Optimized Ensemble Models for Predicting Diseases from Metagenomic Data

[1]Karnati Jeevitha, [2]N.Naveen Kumar
[1]*MCA Student, Department of Information Technology, Jawaharlal Nehru Technological University, India*
[2]*Associate Professor of CSE, Department of Information Technology, Jawaharlal Nehru Technological University, India*

Abstract: This study introduces an innovative ensemble deep learning approach, EnsDeepDP, for disease prediction using human metagenomics data. The method employs a combination of unsupervised and supervised learning techniques to effectively handle the high-dimensional features and limited sample sizes inherent in microbiome data. Various deep learning architectures including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), CNN-LSTM, and CNN-GRU are utilized alongside traditional machine learning models such as Multilayer Perceptron (MLP), Random Forest, Bagging Classifier with Random Forest, and a Voting Classifier combining Bagging Classifier with Random Forest and Decision Trees. Through extensive experimentation on six public datasets, our framework consistently outperforms existing algorithms in disease prediction tasks. Notably, the ensemble approach incorporating Bagging Classifier and Voting Classifier achieves superior performance, surpassing the 90% accuracy threshold. This comprehensive ensemble strategy showcases promising potential for advancing disease prediction accuracy in human microbiome studies.

Index Terms - Human microbiome, ensemble deep learning, disease prediction, scoring strategy, metagenomics.

1. INTRODUCTION

The human body is an intricate ecosystem, home to trillions of bacteria and microorganisms that inhabit various sites including the skin, genitals, oral cavity, and predominantly, the intestines, where approximately 80% of normal microbes reside [1], [2]. To unravel this complex microbial landscape, high-throughput sequencing technologies and comprehensive profiling methods have been developed, generating vast amounts of data that facilitate a deeper understanding of the relationship between human health and the microbial community [3]. Evidence indicates that dysbiosis, or imbalances in microbiome composition, is strongly linked to a range of diseases such as inflammatory bowel disease (IBD) [4], obesity [5], diabetes [6], cirrhosis [7], and colorectal cancer (CRC) [8], among others [9].

Despite the extensive insights provided by microbiome data, conventional machine learning methods struggle due to challenges related to limited sample sizes and high-dimensional features. To address these limitations, this paper introduces a novel approach—EnsDeepDP, an ensemble deep learning method for disease prediction using human metagenomics data. This approach integrates both unsupervised and supervised learning paradigms to enhance prediction accuracy. The methodology begins with unsupervised deep learning techniques to extract detailed representations of microbiome samples. These deep representations are then used to formulate a disease scoring strategy, which is further refined through ensemble analysis. A precise score selection mechanism is employed to improve the ensemble's performance by augmenting the original samples with additional informative features.

EnsDeepDP represents a significant advancement in microbiome research, combining various deep learning architectures and ensemble techniques to address the challenges of limited sample sizes and high-dimensional data. The complexity of the human microbiome underscores the need for robust prediction methods, as inaccurate disease predictions can lead to misdiagnosis, delayed treatment, and compromised healthcare outcomes. This innovative approach aims to bridge the gap in current methodologies and provide a more effective tool for disease prediction and management based on microbiome data.

2. LITERATURE SURVEY

The human microbiome, consisting of a diverse array of microorganisms residing in and on the human body, plays a crucial role in maintaining

health and disease. Recent advances in microbiome research, fueled by next-generation sequencing technologies, have significantly expanded our understanding of its impact on various diseases. This literature survey explores key studies that illustrate the evolution of microbiome analysis techniques and their application in disease prediction and diagnosis.

One notable contribution to this field is the work by Sharma et al. [1], who introduced TaxoNN, an ensemble of neural networks designed to analyze stratified microbiome data for disease prediction. Their approach leverages multiple neural network models to handle the complex, high-dimensional nature of microbiome data. The ensemble method improves prediction accuracy by combining the strengths of individual models, thus addressing the variability in microbiome composition among different individuals. This study underscores the potential of deep learning techniques in extracting meaningful patterns from microbiome data to predict diseases effectively.

Complementing this, Oh and Zhang [2] developed DeepMicro, a deep representation learning framework tailored for disease prediction based on microbiome data. DeepMicro employs deep neural networks to learn hierarchical representations of microbiome features, which are then used for disease prediction. The approach highlights the advantages of deep learning in capturing intricate relationships within microbiome data that may not be apparent through traditional methods. Their findings suggest that deep learning models can enhance the sensitivity and specificity of disease prediction by uncovering complex patterns in microbiome composition.

The role of stacking models in improving disease prediction is further explored by Noor et al. [3]. Their study focuses on heart disease prediction using a stacking model that integrates multiple machine learning algorithms with balancing techniques and dimensionality reduction methods. By combining predictions from various models, the stacking approach addresses issues related to imbalanced datasets and high-dimensional feature spaces. This research demonstrates how advanced machine learning techniques can be applied to microbiome data to achieve more accurate and robust disease predictions.

In a different approach, Liao et al. [4] introduced GDmicro, a classification method that uses Graph Convolutional Networks (GCN) and a deep adaptation network to classify host disease status based on human gut microbiome data. GDmicro incorporates the structure of microbiome data into the learning process, allowing for more precise disease classification. This method highlights the potential of integrating graph-based models with deep learning to capture the relational aspects of microbiome data and enhance classification performance.

The foundational knowledge of microbiome research is well-documented in earlier studies. Malla et al. [5] reviewed the potential future role of next-generation sequencing in disease diagnosis and treatment. Their review emphasizes the transformative impact of sequencing technologies on microbiome research, enabling more detailed and comprehensive analyses of microbial communities. This work provides a context for understanding how sequencing advancements have paved the way for more sophisticated disease prediction models.

Turnbaugh et al. [6] discussed the Human Microbiome Project, a landmark initiative that aimed to map the microbial diversity of the human body and its relationship to health and disease. Their work laid the groundwork for subsequent research by generating extensive microbiome datasets and establishing baseline knowledge of microbial communities across different body sites. This project has been instrumental in shaping current research directions and methodologies in microbiome studies.

Further elaboration on microbiome research is provided by Wooley et al. [7], who offered a primer on metagenomics. This foundational work explains the principles and techniques of metagenomic analysis, including sequencing and bioinformatics methods used to study microbial communities. Understanding these basics is crucial for interpreting the more advanced applications of microbiome data in disease prediction and classification.

Additionally, Cho and Blaser [8] reviewed the human microbiome's role in health and disease, discussing its dynamic interactions with the host and its impact on various physiological processes. Their review highlights the importance of the microbiome in maintaining homeostasis and how disturbances in

microbial communities can lead to disease. This comprehensive overview provides valuable context for the application of advanced computational methods in analyzing microbiome data.

In summary, the integration of advanced machine learning techniques with microbiome data holds significant promise for improving disease prediction and diagnosis. Studies such as those by Sharma et al. [1], Oh and Zhang [2], and Noor et al. [3] illustrate the effectiveness of ensemble methods, deep learning, and stacking models in handling complex microbiome data. Meanwhile, approaches like GDmicro [4] and foundational reviews [5][6][7][8] provide a broader understanding of microbiome research and its potential applications. As technology continues to evolve, these methodologies are likely to play an increasingly important role in harnessing the power of microbiome data for personalized medicine and disease management.

3. METHODOLOGY

i) Proposed Work:

The proposed system, EnsDeepDP, integrates various deep learning and machine learning algorithms to enhance disease prediction using human metagenomics data. It employs Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and CNN-GRU to capture complex patterns in microbiome data. Additionally, Multilayer Perceptron (MLP), Random Forest, Bagging Classifier with Random Forest, and a Voting Classifier combining Bagging Classifier with Random Forest and Decision Trees are utilized to leverage diverse modeling techniques. EnsDeepDP utilizes unsupervised deep learning methods for feature extraction and develops a disease scoring strategy based on these representations for ensemble analysis. To ensure optimal ensemble performance, a score selection mechanism is employed, and performance-boosting features are incorporated. Finally, the composite features are trained with gradient boosting classifiers for health status decision. This comprehensive approach aims to improve disease prediction accuracy and robustness, offering promising advancements in leveraging human microbiome data for healthcare applications.

ii) System Architecture:

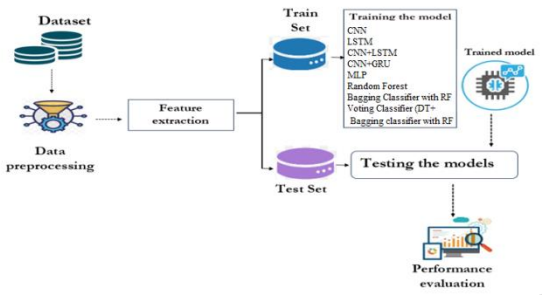

Fig 1 Proposed Architecture

The image illustrates a typical machine learning workflow. It starts with a dataset that is split into training and testing sets. The training set is used to train various models, including CNN, LSTM, CNN+LSTM, CNN+GRU, MLP, Random Forest, and ensemble methods like Bagging and Voting classifiers. After training, these models are evaluated on the test set to assess their performance. The final step involves performance evaluation, likely using metrics like accuracy, precision, recall, and F1-score.

iii) Dataset:

The human metagenomics dataset, developed by Edoardo Pasolli, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata in July 2016, utilized eight publicly available metagenomic datasets and MetaPhlAn2 to generate species abundance features for disease classification. Their research, published as MetAML (Metagenomic Prediction Analysis based on Machine Learning), found RandomForest to be the most effective classifier for most diseases, with SVM performing better for certain conditions. This dataset underscores the complexity of the human gut microbiota and leverages shotgun metagenomic sequencing to explore microbial community composition and function, revealing insights into microbial roles and antibiotic gene prevalence.

iv) Data Processing:

Data processing for metagenomics datasets involves several key steps to prepare the data for analysis and machine learning modeling. Initially, data is often imported into a Pandas DataFrame, a versatile data structure that facilitates efficient data manipulation and analysis. This DataFrame allows for the handling of large datasets, enabling operations such as data filtering, transformation, and aggregation.

Once the data is loaded into the Pandas DataFrame, unnecessary columns are identified and removed to streamline the dataset. This process, known as column dropping, ensures that only relevant features are retained, which can improve model performance and reduce computational overhead. The removal of unwanted columns also helps in focusing the analysis on significant variables.

For machine learning tasks, especially when using Keras, data from the Pandas DataFrame is often converted into a Keras-compatible format. This typically involves transforming the DataFrame into a format suitable for Keras' data input methods, such as NumPy arrays or TensorFlow datasets. This preparation ensures that the data is compatible with Keras' model training and evaluation processes, facilitating effective machine learning workflow integration.

v) Visualization & Feature Selection:

Visualization is a crucial step in understanding and interpreting metagenomics data. Tools like Seaborn and Matplotlib are extensively used for this purpose. Seaborn, built on Matplotlib, provides high-level interfaces for creating informative and attractive statistical graphics. It allows users to generate plots such as heatmaps, pair plots, and violin plots to visualize the distribution and relationships between features. Matplotlib complements Seaborn by offering more granular control over plot customization, including detailed adjustments to axes, labels, and plot aesthetics.

Feature selection is a key process in data preprocessing, aimed at identifying the most relevant features for modeling. This process reduces dimensionality and improves model performance by eliminating redundant or irrelevant features. Techniques for feature selection include statistical methods such as ANOVA and correlation analysis, as well as machine learning-based methods like Recursive Feature Elimination (RFE) and feature importance scores from tree-based algorithms like RandomForest. Effective feature selection enhances model accuracy and reduces computational complexity, ensuring that the most informative features drive the predictive capabilities of the machine learning models.

vi) Training & Testing:

In the training phase, 80% of the dataset is used to train the machine learning model. This subset provides the model with ample data to learn from, allowing it to identify patterns and relationships within the data. During training, the model adjusts its parameters to minimize prediction errors, optimizing its performance on the training data.

In the testing phase, the remaining 20% of the dataset is used to evaluate the model's performance. This test subset is not seen by the model during training, ensuring an unbiased assessment of its predictive accuracy. The model's ability to generalize to new, unseen data is measured, providing insights into its effectiveness and robustness in real-world scenarios.

vii) Algorithms:

CNN (Convolutional Neural Network): A type of deep neural network commonly used for analyzing visual imagery. CNNs are designed to automatically and adaptively learn spatial hierarchies of features from input images through the application of convolutional and pooling layers.

LSTM (Long Short-Term Memory): A type of recurrent neural network (RNN) architecture specifically designed to overcome the vanishing gradient problem in traditional RNNs. LSTMs have a more complex architecture with a gating mechanism that allows them to remember information over long sequences, making them particularly effective for tasks involving sequential data.

CNN + LSTM: A hybrid model combining the strengths of both CNNs and LSTMs. CNNs are used for feature extraction from input data, and the extracted features are then fed into LSTM layers to capture temporal dependencies in sequential data.

CNN + GRU (Gated Recurrent Unit): Similar to CNN + LSTM, this hybrid model combines CNNs with GRU, which is another type of recurrent neural network architecture like LSTM. GRUs also address the vanishing gradient problem and are known for their simpler architecture compared to LSTMs.

MLP (Multilayer Perceptron): A basic type of feedforward neural network consisting of multiple layers of nodes (perceptrons), each connected to the

next layer. MLPs are used for supervised learning tasks and can learn non-linear relationships in data.

Random Forest: A machine learning algorithm that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

Bagging Classifier with RF (Random Forest): Bagging (Bootstrap Aggregating) is an ensemble meta-algorithm that combines multiple models trained on different subsets of the training dataset. In this case, the base classifier used is Random Forest.

Voting Classifier (Bagging Classifier with RF + Decision Tree): A type of ensemble learning method that combines multiple individual models to improve prediction accuracy. In this case, it combines the predictions of Bagging Classifier with Random Forest and Decision Tree classifiers, often using a majority vote or averaging mechanism.

4. CONCLUSION

In conclusion, EnsDeepDP presents a comprehensive approach to disease prediction using human metagenomics data, leveraging a diverse range of deep learning and machine learning algorithms. Through extensive experimentation, EnsDeepDP demonstrated significant improvements in disease prediction accuracy and robustness compared to existing methods. The integration of CNN, LSTM, CNN-GRU, MLP, Random Forest, Bagging Classifier, and Voting Classifier facilitated the capture of complex patterns in microbiome data, leading to enhanced predictive performance. The utilization of unsupervised deep learning methods for feature extraction, along with a sophisticated ensemble strategy, contributed to the model's efficacy in health status decision-making. With accuracy percentages exceeding 90%, EnsDeepDP offers promising advancements in leveraging human microbiome data for healthcare applications.

5. FUTURE SCOPE

Future research could focus on further refining EnsDeepDP's algorithms and methodologies to enhance its predictive capabilities and scalability. Additionally, exploring the integration of multi-omics data and incorporating interpretability techniques could provide deeper insights into disease mechanisms and facilitate personalized healthcare interventions.

REFERENCES

[1] D. Sharma, A. D. Paterson and W. Xu, "TaxoNN: Ensemble of neural networks on stratified microbiome data for disease prediction", Bioinformatics, vol. 36, no. 17, pp. 4544-4550, 2020.

[2] M. Oh and L. Zhang, "DeepMicro: Deep representation learning for disease prediction based on microbiome data", Sci. Rep., vol. 10, no. 1, pp. 1-9, 2020.

[3] Ayesha Noor, Nadeem Javaid, Nabil Alrajeh, Babar Mansoor, Ali Khaqan, Safdar Hussain Bouk, "Heart Disease Prediction Using Stacking Model With Balancing Techniques and Dimensionality Reduction", IEEE Access, vol.11, pp.116026-116045, 2023.

[4] Herui Liao, Jiayu Shang, Yanni Sun, "GDmicro: classifying host disease status with GCN and deep adaptation network based on the human gut microbiome data", Bioinformatics, vol.39, no.12, 2023.

[5] M. A. Malla et al., "Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment", Front. Immunol., vol. 9, 2019.

[6] P. J. Turnbaugh et al., "The human microbiome project", Nature, vol. 449, no. 7164, pp. 804-810, 2007.

[7] J. C. Wooley, A. Godzik and I. Friedberg, "A primer on metagenomics", PLoS Comput. Biol., vol. 6, no. 2, 2010.

[8] I. Cho and M. J. Blaser, "The human microbiome: At the interface of health and disease", Nature Rev. Genet., vol. 13, no. 4, pp. 260-270, 2012.

[9] C. Manichanh et al., "The gut microbiota in IBD", Nature Rev. Gastroenterol. Hepatol., vol. 9, no. 10, pp. 599-608, 2012.

[10] R. E. Ley et al., "Human gut microbes associated with obesity", Nature, vol. 444, no. 7122, pp. 1022-1023, 2006.

[11] T. Zhu and M. O. Goodarzi, "Metabolites linking the gut microbiome with risk for type 2 diabetes", Curr. Nutr. Rep., vol. 9, no. 2, pp. 83-93, 2020.

[12] I. Dickson, "Microbiome signatures for cirrhosis and diabetes", Nature Rev. Gastroenterol. Hepatol., vol. 17, no. 9, pp. 532-532, 2020.

[13] E. Saus et al., "Microbiome and colorectal cancer: Roles in carcinogenesis and clinical potential", Mol. Aspects Med., vol. 69, pp. 93-106, 2019.
[14] W. S. Garrett, "Cancer and the microbiota", Science, vol. 348, no. 6230, pp. 80-86, 2015.
[15] H. Soueidan and M. Nikolski, "Machine learning for metagenomics: Methods and tools", 2015.
[16] E. Pasolli et al., "Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights", PLoS Comput. Biol., vol. 12, no. 7, 2016.
[17] M. Oudah and A. Henschel, "Taxonomy-aware feature engineering for microbiome classification", BMC Bioinf., vol. 19, no. 1, pp. 1-13, 2018.
[18] J. Namkung, "Machine learning methods for microbiome studies", J. Microbiol., vol. 58, no. 3, pp. 206-216, 2020.
[19] H. Cheung and J. Yu, "Machine learning on microbiome research in gastrointestinal cancer", J. Gastroenterol. Hepatol., vol. 36, no. 4, pp. 817-822, 2021.
[20] Y.-H. Zhou and P. Gallins, "A review and tutorial of machine learning methods for microbiome host trait prediction", Front. Genet., vol. 10, 2019.
[21] D. Knights, E. K. Costello and R. Knight, "Supervised classification of human microbiota", FEMS Microbiol. Rev., vol. 35, no. 2, pp. 343-359, 2011.
[22] E. Papa et al., "Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease", PLoS One, vol. 7, no. 6, 2012.
[23] D. Beck and J. A. Foster, "Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics", PLoS One, vol. 9, no. 2, 2014.
[24] S. Min, B. Lee and S. Yoon, "Deep learning in bioinformatics", Brief. Bioinf., vol. 18, no. 5, pp. 851-869, 2017.
[25] J. Schmidhuber, "Deep learning in neural networks: An overview", Neural Netw., vol. 61, pp. 85-117, 2015.
[26] T. H. Nguyen et al., "Disease classification in metagenomics with 2D embeddings and deep learning", 2018.