Random Forest Approach to Identifying Microbial Biomarkers Relating to COVID-19 Severity

Jialin Lyu
School of Computing
Ulster University
Belfast, UK
lyu-j@ulster.ac.uk

Haiying Wang
School of Computing
Ulster University
Belfast, UK
hy.wang@ulster.ac.uk

Taranjit Singh Rai
School of Medicine
Ulster University
Derry/Londonderry, UK
t.rai@ulster.ac.uk

Huiru Zheng*
School of Computing
Ulster University
Belfast, UK
h.zheng@ulster.ac.uk

***Abstract*—The Coronavirus Disease 2019 (COVID-19) pandemic, caused by the novel coronavirus SARS-CoV-2, has led to significant mortality and hospitalizations worldwide. Although the spread of the pandemic has recently slowed, the persistence of long-COVID remains a continuous public health concern. Substantial data have been amassed to delineate and synopsize the fundamental symptoms of COVID-19, encompassing respiratory distress, muscular discomfort, anosmia, limb paresthesia, and fatigue. Notwithstanding the extensive investigations conducted in this domain, a pivotal unresolved issue pertaining to the pandemic is the considerable interindividual variability in the severity of COVID-19 symptoms following SARS-CoV-2 infection. Despite concerted research efforts, the precise biomarkers from the human microbiome that underlie the cause of different severity of COVID-19 symptoms remain elusive. This study aims to develop a Machine Learning (ML) model for COVID-19 severity classification between asymptomatic and moderate and to identify critical microbial biomarkers for COVID-19 severity using previously published 16S rRNA gene sequencing data. The employed Random Forest (RF) model has exhibited outstanding performance with an accuracy of 0.82, an F1-score of 0.88 and an AUC of 0.83. Our results show *Selenomonas*, *Neisseria*, *Prevotella*, *Veillonella* and *Rothia* are critical microbial biomarkers related to COVID-19 severity.**

***Index Terms*—COVID-19, 16S rRNA gene sequencing, COVID-19 Severity Classification, Asymptomatic, Moderate, Machine Learning (ML), Random Forest (RF)**

I. Introduction

The SARS-CoV-2 infection, commonly known as the coronavirus, and the resulting COVID-19 pandemic have emerged as a global crisis impacting both public health and economics across societies worldwide. According to the World Health Organization (WHO), the symptoms of COVID-19 exhibit variability. Individuals with mild cases generally manifest symptoms such as fever, cough, and fatigue. Moderate-severity cases may involve difficulty in breathing and mild pneumonia. In contrast, severe cases may entail acute pneumonia, potential organ failure, and, in some instances, a risk of mortality. In particular, asymptomatic individuals may exhibit lower infectiousness compared to symptomatic ones, but they still constitute a significant portion of all SARS-CoV-2 infections. Additionally, asymptomatic individuals tend to have more social contacts than symptomatic ones, as the latter are more likely to self-isolate when feeling unwell. Finding out and understanding asymptomatic infections is essential for early prevention and control of COVID-19 worldwide [1]. Hence, the contribution of asymptomatic cases to the community transmission of COVID-19 should not be underestimated [2].

Previous research [3]–[9] has mainly concentrated on examining the presence, absence, or differential abundance of particular microbes in COVID-19, while fewer studies have employed ML to uncover the responsible biomarkers between COVID-19-positive or negative [10], [11], severe or moderate [12], [13], intubation or not [13]. However, research using ML to uncover the microbial biomarkers linked to asymptomatic cases of COVID-19 is conspicuously rare.

At the beginning of 2020, 831 adults without known SARS-CoV-2 infection were followed weekly for six months to study the occurrence and progression of SARS-CoV-2 infection [14]. A relative stability of the oral microbiome profile throughout the duration of SARS-CoV-2 infection was reported, with mild to moderate cases showing minimal alteration while severe cases exhibiting notable early shifts. Nevertheless, the cause of having asymptomatic or other symptomatic severity in COVID-19 remains unclear. Further research specifically on asymptomatic cases based on this published dataset is still needed to elucidate the underlying causes and identify relevant microbial biomarkers.

Addressing the gaps in the field and adding on values to the previously published work [14], an ML model has been employed in our work to classify between asymptomatic or moderately symptomatic instances from SARS-CoV-2 exposed individuals. 16s rRNA gene sequencing data was preprocessed to obtain Amplicon Sequence Variants (ASV) features for training the model. The optimal RF model was identified by grid-search. Critical microbial biomarkers were ranked by the optimal RF model to provide biological insights.

II. Related Works

Explainable artificial intelligence (XAI) methods were proposed based on conventional ML using COVID-19 metagenomic next-generation sequencing (mNGS) data to classify between COVID-19 negative and positive cohorts [10]. The Extreme Gradient Boosting (XGBoost) model was reported

This research is funded by Ulster University VCRS PhD Studentship.

979-8-3503-5298-6/24/$31.00 ©2024 IEEE

to outperform the Logistics Regression (LR), Support Vector Machine (SVM), and RF models, with an accuracy of 0.93. An explainable strategy that leverages the Local Interpretable Model-agnostic Explanations (LIME) and the SHAPley Additive Explanations (SHAP) techniques was applied to identify potential COVID-19 biomarker genes of the model. The SHAP identified IFI27, LGR6, and FAM83A as the three most critical genes responsible for COVID-19. According to the LIME, an elevated level of IFI27 gene expression in particular increased the likelihood of a positive class.

The RF model has been applied to discriminate between COVID-19 patients who need intubation or not [13], as well as those who have experienced severe symptoms or moderate symptoms [12], [13]. Oropharyngeal (OP) and Nasopharyngeal (NP) samples were collected from COVID-19 samples, integrating bacteria relative abundance data on genus level and small circular DNA virus copy numbers. As a result, the ML model exhibited an effective performance with an AUC of 0.86 using OP data for discriminating between intubation/none-intubation patients, and an AUC of 0.82 using OP data for discrimination moderately/severely diseased patients. Moreover, small circular DNA viruses of OP samples were reported to contribute the most to distinguishing moderately/severely diseased patients, Mycoplasma and Prevotella were found to be positively responsible for intubation [13]. 241 stool samples were collected from 127 hospitalized patients, and gut microbiome taxonomic was profiled. The RF model exhibited an excellent result with an AUC of 0.925 to distinguish between severely and moderately diseased patients [12].

Machine learning has proven its effectiveness in identifying patterns within vast, disorganized, and intricate datasets. Nonetheless, constructing an accurate prediction model based on genotype data presents a formidable challenge, primarily due to the "curse of dimensionality" – wherein the number of features far exceeds the number of samples [15]. Consequently, the performance of ML models necessitates feature selection. This process targets the extraction of informative features while eliminating irrelevant and redundant ones. In the pertinent domain, it is necessary not only to obtain a precise ML model but also to interpret it to identify the critical microbial biomarkers, i.e. the most illustrative features of the issue. In such instances, although feature selection methods such as Lasso [10], Principal component analysis (PCA) [16] and differential abundance analysis [12] could lead to better performance, they still pose challenges in identifying critical microbial biomarkers from trained classification models. LASSO tends to arbitrarily select one feature from a group of highly correlated features and shrink the coefficients of the others [17], [18] in which way bias can be introduced into the model interpretation. PCA transforms original features into orthogonal variables, known as principal components, through linear combinations. While these components preserve much of the original information, interpreting them in the context of the original features can be difficult, especially when they lack clear semantic significance [19]. Differential abundance analysis selects a certain number of features differentially abundant features for training the model. However, it may retain certain highly correlated features, causing data redundancy. Moreover, features selected without the knowledge of the model's performance do not necessarily make them important for training, hence it would be less convincing that ranked features have a critical impact on the model's decisions.

The field of biology has seen a surge in the adoption of ML methods to handle the vast and intricate datasets prevalent in the domain. Within Bioinformatics specifically, the RF model has emerged as a favoured approach [10]–[13], [20]. Its interpretability and distinct strengths in handling small sample sizes, and high-dimensional feature spaces, propelled its increased utilization in computational biology research [21]. Addressing the identified issues of feature selection, we trained an optimal RF model by grid-search using a comprehensive range of correlation thresholds for feature selection, along with a wide range of tree numbers. By doing so, features making the RF model produce optimal results could be obtained, yielding reliable model interpretation results.

III. Methodology

A. Overview

Fig. 1 shows the pipeline of our proposed method. In each training process, the RF model is trained with a certain tree number T_n, and a feature-selected dataset using a certain correlation threshold C_n, the training process enumerates using all possible combinations of T_n and C_n, namely the parameter grid. The model's performance is monitored by its F1 Score, T_n and C_n which provide the highest F1 Score will be noted to get the optimal model and the best-selected dataset. With the optimal RF model and the best-selected dataset, features will be ranked by feature importance scores.

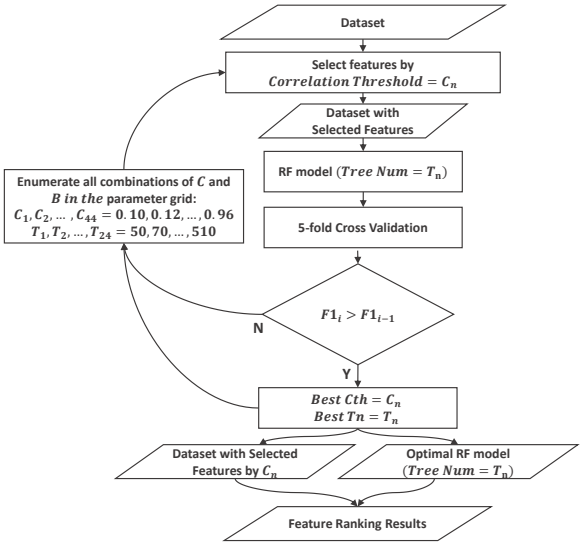

Fig. 1. Proposed Method

B. Dataset

In this study, we employed a publicly available dataset from a study conducted by Armstrong et al. [14]. The data have been obtained from the European Nucleotide Archive (ENA) under accession number (PRJEB62655). A subset was subsequently extracted from the initial dataset, comprising a total of 158 saliva samples subjected to 16S rRNA sequencing. Among these, 47 samples were derived from asymptomatic individuals, serving as the negative samples, while the remaining 111 samples were sourced from individuals exhibiting moderate symptoms, serving as the positive samples. It is pertinent to note that all samples were collected subsequent to individuals testing positive for COVID-19.

C. Data Pre-processing

We used QIIME2 [22] for data pre-processing. ASV features were extracted from the sequence leveraging the DADA2 [23] package in QIIME2. ASV (Amplicon Sequence Variants) features represent unique biological sequences present in the sample derived from sequencing reads generated from PCR (polymerase chain reaction) amplification of specific genomic regions, such as the 16S rRNA gene for bacteria and archaea. A pre-trained Naive Bayes classifier developed by SILVA [24] was applied for taxonomic assignment. As a result, 14222 ASV features were identified from the domain level to the species level. 6713 taxonomic labels on the genus level were subsequently extracted from all the ASV features, they are further grouped by summing the abundance values of taxons on the same genus level, resulting in 203 features.

D. Feature Selection

Feature selection is a crucial step in ML and data analysis tasks, aimed at identifying the most informative features while discarding redundant or irrelevant ones. Correlation-based feature selection is a commonly employed technique that leverages the relationships between features to identify and remove redundant ones. Our correlation-based feature selection method is shown in Fig. 2 and the pipeline of obtaining features to be dropped is shown in Fig. 3.

As Fig. 3 shows, Pearson Correlation coefficients $Corr_{ij}$ are calculated pairwisely among all the features from the original dataset. By enumerating over all the $Corr_{ij}$, a list aimed to store the indexes of features to be dropped appends the second index j which appeared for the first time across the enumeration. By doing so, the second index j of $Corr_{ij}$ of the pairs of the same feature (marked in green in Fig. 2) are avoided from being added to the list across the enumeration, and the second index j of $Corr_{ij}$ which are larger than the correlation threshold (marked in red in Fig. 2) are added in the list, but without any repeating. Finally, features can be selected by dropping the columns of the original dataset with indexes stored in the list.

E. Machine Learning Model

The RF model was trained by 5-fold cross-validation for binary classification. The optimal model was identified through a parameter grid-search encompassing tree numbers within the RF model and correlation thresholds for feature selection. Finally, the selected ASV features utilized for training were ranked based on their feature importance scores obtained from the optimal RF model to provide biological insights into critical microbial biomarkers.

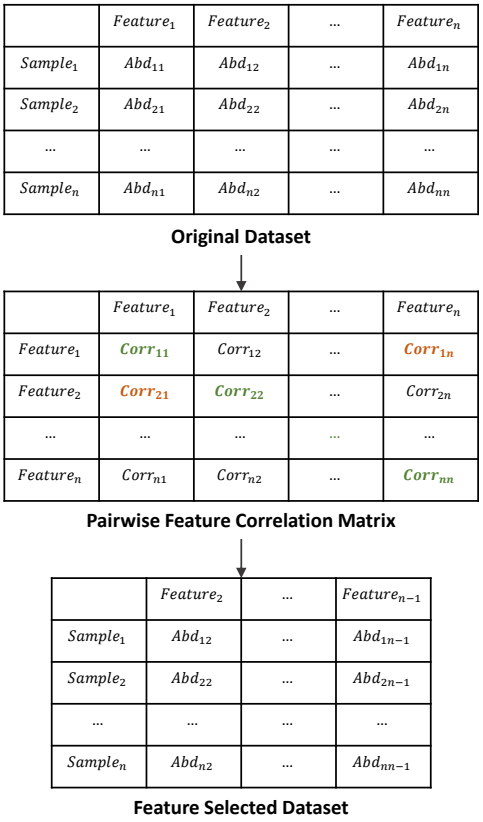

Fig. 2. Feature Selection

F. Model Evaluation

Precision, Recall, F1-score, Receiver Operating Characteristic (ROC) Curve and Area under the ROC Curve (AUC) were used to evaluate the optimal RF model identified by grid-search.

IV. Results and Discussion

A. Grid-search Results

The grid of parameters is formed with correlation thresholds ranging from 0.1 to 0.96 with an interval of 0.02, and tree numbers ranging from 50 to 510 with an interval of 20. The RF model's F1 Score is used to determine the best parameter combination. Fig. 4 shows the grid-search result of the RF model's F1 Score under different correlation threshold-tree number combinations. The data points are scaled to deliver a discernible variation when being mapped in turbo colour. Consequently, an optimal F1 Score of 0.88 was attained with a correlation threshold set to 0.86 and a tree number set to

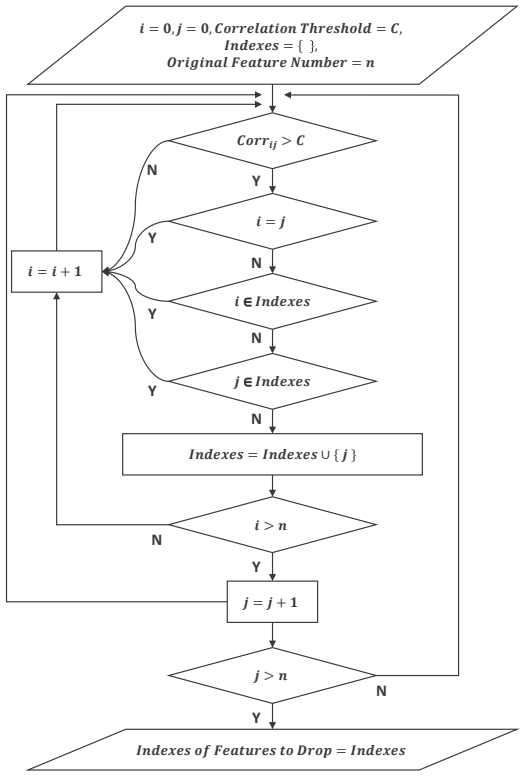

Fig. 3. Pipeline of Getting Features to Drop

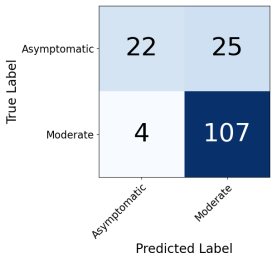

Fig. 5. Confusion Matrix

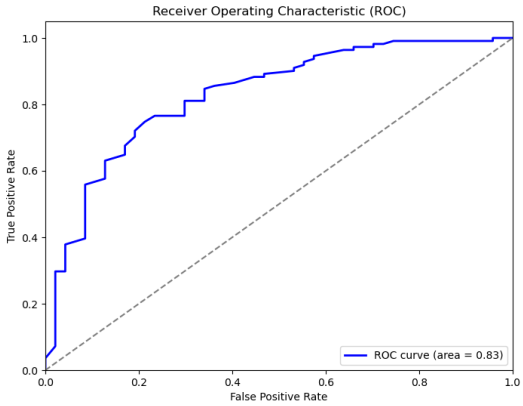

Fig. 6. ROC Curve

170. Under the correlation threshold set to 0.86, 99 features out of 203 were selected.

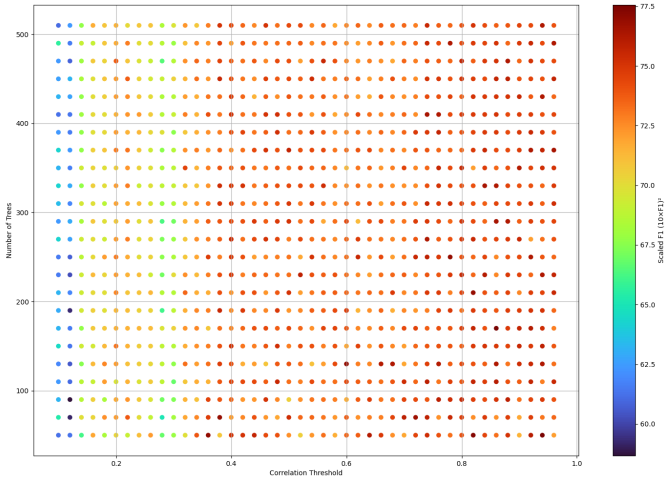

Fig. 4. Grid-search Result

B. Evaluation

Fig. 5 shows the Confusion Matrix of the optimal RF model, yielding a precision of 0.81, a recall of 0.96, an F1-score of 0.88 and an accuracy of 0.82. Fig. 6 shows the ROC Curve of the optimal RF model, yielding a mean AUC across folds of 0.83. To assess the consistency of the model's performance across different folds and provide insights into the variability of the dataset itself, Fig. 7 provides visualization of AUC scores across each fold.

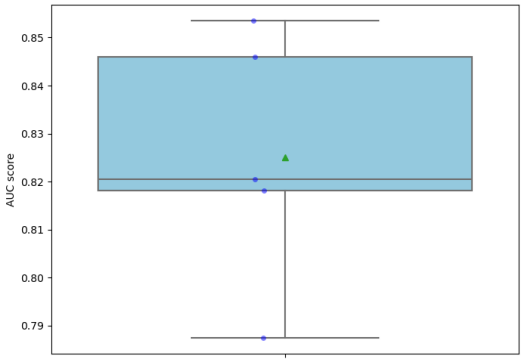

Fig. 7. AUC Scores Across Folds

C. Feature Ranking

Feature importance scores provided by the RF model represent a measure of the relative contribution of each feature to the overall predictive accuracy of the model. These values are derived from the extent to which each feature contributes to the reduction in prediction error when making decisions within the ensemble of decision trees that comprise the RF model. Fig. 8 shows the importance scores of the top 10 ASV features derived from the trained optimal RF model.

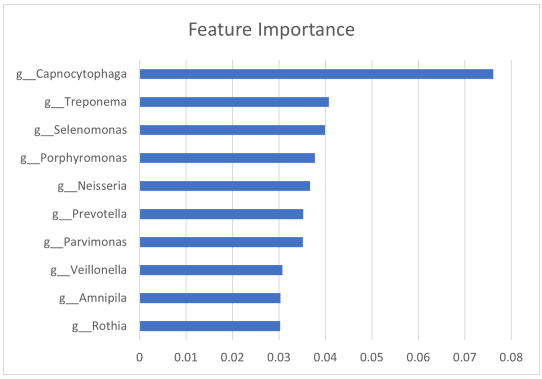

Fig. 8. ASV Feature Importance Scores (Top 10)

D. Relative Abundance of Top10 Features

To further discover the alternations of the top 10 features ranked by the optimal RF model within the Asymptomatic and Moderate cohorts, we have calculated the mean relative abundance within both cohorts, as Fig. 9 shows.

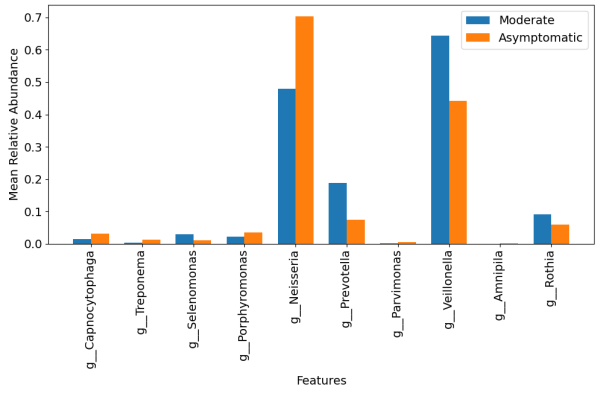

Fig. 9. Mean Relative Abundance of the Top 10 ASV Features

E. Discussion

1) Model Performance: During the grid-search, the F1 Score remains relatively consistent across varying tree numbers. However, a notable increase is observed as the correlation threshold is raised. The findings suggest that the feature selection exerts a more pronounced influence on the performance of the RF model. Given the circumstances pertaining to the imbalanced nature of the dataset, the optimal RF model demonstrated commendable performance in terms of AUC and F1 Score, with a slight flaw in precision as we could anticipate.

The optimal RF model has shown consistency across 5 folds with a concentrated distribution of AUC scores (Fig. 7), suggesting that the model is robust and can generalize well across folds, the features are informative as they have consistent relationships with the target variable across subsets. A concentrated distribution of AUC scores further underscores the dataset likely contains a relatively homogenous set of samples, as each fold contains similar characteristics of the data, leading to consistent model performance.

2) Biological Insights: As Fig. 8 shows, several critical microbial biomarkers were identified to play a significant role in classification. Our discoveries have been corroborated by a body of research. *Prevotella*, *Veillonella* and *Rothia* have been reported to appear more frequently in samples of those who used breathing assistive devices [8], [25], more specifically, 2 species of bacteria within the genus *Prevotella* and genus *Veillonella*, i.e. the *Prevotella histicola* and the *Veillonella dispar* have been reported to increase in the pharynges of COVID-19 patient compared to non-COVID ones [26]. Similarly, *Selenomonas* has been found to be enriched in patients with COVID-19 but reduced in the healthy controls [27]. On the other hand, *Neisseria* was found to decline with COVID-19 severity [13], [28]. The alternation of *Selenomonas*, *Neisseria*, *Prevotella*, *Veillonella* and *Rothia* in our studied samples (Fig. 9) is consistent with findings from above, underlying that the microbiome alternation patterns of Asymptomatic to Moderately symptomatic COVID-19 are similar to those of Non-COVID to COVID-19 and mild to severe.

Rare cases have reported the relationship between COVID-19 infection and *Capnocytophaga* and *Parvimonas*. *Capnocytophaga canimorsus* bacteremia was revealed in a post-mortem of a COVID-19 pneumonia patient. While *Capnocytophaga canimorsus* infections are rare, they can lead to fulminant septic shock, as observed in this case [29]. Cultures obtained after bedside drainage of a COVID-19 patient grew *Parvimonas Micra*, underscoring a rare instance of *Parvimonas Micra*-associated pleural empyema complicating COVID-19 pneumonia [30].

To our best knowledge, no previous studies have examined the alternation of *Treponema*, *Porphyromonas* and *Amnipila* in COVID-19 positive individuals or reported any relationship between them and COVID-19 infection.

In summary, supported by adequate cohort studies, *Selenomonas*, *Neisseria*, *Prevotella*, *Veillonella* and *Rothia* are considered to be critical microbial biomarkers responsible for causing asymptomatic or moderately symptomatic COVID-19. Conversely, although *Capnocytophaga* and *Parvimonas* are ranked as top features in the studied samples, occasional cases reporting them in relation to COVID-19 do not constitute compelling evidence to demonstrate their significance. Along with *Treponema*, *Porphyromonas* and *Amnipila* which were not studied in relation to COVID-19 infection, these microbial biomarkers need to be further validated for their relationships to COVID-19 infection in the future.

V. Conclusion and Future Works

This study developed a Machine Learning-based approach using 16s rRNA gene data for classifying asymptomatic and moderately symptomatic instances of COVID-19 patients. The RF model achieved commendable performance in classification and interpretation. Critical microbial biomarkers that align with findings from existing studies have been identified.

Nevertheless, our work has certain limitations. The dataset is notably of limited size and imbalanced, presenting a barrier for the model to achieve better performance. Certain ranked microbial biomarkers lacking support from other studies are not convincing enough to exhibit their relationships to COVID-19 infection. Although the RF model has been a favoured solution in the industry and research field, the performance of other ML models also warrants verification in this study context. In the future, we will use multiple datasets to benchmark mainstream ML models across various tasks, focusing on addressing imbalanced data categories. Currently, we solely employ the abundance to train the model. There is still potential for enhancement by enriching and integrating additional features. In the future, diverse types of features could potentially be derived and integrated. Models that comprehensively assimilate information from various types of features could be purposefully developed. Consequently, the outcomes of model interpretation would also derive greater biological relevance, offering insights that align more coherently with the underlying biological context, which would potentially pave the way for personalized medicine, predictive diagnostics, and targeted treatments.

REFERENCES

[1] Z. Gao, Y. Xu, C. Sun, X. Wang, Y. Guo, S. Qiu, and K. Ma, "A systematic review of asymptomatic infections with covid-19," *Journal of Microbiology, Immunology and Infection*, vol. 54, no. 1, pp. 12–16, 2021.

[2] W. Gao, J. Lv, Y. Pang, and L.-M. Li, "Role of asymptomatic and pre-symptomatic infections in covid-19 pandemic," *bmj*, vol. 375, 2021.

[3] Z. Ren, H. Wang, G. Cui, H. Lu, L. Wang, H. Luo, X. Chen, H. Ren, R. Sun, W. Liu *et al.*, "Alterations in the human oral and gut microbiomes and lipidomics in covid-19," *Gut*, vol. 70, no. 7, pp. 1253–1265, 2021.

[4] T. Zuo, H. Zhan, F. Zhang, Q. Liu, E. Y. Tso, G. C. Lui, N. Chen, A. Li, W. Lu, F. K. Chan *et al.*, "Alterations in fecal fungal microbiome of patients with covid-19 during time of hospitalization until discharge," *Gastroenterology*, vol. 159, no. 4, pp. 1302–1310, 2020.

[5] T. Zuo, Q. Liu, F. Zhang, G. C.-Y. Lui, E. Y. Tso, Y. K. Yeoh, Z. Chen, S. S. Boon, F. K. Chan, P. K. Chan *et al.*, "Depicting sars-cov-2 faecal viral activity in association with gut microbiota composition in patients with covid-19," *Gut*, vol. 70, no. 2, pp. 276–284, 2021.

[6] T. Zuo, F. Zhang, G. C. Lui, Y. K. Yeoh, A. Y. Li, H. Zhan, Y. Wan, A. C. Chung, C. P. Cheung, N. Chen *et al.*, "Alterations in gut microbiota of patients with covid-19 during time of hospitalization," *Gastroenterology*, vol. 159, no. 3, pp. 944–955, 2020.

[7] Y. K. Yeoh, T. Zuo, G. C.-Y. Lui, F. Zhang, Q. Liu, A. Y. Li, A. C. Chung, C. P. Cheung, E. Y. Tso, K. S. Fung *et al.*, "Gut microbiota composition reflects disease severity and dysfunctional immune responses in patients with covid-19," *Gut*, vol. 70, no. 4, pp. 698–706, 2021.

[8] S. Gu, Y. Chen, Z. Wu, Y. Chen, H. Gao, L. Lv, F. Guo, X. Zhang, R. Luo, C. Huang *et al.*, "Alterations of the gut microbiota in patients with coronavirus disease 2019 or h1n1 influenza," *Clinical Infectious Diseases*, vol. 71, no. 10, pp. 2669–2678, 2020.

[9] X. Xu, W. Zhang, M. Guo, C. Xiao, Z. Fu, S. Yu, L. Jiang, S. Wang, Y. Ling, F. Liu *et al.*, "Integrated analysis of gut microbiome and host immune responses in covid-19," *Frontiers of medicine*, vol. 16, no. 2, pp. 263–275, 2022.

[10] F. H. Yagin, İ. B. Cicek, A. Alkhateeb, B. Yagin, C. Colak, M. Azzeh, and S. Akbulut, "Explainable artificial intelligence model for identifying covid-19 gene biomarkers," *Computers in Biology and Medicine*, vol. 154, p. 106619, 2023.

[11] S. Ke, S. T. Weiss, and Y.-Y. Liu, "Dissecting the role of the human microbiome in covid-19 via metagenome-assembled genomes," *Nature Communications*, vol. 13, no. 1, p. 5235, 2022.

[12] L. H. Nguyen, D. Okin, D. A. Drew, V. M. Battista, S. J. Jesudasen, T. M. Kuntz, A. Bhosle, K. N. Thompson, T. Reinicke, C.-H. Lo *et al.*, "Metagenomic assessment of gut microbial communities and risk of severe covid-19," *Genome Medicine*, vol. 15, no. 1, p. 49, 2023.

[13] C. Merenstein, G. Liang, S. A. Whiteside, A. G. Cobián-Güemes, M. S. Merlino, L. J. Taylor, A. Glascock, K. Bittinger, C. Tanes, J. Graham-Wooten *et al.*, "Signatures of covid-19 severity and immune response in the respiratory tract microbiome," *MBio*, vol. 12, no. 4, pp. 10–1128, 2021.

[14] A. J. Armstrong, D. B. Horton, T. Andrews, P. Greenberg, J. Roy, M. L. Gennaro, J. L. Carson, R. A. Panettieri, E. S. Barrett, and M. J. Blaser, "Saliva microbiome in relation to sars-cov-2 infection in a prospective cohort of healthy us adults," *EBioMedicine*, vol. 94, 2023.

[15] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, p. 927312, 2022.

[16] L. Romani, F. Del Chierico, G. Macari, S. Pane, M. V. Ristori, V. Guarrasi, S. Gardini, G. R. Pascucci, N. Cotugno, C. F. Perno *et al.*, "The relationship between pediatric gut microbiota and sars-cov-2 infection," *Frontiers in cellular and infection microbiology*, vol. 12, p. 908492, 2022.

[17] Z. Zhang, Y. Tian, L. Bai, J. Xiahou, and E. Hancock, "High-order covariate interacted lasso for feature selection," *Pattern Recognition Letters*, vol. 87, pp. 139–146, 2017.

[18] Y. Sun, B. Chain, S. Kaski, and J. Shawe-Taylor, "Correlated feature selection with extended exclusive group lasso," *arXiv preprint arXiv:2002.12460*, 2020.

[19] J. R. Beattie and F. W. Esmonde-White, "Exploration of principal component analysis: deriving principal component analysis visually using spectra," *Applied Spectroscopy*, vol. 75, no. 4, pp. 361–375, 2021.

[20] J. T. Wassan, H. Wang, and H. Zheng, "Developing a new phylogeny-driven random forest model for functional metagenomics," *IEEE Transactions on NanoBioscience*, 2023.

[21] Y. Qi, "Random forest for bioinformatics," *Ensemble machine learning: Methods and applications*, pp. 307–323, 2012.

[22] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar *et al.*, "Reproducible, interactive, scalable and extensible microbiome data science using qiime 2," *Nature biotechnology*, vol. 37, no. 8, pp. 852–857, 2019.

[23] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "Dada2: High-resolution sample inference from illumina amplicon data," *Nature methods*, vol. 13, no. 7, pp. 581–583, 2016.

[24] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.

[25] A. K. Feehan, R. Rose, D. J. Nolan, A. M. Spitz, K. Graubics, R. R. Colwell, J. Garcia-Diaz, and S. L. Lamers, "Nasopharyngeal microbiome community composition and structure is associated with severity of covid-19 disease and breathing treatment," *Applied Microbiology*, vol. 1, no. 2, pp. 177–188, 2021.

[26] J. Liu, S. Liu, Z. Zhang, X. Lee, W. Wu, Z. Huang, Z. Lei, W. Xu, D. Chen, X. Wu *et al.*, "Association between the nasopharyngeal microbiome and metabolome in patients with covid-19," *Synthetic and Systems Biotechnology*, vol. 6, no. 3, pp. 135–143, 2021.

[27] A. Hernández-Terán, F. Mejía-Nepomuceno, M. T. Herrera, O. Barreto, E. García, M. Castillejos, C. Boukadida, M. Matias-Florentino, A. Rincón-Rubio, S. Avila-Rios *et al.*, "Dysbiosis and structural disruption of the respiratory microbiota in covid-19 patients with severe and fatal outcomes," *Scientific reports*, vol. 11, no. 1, p. 21297, 2021.

[28] C. Merenstein, F. D. Bushman, and R. G. Collman, "Alterations in the respiratory tract microbiome in covid-19: current observations and potential significance," *Microbiome*, vol. 10, no. 1, p. 165, 2022.

[29] E. C. Meyer, S. Alt-Epping, O. Moerer, and B. Büttner, "Fatal septic shock due to capnocytophaga canimorsus bacteremia masquerading as covid-19 pneumonia-a case report," *BMC Infectious Diseases*, vol. 21, pp. 1–6, 2021.

[30] S. Gumbs, I. Kwentoh, E. Atiku, W. Gikunda, and A. Safavi, "Parvimonas micra: A rare cause of pleural empyema with covid-19 co-infection," *Cureus*, vol. 16, no. 1, 2024.