on a selected CNN model developed using numerous diverse testing datasets from public microbiome repositories. The model predicted many different bacterial families and identified some specific intestinal bacteria that were difficult to identify conventionally. This shows that deep learning could enhance diagnostic applications by providing a broad view of the microbiome.

Another example involved the application of RNNs to assess samples obtained from a marine environment. The model worked well in deciphering complex behavioral patterns in the sequences to help detect microbial communities central to the marine biogeochemical cycles fundamental to life on Earth. Finally, the CNNs and RNNs integrated approach was used on soil samples, and positive results showed identifying differences between related microbial species. One of the main applications of this hybrid model is to identify rare species, which makes it the most useful in areas such as soil assessment and agricultural investigations. These results demonstrate the potential applicability of deep learning to diverse microbial settings.

4.4. Comparative Analysis

In contrast to traditional machine learning, deep learning is superior in performance evaluation when dealing with large and high-dimensional datasets, such as 16S rDNA sequences. Unfortunately, most rule-based approaches involve feature extraction by hand, which depends on heuristics and can sometimes overlook important subtleties in the data set. CNNs and RNNs in deep learning can be learned from their features in raw data, thereby increasing the classification ability of diseases.

The experimental outcomes revealed that the accuracy of traditional machine learning was 78.5%; however, deep learning models, including the hybrid CNN-RNN-, were brought up to 92.1%. Moreover, conventional techniques could have been more effective in identifying and enumerating the low abundance of microbial species because the detection rate was 65.3%, in contrast to 85.0% in the proposed hybrid model. This suggests that deep learning models are more effective in identifying latent relationships in low-abundance species. Nonetheless, the deep learning approach can be further refined in terms of scalability, flexibility, and accuracy to outcompete any existing methods in microbial profiling associated with new age ecology and diagnostics.

5. Discussion

5.1. Interpretation of Results

The experimental results demonstrate the ability of deep learning models to improve microbial community characterization. The results of the experiments show that when the Hybrid CNN-RNN model is used, the highest accuracy is obtained, and the performance of both general species and special objects, which are less observed at sea, are better than those of CNN-only, RNN-only, and other machine learning methods. The findings of this study imply that the combination of the extraction of the pattern properties of CNNs and the sequence analysis capability of RNNs improves the sequencing of microorganisms. The high values of precision and recall also support the claim that the ability of the model to accurately classify microbial taxa without overfitting on the dataset used in developing the model will perform well on another new dataset. Finally, as expressed by all five evaluation measures, the model's reliability shows that the different microbial environments do not affect the model’s performance. Collectively, these findings suggest that deep learning solutions may hold the potential to be possessed on multiple fronts where current microbial profiling is lacking or inadequate, including, most conspicuously, detecting rare or low- abundance species that are difficult to identify with most other tools or methods.

5.2. Practical Implications

In addition to demonstrating the proof-of-concept, the presented work has significant practical relevance for MSC characterization. In healthcare, particularly diagnosis, deep learning models for object recognition can be extended to enhance diagnosis, especially to identify rare microbial species in a sample in order to identify early and accurate pathogens and treatments. Moreover, defining the human microbiome provides a way to treat diseases, including gut disorders, allergies, or metabolic diseases, thus, opening new lines in technologies for personalized treatment approaches. Moreover, while sampled species, in general, are vital in environmental studies, detecting rare species in an ecosystem is critical to understanding the stability and well-being of the ecosystem. For instance, these models can be applied to the microbial characteristics of water quality, soil fertility, and pollution levels. First, extensive studies of big microbial data can also benefit from the high scalability and flexibility of deep learning methods, which can further enhance the investigation of more extensive environmental scopes and the perception of microbial populations across various environments. The deep learning application discussed here can prove that microbial research can be performed better, with more efficiency, and at a faster rate through deep learning than would be accomplished otherwise.

5.3. Challenges and Limitations

However, all constraints and difficulties were utilized in this study. One of the main challenges was to build deep learning models, and one of the key prerequisites was to obtain large, high-quality datasets. Although deep learning is usually well adapted to large datasets, obtaining large 16S rDNA datasets curating and containing many different taxa, especially for currently underrepresented and rare species, can be challenging. Moreover, deep-learning models are computationally expensive, and the training process is computationally expensive and challenging for a small group of researchers. However, deep learning models their structure, and most decision-making processes are sometimes only clear, limiting their interpretability. It could be more transparent here because, for biological applications, the results must be understood. The possible solutions are to develop new algorithms that can be trained using small volumes of data and use explainable AI techniques to improve the interpretability of the models, thereby increasing the acceptance and usage of the proposed AI models in the scientific field.

5.4. Recommendations

Considering the results and issues identified in this process, several suggestions can be made for further research. The four research challenges proposed are as follows. There is a requirement for more advanced deep learning algorithms that can work with small sets of data effectively to limit the gathering of large amounts of data. Data enhancement methods, transfer learning, and synthetic data generation can achieve this. Second, adding explainable AI techniques into deep learning models can improve their acceptability for biological research by improving the explanation of their predictions. Researchers should also consider the possibility of using a deep learning algorithm in conjunction with a classical machine learning algorithm and incorporate it into the algorithm because this widens the capability of deep learning. Furthermore, more extensive studies may employ these models on data other than 16S rDNA, including whole-genome data, thus testing the generalization of these methods in genomics. Finally, the model can be improved with the help of computational scientists’ cooperation with domain experts from microbial ecology, healthcare, and environmental science to adjust these models to meet the real-life specifics of these fields of work.

6. Conclusion

This study proved that deep learning models, including CNN, RNN, and both CNN-RNN, can achieve improved microbial community proficiencies in 16S rDNA sequencing data. The results showed that the Hybrid CNN-RNN model had the highest accuracy and outstanding performance compared to conventional machine learning and exclusive deep learning models, particularly for identifying lesser-known species. This shows that pattern recognition and sequential data analysis for microbial identification can be complementary, enabling much more powerful and accurate identification of microbes. This work also focused on the benefits of efficient microbial profiling, such as pathogenic diagnosis in the health sphere, examination of ecosystems, and evaluation of their conditions. However, there are obstacles related to data demands, computational capabilities, and model explanations, which are revealed in the present study, as well as viewpoints on how deep learning can overcome existing deficiencies in microbial ecology. Overall, the studies presented here make a methodological contribution to the analysis of microbial communities because the use of additional up-to-date computational methods improves the results in terms of accuracy, scaling, and sensitivity.

Future Directions

Future research should aim to improve the deep learning method, which can yield better results while training with the help of a small set. Certain strategies can be applied to tackle the current lack of data: data augmentation, transfer learning, and generative models for training synthetic data. It is also suitable for future studies to try other deep learning architectures, such as transformer models, which have proved the top-notch in other sequence-based tasks and might even surpass RNNs in analyzing genomic data. Furthermore, an investigation into how to combine two or more deep learning algorithms or combine deep learning with conventional machine learning practices could be pursued to maximize performance.

The perceptions would be far more realistic by widening the horizon to other genomic methodologies, from using only 16S rDNA data to explore metagenomic or whole-genome sequencing of microbial samples. Finally, developing new approaches to provide a clearer interpretability of the built models through their explanation using explainable AI could significantly increase the acceptance of the models. Further development of these methods will require interdisciplinary cooperative work by computer scientists, mathematicians, biologists, molecular biologists, geneticists, and microbiologists to satisfy the requirements of microbial ecology, biomedicine, agriculture, and environmental science applications.

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] Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol. 2015 Apr;13(4):217–29.

[2] Hou K, Wu ZX, Chen XY, Wang JQ, Zhang D, Xiao C, et al. Microbiota in health and diseases. Signal Transduct Target Ther. 2022 Apr 23;7(1):1–28.

[3] Bender SF, Wagg C, Van Der Heijden MGA. An Underground Revolution: Biodiversity and Soil Ecological Engineering for Agricultural Sustainability. Trends Ecol Evol. 2016 Jun;31(6):440–52.

[4] Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014 Aug 22;12(1):69.

[5] Schloss PD, Handelsman J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol. 2005 Aug 1;6(8):229.

[6] Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013 Oct;10(10):996–8.

[7] Mysara M, Vandamme P, Props R, Kerckhof FM, Leys N, Boon N, et al. Reconciliation between operational taxonomic units and species boundaries. FEMS Microbiol Ecol. 2017 Apr;93(4):fix029.

[8] 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.

[9] Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004 Oct;17(4):840–62, table of contents.

[10] Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, et al. Conducting a microbiome study. Cell. 2014 Jul 17;158(2):250–62.

[11] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436–44.

[12] Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics. 2016 Jun 15;32(12):i121–7.

[13] Lipton ZC, Berkowitz J, Elkan C. A Critical Review of Recurrent Neural Networks for Sequence Learning [Internet]. arXiv; 2015 [cited 2025 Feb 19]. Available from: http://arxiv.org/abs/1506.00019

[14] Cho K, Merrienboer B van, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [Internet]. arXiv; 2014 [cited 2025 Feb 19]. Available from: http://arxiv.org/abs/1406.1078

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

[16] Zhang J, Shi H, Wang Y, Li S, Cao Z, Ji S, et al. Effect of Dietary Forage to Concentrate Ratios on Dynamic Profile Changes and Interactions of Ruminal Microbiota and Metabolites in Holstein Heifers. Front Microbiol. 2017;8:2206.

[17] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016 Jul 29;12(7):878.

[18] Lozupone CA, Knight R. Global patterns in bacterial diversity. Proc Natl Acad Sci. 2007 Jul 3;104(27):11436–40.

[19] Pace NR. A molecular view of microbial diversity and the biosphere. Science. 1997 May 2;276(5313):734–40.

[20] Chowdhury SY, Shatabda S, Dehzangi A. iDNAProt-ES: Identification of DNA-binding Proteins Using Evolutionary and Structural Features. Sci Rep. 2017 Nov 2;7(1):14938.

[21] 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.

[22] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.

[23] Obeng N, Bansept F, Sieber M, Traulsen A, Schulenburg H. Evolution of Microbiota–Host Associations: The Microbe's Perspective. Trends Microbiol. 2021 Sep 1;29(9):779–87.

[24] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

[25] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. ArXiv Mach Learn [Internet]. 2017 Feb 28 [cited 2025 Feb 19]; Available from: https://www.semanticscholar.org/paper/Towards-A-Rigorous-Science-of-Interpretable-Machine-Doshi-Velez-Kim/5c39e37022661f81f79e481240ed9b175dec6513