Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 239 (2024) 1452–1459

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2023

ML-based predictive gut microbiome analysis for health assessment

Manel Gil Sorribes[a], Gabriele Leoni[b], Antonio Puertas Gallardo[b], Mauro Petrillo[c], Sergio Consoli[b], Vicenç Gómez[a], Mario Ceresa[b]*

[a]*Universitat Pompeu Fabra, Barcelona, Spain*
[b]*European Commission, Joint Research Centre (JRC), Ispra (VA), Italy*
[c]*Seidor Italy SRL, 20219 Milano, Italy*

Abstract

Personalised medicine is a rapidly evolving field to which many resources have been devoted recently. It represents a paradigm shift from a one-size-fits-all approach to healthcare, focusing instead on tailoring treatments and diagnoses to individual patients. This study aims to contribute to this transition by leveraging recent advancements in microbiome research. An earlier study that computed the Gut Microbiome Health Index (GMHI), a potent indicator capable of predicting disease presence with approximately 70% of accuracy, serves as the foundation for this research. Positive values of the GMHI are associated with healthy subjects and negative values to non-healthy ones. The objective of this study is twofold: firstly, to advance the use of the GMHI through the application of two distinct machine learning techniques, a Fully Connected Neural Network and an Autoencoder, secondly, to adapt the GMHI for a unique dataset of COVID-19 patients. The employment of these two neural network architectures facilitated an enhancement in predictive accuracy to approximately 74.5%, surpassing the GMHI baseline accuracy of 70.95%. Simultaneously, the application of the GMHI, recalibrated using species specifically identified in the COVID-19 dataset, demonstrated a substantial increase in accuracy by 17% (achieving an accuracy of 76%). These promising results in distinguishing COVID-19 patients, highlight the potential and broad applicability of this approach in the sphere of personalized medicine and directly relating the COVID-19 disease to one's microbiome and its composition.

Keywords: GMHI, Gut Microbiome, Metagenomics, Machine Learning, Neural Network, COVID-19

* Corresponding author. Tel.: +39 0332 78 9634;
E-mail address: mario.ceresa@ec.europa.eu

10.1016/j.procs.2024.06.318

1. Introduction

The gut microbiome is the totality of microorganisms, bacteria, viruses, protozoa, and fungi, and their collective genetic material present in the gastrointestinal tract [1]. Several works described correlations between gut microbiome composition and the occurrence of diseases [2-5] and, more recently, Gupta et al. [6] introduced the Gut Microbiome Health Index (GMHI) to translate the complex microbial species present in patients into a useful indicator of an individual's health status.

There has been a general interest in applying Machine Learning (ML) techniques such as Support Vector Machine (SVM) and Neural Networks (NN) to improve the analysis of gut microbiome [5,7,8] and in this study we aim to shed light on the SARS-CoV-2 influence on the gut microbiome [9,10]. Gaining these insights could improve patient management and potentially foster the creation of new therapeutic strategies [11,12].

Two primary objectives guide the study: enhance the GMHI predictive accuracy utilizing ML algorithms and apply a rescaled GMHI model to a COVID-19 dataset to assess patient health conditions and pinpoint specific diseases. By achieving these objectives, this research contributes to the advancement of personalized medicine and enhances our understanding of the gut microbiome's role in health and disease [13].

The structure of this contribution is as follows. In Section 2 we describe the used methodologies to advance the use of the GMHI, and to adapt it to the dataset of COVID-19 patients. Our computational experiments are reported in Section 3 with discussions related to four distinguishable experiments and conclusions and future works in Section 4.

2. Methodology

GMHI: The formula used to calculate the GMHI is as follows [6]:

$$h_{i,M_H,M_N} = \log_{10}\left(\frac{\frac{R_{M_H}}{|M_H|}\sum_{j\in I_{M_H}}|n_j\ln(n_j)|}{\frac{R_{M_N}}{|M_N|}\sum_{j\in I_{M_N}}|n_j\ln(n_j)|}\right)$$

Where $R_{M\_N}/ |M_H|$ represents the richness of health-prevalent species divided by the maximum richness that can be obtained by M_H species in a particular microbiome sample. The term, $\Sigma_{j\in I\_M\_H}$ $|n_j \ln(n_j)|$, represents the geometric mean, where I_{M_H} is the index set of M_H, and n_j represents the relative abundance of species j in I_{M_H}. The identical rationale applies to the denominator as well, but it relates to health-scarce species.

In the study the number of species (initially exceeding 800) was strategically pared down to a targeted group of 50 species, identified as significant for differentiating between healthy and non-healthy patients. This reduction, grounded on a prevalence-based strategy, allowed for a clearer distinction between health conditions, thereby providing valuable insights into their correlation with the selected microbial species.

Metagenomic Analysis Using the HTCondor Pipeline: The research involved a metagenomic analysis of 426 samples, obtained as a subset of the 4,347 samples from the validation dataset of [6]. These samples were examined using the MetaPhlAn software[†], a tool designed for profiling the composition of microbial communities from metagenomic shotgun sequencing data. To manage the large computational requirements of processing these samples, HTCondor high-throughput computing software was employed. This software provides a job queueing mechanism, scheduling policy, resource monitoring, and resource management. The results generated by MetaPhlAn for each individual sample were then merged and processed into a single .txt file. Subsequently, this file was used to calculate the GMHI as detailed in [6]. The robustness of the pipeline was tested with two different versions of the MetaPhlAn software, namely MetaPhlAn2 (version 2) and MetaPhlAn4 (version 4), to compare their efficacy and accuracy. MetaPhlAn was run with default parameters.

[†] https://huttenhower.sph.harvard.edu/metaphlan/

Variable Analysis and Species Identification: The dataset, comprising 4,347 samples from the original GMHI study, formed the basis for this part of the investigation. With a significant number of species found in these samples, the primary challenge was to discern the most influential species and effectively manage the high dimensionality.

The process for identifying the most representative species involved calculating the Analysis of Variance (ANOVA) F-values for all species and arranging them in descending order. Species with ANOVA F-values exceeding 10% of the top species' value were retained, yielding a total of 130 species. These species were then paired amongst themselves, and each pair was tested using a SVM to pinpoint the pair that could predict with the highest accuracy.

In order to manage the high-dimensional nature of the data, the Uniform Manifold Approximation and Projection (UMAP) was utilized, proving more effective than traditional dimensionality reduction methods [14]. In the next phase, a process was executed to determine the optimal number of UMAP dimensions to retain for accurate predictions. A consistent input of 558 species that had an ANOVA F-value greater than 1, was used for this operation. The UMAP method was used iteratively, where each iteration adjusted the number of UMAP dimensions. These adjusted dimensions were then used to train the SVM algorithm to calculate accuracy. This sequence of steps was repeated, progressively fine-tuning the number of UMAP dimensions until the highest achievable accuracy was reached. This methodology was crucial in establishing the ideal balance between the reduction of UMAP dimensions and prediction accuracy.

Simultaneously, a SVM classification algorithm was initiated with 558 species, selected based on the previous criteria. Throughout each subsequent iteration, the last 50 species were eliminated until an optimal combination was found that yielded the highest accuracy and F1 score. This process helped establishing the optimal number of species to retain for the most accurate results. Additionally, hyperparameter tuning was integrated during this phase to prevent any correlation between different runs, thus ensuring the reliability of the results.

Implementing Neural Networks: In the third experiment of the study, the focus was on fitting a NN model for improving accuracy on the GMHI. For the training of these NN models, inputs that optimally differentiated between healthy and non-healthy groups were utilized. This included the results obtained from the previous section: the top 300 species ordered by descending ANOVA F-values, four UMAP variables, and the original GMHI index computed from the samples.

For the creation and definition of all NN features, popular machine learning libraries, such as TensorFlow, sklearn, pandas, and numpy, were utilized.

Two types of NN models were explored in this study, namely the Fully Connected (FC) and the Autoencoder (AE) models [15]. The AE was primarily used for dimensionality reduction by capturing and retaining the most significant features from the input data. Conversely, the FC network was designed to examine and process all features provided as input directly. Both types of networks processed three different inputs - the GMHI, UMAP, and microbial species abundance data - simultaneously, using them as a unified tensor.

For adaptively adjusting the learning rate from an initial rate, the Adam optimizer was employed. Considering the binary nature of the target variable, the binary cross-entropy function [16] was chosen as the model's loss function. Training was conducted up to a set maximum of epochs, with an early stopping mechanism activated if the validation loss did not decrease for 50 consecutive epochs. Accuracy served as the primary metric for model evaluation.

Common features were implemented for both NN, including elements such as LeakyReLU as the activation function, a validation split of 0.1, and a batch size of 8. Other features were defined distinctly in Table 1:

Table 1. Singular features of the studied NN

Features	FC	AE
Hidden Layers	5	3 encoder layers
Learning Rate	10^{-5}	10^{-4}
Units Per Layer	305, 305, 305, 305, 305	305, 150, 50
Maximum Epochs	1000	2000

Architectures of both NN were represented using TensorFlow for enhanced understanding, and the model with the highest accuracy was saved after 20 iterations.

COVID-19 Dataset Analysis and GMHI Recalculation: Finally, an analysis was performed on a new dataset comprising patients either affected or unaffected by COVID-19, aiming at understanding the impact of COVID-19 on the gut microbiome. Data for this analysis was obtained from [9] where links for the 865 samples utilized are indicated.

MetaPhlAn2 and MetaPhlAn4 software were compared to ascertain which offered better precision in GMHI calculation for this specific dataset, using the reference GMHI set of species. This approach involved computing the accuracy and the p-value between the Control (healthy) and Covid (non-healthy) groups for each version, discerning as the best version the one with higher accuracy and lower p-value.

Significant species within this new dataset were determined using an ANOVA test: ordering by descending ANOVA F-value and retaining species with values greater than 1. To determine the association of each species with healthy or non-healthy patients, the mean concentration of each species was calculated and compared between the two groups. Species with a higher mean concentration in the COVID-19 group were labelled as non-healthy species, while species with higher mean concentration in the Control group were labelled as healthy species.

A recalculation of GMHI was conducted using different combinations of 50 selected species, with each combination including the highest-ranking species in each labelled group according to the ANOVA analysis. Specifically, the optimal criteria were found when keeping the top 20 to 35 species in the ordered list associated with non-healthy conditions. All possible combinations within this range were explored, adjusting the number of species related to healthy conditions. The method used for this combination exploration is clarified in the following table:

Table 2. Combinations studied for the COVID-19 dataset

Configuration step	Non-healthy species M_N	Healthy species M_H
1	First 20 species	First 30 species
2	First 21 species	First 29 species
…	…	…
15	First 34 species	First 16 species
16	First 35 species	First 15 species

Before providing a detailed account of the project's results, the following figure presents the pipeline that was employed. This illustration clarifies the simultaneous tasks executed during the research.

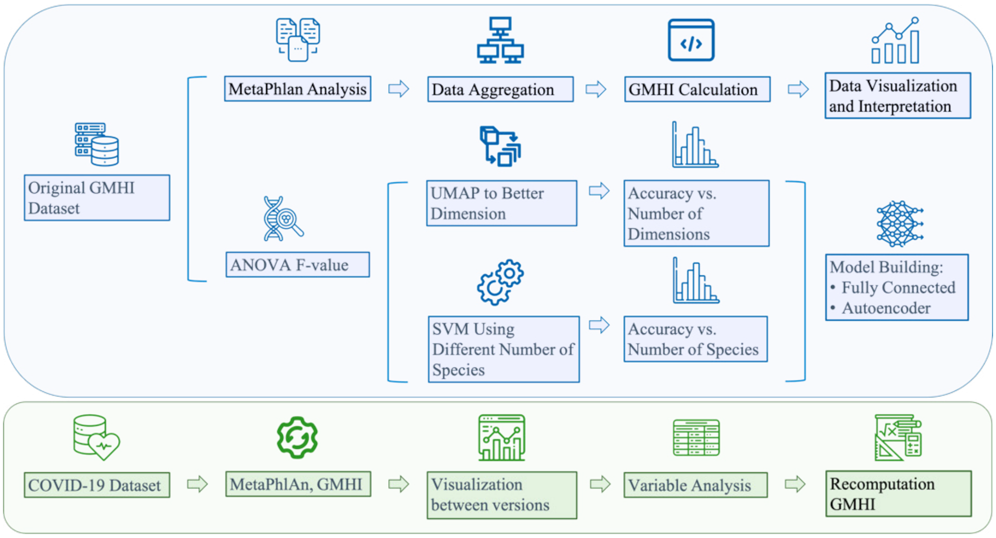

Fig. 1. Schematic representation of the experimental pipeline. Two parallel paths were pursued, each stemming from distinct source data: the top branch represents the original GMHI study data, while the bottom branch corresponds to a COVID-19 dataset.

3. Results

Metagenomic Analysis Using the HTCondor Pipeline: This research effectively utilized a HTCondor pipeline for metagenomic analysis from [6], specifically examining a subset of 426 samples drawn from the broader validation dataset used in the original study. A comparative graph of the GMHI index values obtained with MetaPhlAn2 and MetaPhlAn4 versions were compared, segregated by health condition.

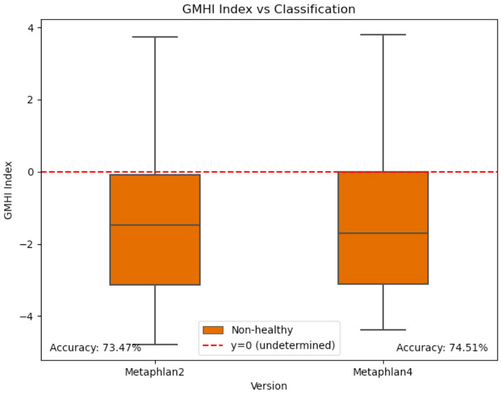

Fig. 2. Comparison of GMHI Index: This box plot represents the distribution of GMHI index values for both MetaPhlAn2 and MetaPhlAn4 softwares, categorized by health condition (all Non-healthy). The accuracy of classification for each method is also displayed. The red dashed line at y=0 serves as the boundary between Control and Non-healthy classifications, the differences in GMHI values across the two groups are not significative

The plot in Fig. 1 provided limited information. The accuracies aligned with those acquired by the authors of [6] across the entire validation dataset are similar for both versions. For this reason, a deeper analysis between both versions was specifically conducted when analysing the COVID-19 dataset.

Variable Analysis and Species Identification: The relation between microbial species and the patients' health conditions was examined using data from [6], which incorporated 4,347 labelled samples. Using the two most representative species (*Alistipes shahii* and *Bifidobacterium adolescentis*), a SVM algorithm was trained, achieving a predictive accuracy of 75%.

As shown in Fig.2a, the optimal number of 300 species can be used in subsequent sections by training SVM algorithms with varying species numbers and representing the accuracies obtained. In addition, the optimal dimensionality for UMAP reduction in this dataset is found to be 4, as shown in Fig.2b

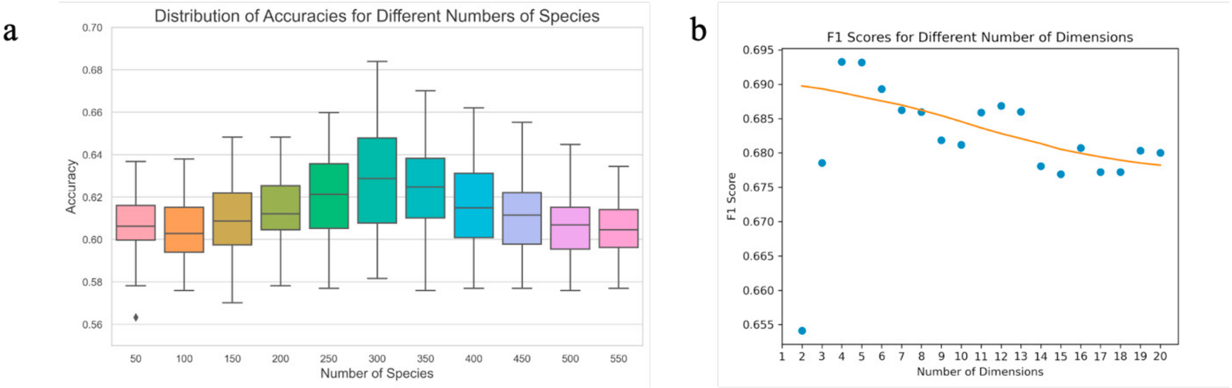

Fig. 3. Comparison of Accuracies and F1 Scores for Variable Species Numbers and Dimensions: (a) Boxplot showing the accuracies achieved when training an SVM algorithm with a varying number of species. (b) The F1 score of the SVM algorithm plotted against the number of dimensions to which UMAP reduced the dimensionality.

Implementing Neural Networks: FC and AE models were extensively tested across 20 iterations each, using 5-fold cross validation in each iteration to ensure robust and diverse results. Both the FC and AE models surpassed the baseline accuracy of the GMHI—70.95% when computed for the same samples—achieving similar mean accuracy rates of 74.57% and 74.34% respectively.

By analysing the accuracies achieved by each iteration due to the inherent variability of the cross validation, we observe that 50% of the results fell within the 73% to 75.5% accuracy range for NN models. The increment, though numerically small, is highly significant in practice, promising more reliable predictions, which could improve diagnoses, treatment outcomes, and overall health for patients with gut microbiome-related conditions.

Furthermore, the validation and training loss behaviours were examined throughout the epochs as shown in Fig 3. While the FC model indicated effective learning without overfitting, it displayed potential limitations in generalization. On the other hand, the AE model, despite faster convergence, showed potential generalization constraints, thus highlighting the inherent trade-off between model complexity, training duration, and generalization performance in NN.

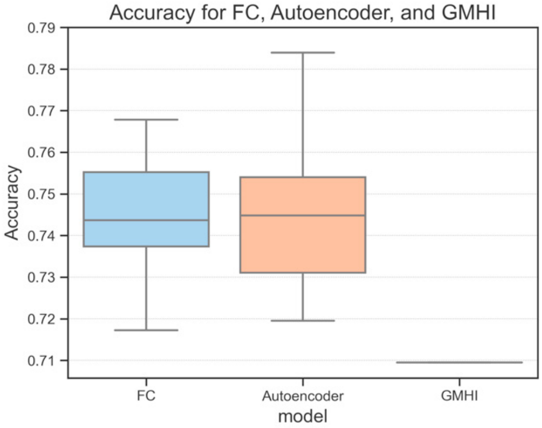

Fig. 4. Accuracy Comparison between models: Accuracies obtained when training the NN models and using the original GMHI.

COVID-19 Dataset Analysis and GMHI Recalculation: In the analysis of the COVID-19 dataset, the performance of the different versions of MetaPhlAn was assessed using the reference set of species. Boxplots were produced to display GMHI index distributions for both versions. While the accuracy of MetaPhlAn2 and MetaPhlAn4 was found to be similar (approximately 59%), MetaPhlAn4 performed slightly better when the *t*-statistics and associated *p*-values were examined, indicating a statistically significant distinction between COVID-19 and Control groups.

The overall accuracies were significantly lower than the authors' reported results. The discrepancy can be attributed to the lack of consideration of the overall health condition of the patients. Patients without COVID-19 may have incurred other health conditions that influenced the GMHI index, leading to misclassification.

Despite these limitations, the GMHI was successful in detecting COVID-19 patients, correctly identifying about 65% of the patients. However, the performance was not as strong in distinguishing patients without COVID-19.

Subsequently, an ANOVA test pipeline was employed for the dataset. The findings suggested an optimal configuration between 20 M_N and 30 M_N, since these configurations demonstrated higher accuracy and mean GMHI values that aligned with health conditions.

To confirm these findings, further investigation was undertaken to identify the optimal number of species for the best prediction. The configurations of 31 M_N and 32 M_N were found to be the most effective, offering superior performance in both accuracy and mean GMHI values.

Table 3. GMHI Calculation Variations Range: Comparison of accuracy and mean GMHI index values for different configurations. Showcases the variations in model performance and GMHI index values for Control and Covid cases

Version	Accuracy	Mean Control	Mean Covid
29 M_N	0.734	0.285	- 0.920
30 M_N	0.755	0.178	- 1.005
31 M_N	**0.756**	**0.198**	**- 0.996**
32 M_N	**0.774**	**0.054**	**- 1.102**
33 M_N	0.772	- 0.125	- 1.367

A comparison of the species identified in the COVID-19 dataset and the Gupta *et al.*, [6] reference set of species was conducted. The comparison revealed an overlap of 6 %. A high percentage (94 %) of unique species from the COVID-19 dataset (e.g., *Faecalibacillus intestinalis*, full list is available in Table S1[17]) were identified, suggesting the potential for unique microbial markers to be associated with COVID-19.

In Fig. 4 we present a final comparison of the GMHI index results obtained using MetaPhlAn4 and the selected 31 M_N approach showed a clear distinction between healthy and COVID-19 or Pneumonia conditions. The newly selected approach resulted with the Control group consistently above zero and the COVID-19 group consistently below zero. It also revealed correlations between COVID-19 and the microbiome data of patients, confirming that COVID-19 can influence the gut microbiome equilibrium.

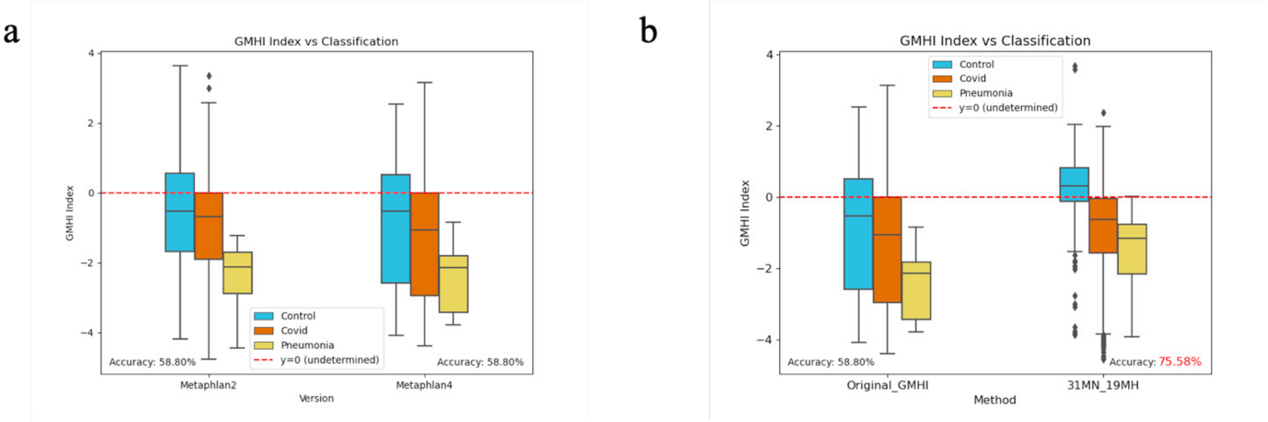

Fig. 5. Image (a) displays the GMHI Index values for MetaPhlAn2 and MetaPhlAn4 in Covid and non-Covid samples, while Image (b) contrasts the GMHI Index between the species selected by authors and those from the current methodology in COVID-19 and non-COVID-19 samples. Both images use boxplots to show data distribution and a red dashed line to indicate the healthy/non-healthy boundary.

4. Conclusions

In this study we presented a flexible pipeline for the combination of classical and ML methods in the analysis of gut microbiome and its relationship with general Health. We found that applying ML in microbiome studies presents unique challenges, primarily due to the variability of the bacteria populations and the complexities involved in sample processing. Additionally, acquiring high-quality, comprehensive datasets—crucial for effective model training—often proves to be complex and inaccessible. These difficulties are further amplified when it is difficult to retrieve harmonised associated metadata of patients, like in the case of individuals affected by COVID-19.

Significant correlations between a viral disease such as COVID-19 and alterations in the microbiome were confirmed. A high percentage of species showing substantial differentiation between COVID-19 patients and non-COVID-19 individuals, not identified by the authors of the dataset, suggests that additional unique microbial markers might be associated with COVID-19. Validation of these findings would requiring further investigations, which are outside the scope of this work. Nonetheless, rescaling of the GMHI methodology using these identified species resulted in a 17 percent point improvement in accuracy, underscoring the adaptability of GMHI for specific datasets and disease classifications.

The study also faced challenges in defining what constitutes a healthy microbiome and predicting healthy samples, as few health-related species were identified. Nevertheless, the research achieved approximately 75% accuracy in predicting the presence of COVID-19 or pneumonia. Potential risks of overfitting were offset by integrating nine independent projects into the study.

Even if the accuracy improvements provided by the two proposed NN were small, these could be substantial when applied to large scale population.

Future work includes refining our NN models with additional gut microbiome data, exploring alternative architectures, evaluating the recalibrated GMHI on a larger COVID-19 cohort, and assessing the rescaling pipeline's effectiveness in the context of other viral illnesses. Additionally, the biological implications of the significant species identified will be explored in relation to COVID-19 and the gut microbiome. These insights encourage future exploration into the rich information the microbiome can provide, potentially benefiting society at large.

References

[1] Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J.J., Tripathi, A., Brenner, D.A., Loomba, R., Smarr, L., Sandborn, W.J., Schnabl, B., Dorrestein, P., Zarrinpar, A., & Knight, R. (2019). Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. Clin Gastroenterol Hepatol, 17(2), 218-230.

[2] McBurney, M.I., Davis, C., Fraser, C.M., Schneeman, B.O., Huttenhower, C., Verbeke, K., Walter, J., & Latulippe, M.E. (2019). Establishing What Constitutes a Healthy Human Gut Microbiome: State of the Science, Regulatory Considerations, and Future Directions. The Journal of Nutrition, 149(11), 1882-1895.

[3] Hagerty, S.L., Hutchison, K.E., Lowry, C.A., & Bryan, A.D. (2020). An empirically derived method for measuring human gut microbiome alpha diversity: Demonstrated utility in predicting health-related outcomes among a human clinical sample. PLOS ONE, 15(3).

[4] Ruuskanen, M.O., Erawijantari, P.P., Havulinna, A.S., Liu, Y., Méric, G., Tuomilehto, J., Inouye, M., Jousilahti, P., Salomaa, V., Jain, M., Knight, R., Lahti, L., Niiranen, T.J. (2022). Gut Microbiome Composition Is Predictive of Incident Type 2 Diabetes in a Population Cohort of 5,572 Finnish Adults. Diabetes Care, 45(4), 811-818. doi: 10.2337/dc21-2358

[5] Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M.J., Gasbarrini, A., Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nature Reviews Gastroenterology & Hepatology, 17(10), 635-648. doi: 10.1038/s41575-020-0327-3

[6] Gupta, V.K., Kim, M., Bakshi, U., Cunningham, K.Y., Davis, J.M., Lazaridis, K.N., Nelson, H., Chia, N., Sung, J. (2020). A predictive index for health status using species-level gut microbiome profiling. Nature Communications, 11(1), 4635. doi: 10.1038/s41467-020-18476-8

[7] Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., Kawakami, K., Aikawa, Y., Oda, N., Arai, K., Matsunaga, A., Nonaka, M., Nakai, K., Shinmura, W., Matsumoto, M., Morishita, S., Takeda, A.K., Miwa, H. (2020). Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. J Clin Med, 9(8), 2491. doi: 10.3390/jcm9082491

[8] Vilne, B., Kibilds, J., Siksna, I., Lazda, I., Valciņa, O., Krūmiņa, A. (2022). Could Artificial Intelligence/Machine Learning and Inclusion of Diet-Gut Microbiome Interactions Improve Disease Risk Prediction? Case Study: Coronary Artery Disease. Frontiers in Microbiology, 13, 627892. doi: 10.3389/fmicb.2022.627892

[9] Ke, S., Weiss, S.T., & Liu, Y.Y. (2022). Dissecting the role of the human microbiome in COVID-19 via metagenome-assembled genomes. Nature Communications, 13(1), 5235.

[10] Karthikeyan, S., Levy, J.I., De Hoff, P., et al. (2022). Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature, 609(7925), 101-108. doi: 10.1038/s41586-022-05049-6

[11] Kashyap, P.C., Chia, N., Nelson, H., Segal, E., & Elinav, E. (2017). Microbiome at the Frontier of Personalized Medicine. Mayo Clin Proc, 92(12), 1855-1864.

[12] Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., Maathuis, M.H., Moreau, Y., Murphy, S.A., Przytycka, T.M., Rebhan, M., Röst, H., Schuppert, A., Schwab, M., Spang, R., Stekhoven, D., Sun, J., Weber, A., Ziemek, D., & Zupan, B. (2018). From hype to reality: data science enabling personalized medicine. BMC Medicine, 16(1), 150.

[13] Clerbaux, Laure-Alix, et al. "Mechanisms leading to gut dysbiosis in COVID-19: Current evidence and uncertainties based on adverse outcome pathways." *Journal of Clinical Medicine* 11.18 (2022): 5400.

[14] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).

[15] Thaler, Stefan, and Vlado Menkovski. "The role of deep learning in improving healthcare." *Data Science for Healthcare: Methodologies and Applications* (2019): 75-116.

[16] Ho, Yaoshiang, and Samuel Wookey. "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling." *IEEE access* 8 (2019): 4806-4813.

[17] Sorribes, M. G. et al. Table S1, Zenodo (2023). https://doi.org/10.5281/zenodo.8369884.