Journal of Knowledge Learning and Science Technology
ISSN: 2959-6386 (Online)
2024, Vol. 3, No. 4, pp. 35–47
DOI: https://doi.org/10.60087/jklst.v3.n4.p35

Research Article

Machine Learning Models for Predicting Susceptibility to Infectious Diseases Based on Microbiome Profiles

[1]Nasrullah Abbasi, [2]Nizamullah FNU, [3]Shah Zeb, [4]Muhammad Fahad , [5]Muhammad Umer Qayyum

[12,3,4,5] Washington University of Science and Technology, Alexandria, Virginia, USA

Abstract

The human microbiome comprises complex ecosystems of microorganisms inhabiting different parts of the body and plays a very important role in sustaining health and dictating disease vulnerability. On the basis of this continuous generation of data on the microbiome, interest is developing in their use for disease risk prediction. Machine learning provides an extremely robust way of doing so because of its ability to handle complex and high-dimensional data. In this research article, the authors compared the efficiency of random forest, support vector machines, and neural network machine learning models in predicting infectious diseases via a microbiome profile. This review provides a comprehensive overview of various studies that were published in the recent past that used these machine learning techniques for microbiome data analysis. It further assesses the degree to which each model captures the intrinsic complexity and variability of the microbiome, which holds the key to accurately predicting diseases. Moreover, this review highlights the importance of feature selection and data preprocessing in enhancing the performance of machine learning models. By selecting the most relevant features and properly preprocessing the data, one can train better models and hence make better predictions. Our results provide very good potential for machine learning models in predicting susceptibility to infectious diseases and, at the same time, show that there is indeed potential for further improvement. Multiomics data integration should increase predictive power—incorporation of microbiome data with other kinds of biological information. Model interpretability can be important for enhancing clinicians' understanding of and trust in the prediction, which is critical to the successful integration of these tools into truly personal healthcare.

Keywords

Microbiome, Machine Learning, Disease Prediction, Infectious Diseases, Random Forest, Support Vector Machine (SVM), Personalized Medicine

Introduction

Trillions of microorganisms inhabiting different human body sites form the human microbiome, which has an important role in health and is implicated in the genesis of disease. The greatest diversity of microbial communities resides

*Corresponding author: Nasrullah Abbasi

Email addresses:

nabbasi.studnet@wust.edu (Nasrullah Abbasi), Nizamullah.student@wust.edu (Nizamullah FNU), Szeb.student@wust.edu (Shah Zeb), fahad.student@wust.edu (Muhammad Fahad), qayyum.student@wust.edu (Muhammad Umer Qayyum)

Received: 12-07-2024; **Accepted:** 19-08-2024; **Published:** 25-08-2024

in the gut, skin, mouth, and other mucosal surfaces. These microbial communities have a symbiotic relationship with the host and contribute to vital physiological processes such as digestion, immune modulation, and pathogen resistance. Specifically, the composition and diversity of the microbiome are determined by a few parameters, such as genetics, diet, environment, and lifestyle, leading to unique microbial signatures in individuals (Ren et al., 2022). Current research has shed more light on the deep effects of microbiome imbalance—so-called dysbiosis—on the initiation and course of a number of diseases, most notably infectious diseases. Dysbiosis modulates the host immune response and predisposes individuals to infections; it can also impact the severity and outcome of diseases caused by pathogenic microorganisms. Certain changes in the gut microbiota lead to an increased risk of Clostridioides difficile infection, a very prominent cause of morbidity and mortality, especially among hospitalized patients. Periodontitis is related to shifts in the oral microbiota and might exacerbate some systemic conditions, such as cardiovascular disease.

As such, with the critical role of the microbiome in health and disease, increasing interest has been given to the use of microbiome profiles in predicting susceptibility to infectious diseases. In addition to offering very early points of detection and prevention, this approach opens up avenues for personalized medicine where interventions can be based on an individual's microbiome composition.

Role of Machine Learning in Analyzing Complex Biological Data

Owing to the high dimensionality and complexity of microbiome data, classical statistical analysis is challenging. Each host, be it human or otherwise, harbors thousands of microbial species whose relative abundances differ enormously between and within hosts over time. In addition to this already enormous level of complexity, an extra layer of intricacy arises from interactions between microbes and the immune system of their host, which can hardly be decoded in the traditional way of doing things. One of the very strong tools for analyzing such complex, multidimensional biological data is machine learning, a subdomain of artificial intelligence. Machine learning algorithms learn from large datasets to identify patterns, classify data, and make accurate predictions, making them of particular relevance in microbiome research. These models can capture intrinsic variability in microbiome data to identify critical microbial signatures associated with disease and provide high-accuracy predictions of outcomes.

Machine learning models have been applied to predict an individual's susceptibility to infectious diseases from the profile of their microbiome. These models create predictive signatures from abundance and diversity values for certain microbial taxa, yielding predictive biomarkers of increased risk for infection in a person. Furthermore, machine learning can enable the discovery of new microbial interactions and pathways contributing to disease pathogenesis, opening new avenues for therapeutic intervention.

Purpose and Scope of the Article

This paper presents a review of the status of machine learning models for susceptibility to infectious diseases on the basis of their microbiome profiles. This section discusses the types of machine learning algorithms used in this field, challenges with microbiome data, and strategies employed to address such challenges. Building on a number of recent case studies and examples, the following article will identify effective applications in the use of machine learning for infectious disease prediction, including what is possible and what has its limits. It will also provide a comparative analysis of different machine learning models with respect to their performance, accuracy, and appropriateness for different types of microbiome data. Furthermore, the challenges and limitations within the field pertaining to data heterogeneity, model interpretability, and requirements for large, good-quality datasets are discussed.

Finally, the review will cover current trends and future directions in this rapidly evolving field. This will include views on how to meaningfully integrate machine learning with other omics data, the role of microbiome-based predictions in personalized medicine, and ethics arising from the application of microbiome data in clinical settings. In this context, the authors, through this review, strive to attract the attention of researchers, clinicians, and policymakers to the current capacities and future prospects of machine learning in leveraging microbiome data to predict susceptibility to infectious diseases.

Background

A. Microbiome and Its Relationship with Health

The human microbiome considers a collective genome of all the microorganisms living in the human body. It is a very complex ecosystem of bacteria, viruses, fungi and other microorganisms that have already colonized areas in the body, such as the gut, skin, oral cavity and respiratory tract. To date, the greatest emphasis has been placed on the gut microbiome, which takes the first stance in a wide array of physiological processes, including digestion and metabolism, the modulation of the immune system, and protection against pathogenic bacteria. The microbiome composition is very person-specific and depends on genetic, dietary, environmental, and lifestyle factors. On the other hand, a normal microbiome would connote diversity and a balanced microbial community that reflects good health in its general aspects. It is suspected that the condition of dysbiosis is an imbalance in the microbiome, leading to inflammatory bowel diseases, obesity, diabetes, and mental disorders.

Microbiome Profiles in Relationships with Infectious Diseases

In recent years, the abovementioned mechanisms of infectious diseases have emerged because of the significant contributions of the human microbiome. One basis for the association was the discovery of interactions between the microbiome and the host immune system. A balanced microbiome thus enhances the potential of the host immune system to ward off infection, and an imbalanced microbiome impairs immune function and renders the host very susceptible to infection. Some microbiomes have been shown to lead to general susceptibility to many infectious diseases. For example, alterations in the gut microbiota composition are associated with an increased risk of Clostridioides difficile and Escherichia coli infection in the gut. Similarly, disrupted respiratory microbiomes have been linked to an increased risk of respiratory infections such as influenza and pneumonia. These findings indicate the potential of the microbiome with respect to vaccines and the treatment of infectious diseases. For example, a divergent gut microbiome has been shown to be associated with immune responsiveness to vaccination and its effectiveness. Therefore, recently, research has focused on exploiting microbiome profiles in the prediction and mitigation of potential risk with individual interventions.

B. Machine Learning in Health Care

Overview of Applications of Machine Learning in Health Care

ML, a subdomain of artificial intelligence, has now emerged as a very powerful tool in health care, with new avenues for analyzing complex datasets, identifying patterns, and making predictions. Only a handful of ML algorithms have been applied to many tasks in the domain of health care, ranging from diagnostic imaging and drug discovery to personalized medicines and predictive analytics. The ability to process and analyze volumes of data beyond any high dimensionality and heterogeneity, be it from EHRs, genetic information, and medical imaging, which lies within healthcare, is important for realizing the key strengths of ML. Learning from all these data, hence making a prediction or identification of disease biomarkers, or even therapy plans recommended to the individual patient, lies in the hands of ML algorithms. Within the microbiome research community, ML has helped to discover the complex interrelations between human health and microbiome profiles. Specifically, ML applied to microbiome data may be able to localize microbial signatures for specific diseases, predict disease risk, and explore possibilities for microbiome-based therapy.

Previous Work in Disease Susceptibility Prediction

Another new application of ML is in susceptibility prediction for infectious diseases from microbiome profiles. Preliminary evidence suggests that ML algorithms could be used to predict the risk of a number of diseases on the basis of the composition and diversity of the microbiota. For example, machine learning models have been developed to predict Clostridioides difficile infection from gut microbiome data, wherein bacterial taxa that are either protective or predictive of CDI can be identified, leading toward the development of predictive tools that inform preventive strategies. Similarly, ML has also been applied in trying to predict, on the basis of upper respiratory tract microbiome analysis, which individuals are likely to develop respiratory infections. Some studies have demonstrated that specific patterns of microbes in one's nasal microbiome are able to predict whether one will develop respiratory infections, such as flu. Therefore, in addition to infectious diseases, ML models can already be applied for the prediction of risk for susceptibility to noninfectious diseases with known microbiome associations, which include inflammatory bowel disease and colorectal cancer. In these studies, the models using microbiome data and other biological data also differentiated high-risk individuals, thus suggesting an opportunity for earlier intervention and improved management of disease. In this direction, the further integration of machine learning into microbiome research holds very good potential to grow into an increasingly personalized approach involving healthcare strategies on the basis of an individual's microbiome profile, for instance, highly precise predictions of susceptibility to diseases and their targeted interventions, which ultimately result in improved patient outcomes.

Machine Learning Models

The use of machine learning models has substantially impacted our ability to analyze and interpret intricate biological data in general and to study the human microbiome in particular. The human body bears a large and highly complex number of different microbial communities distributed spatially across the body. The relatively dynamic and diverse microbial ecosystems interact not only with each other but also with the host in ways that can influence an individual's susceptibility to a variety of infectious diseases. Understanding these interactions is essential for improving medical science, but the complexity and volume of microbiome data present challenges for traditional analytical approaches. This is where machine learning models take over. ML might be applicable for the disclosure of cryptic patterns in microbiome profiles, which are otherwise out of reach of classic approaches. Such profiles could include, but are not limited to, bacterial population data regarding species, relative abundance, gene expression, etc., making them key contenders for data rich in information relevant to biomarkers of disease susceptibility. ML models work very well because the data are high-dimensional, they identify a nonlinear relationship, and on this basis, the prediction is able to find signals in small variations in microbial composition. The use of ML models in microbiome research opens several opportunities for personalized and precision medicines. This means that predictive analytics of individual risk predispositions to develop certain infectious diseases allow health professionals to implement the right interventions to avoid the manifestation of the disease. Notably, this approach primarily benefits patients and facilitates cost optimization in health care by shifting the focus to prevention and early intervention. In the following sections, we discuss three of the most widely used ML models in the realm: random forest, support vector machines, and neural networks. These models each have their own specifics and properties applied for the purpose of microbiome data analysis. Together, these models combine random forests for the ability to manage high-dimensional data and missing data, support vector machine robustness in high-dimensional space, and neural networks' powerful pattern recognition capability in predicting disease susceptibility for researchers to harness the maximal power of microbiome data.

Random forest

The random forest is an ensemble learning approach that trains on many decision trees. Each tree of the method trains with a random selection of training samples, and the last output is the average or mode of class labels from many trees. Usually, a majority voting rule is used for aggregating the class label. This ensemble model lessens the variance of the model in such a way that if it was on a single decision tree, it would be prone to overfitting.

Key features

Ensemble Method: RF combines the predictions from multiple decision trees, enhancing overall model stability and accuracy.

Bootstrap aggregation (bagging): RF uses bootstrapping to generate multiple training sets, further enhancing model performance by reducing overfitting.

Random feature selection: The RF selects a random sample of a feature subset at the construction of every tree; thus, the model does not rely on one feature too much, and there is increased diversity between the trees.

RF performs very well when trained on microbiome data, most likely because data are often high dimensional, consisting of many more features than samples. Microbiome datasets seem to have a very large number of variables, and most of the features can have either complicated or nonlinear relationships with the outcome of the disease. One of the most appealing characteristics of RF lies in its ability to manage such complexity effectively without overfitting. Random forest can handle missing data effectively, which is very common when microbiome data are used, through the use of surrogate splits or the construction of trees with only a fraction of the available data. Although RF is an ensemble method, it maintains some interpretability and can even provide help in ranking features through an important procedure that has particular importance during the identification of microbial taxa that are key to being associated with disease susceptibility.

Support Vector Machine (SVM)

Support vector machines (SVMs) can be defined as an excellent class of models for supervised learning. SVM identifies the most suitable hyperplane to separate the classes in the feature space. The optimum hyperplane can be defined as the hyperplane that maximizes the margin between the closest points of the classes, which are well known as support vectors. SVM is able to perform linear and nonlinear data by actually using kernel functions; therefore, the original feature space is mapped onto a dimension where linear separation is possible.

Patient characteristics

Margin Maximization: SVM maximizes the margin through the hyperplane to achieve better generalization performance on unseen data.

Kernel Trick: SVMs can be endowed with the ability to work in a transformed attribute space via different kernel functions, such as linear, polynomial, and radial basis functions. This makes the SVM capable of working in the solution to very tight and deep nonlinear problems.

Robustness: SVMs are not as sensitive to overfitting, especially in high-dimensional spaces. This makes the SVM equally attractive for cases in which the feature space is very large compared with the observations.

Applicability to microbiome data

SVMs are outliers in the development of models applied to microbiome data analysis owing to their compatibility with high-dimensional data for analysis, in which the number of features is much greater than the number of samples. This added advantage becomes significant in situations where the relationship between the microbiome profile and susceptibility to disease is nonlinear. SVMs can be customized to handle imbalanced data. This is common in microbiome datasets if some outcomes are quite rare, such as presenting the existence of a disease. The flexibility of SVMs in choosing diverse kernel functions helps SVMs adjust to the idiosyncratic feature space of microbiome data that otherwise would be inapplicable for linear models.

Neural Networks (NNs)

When layers of nodes interconnect with each other, neural networks model human brain-like structures. Each node, or neuron, processes the input data and sends the result further on to the next layer of nodes. They are good at determining patterns and obtaining complex, often nonlinear relationships between variables. The simplest form of a neural network is a feedforward network, in which information travels in just one direction: through the input, through some hidden layers, to the output layer. Other more complex architectures that have been introduced are the convolutional neural network and the recurrent neural network; these architectures solve very specific types of problems. These types of data include image, text, and sequence data.

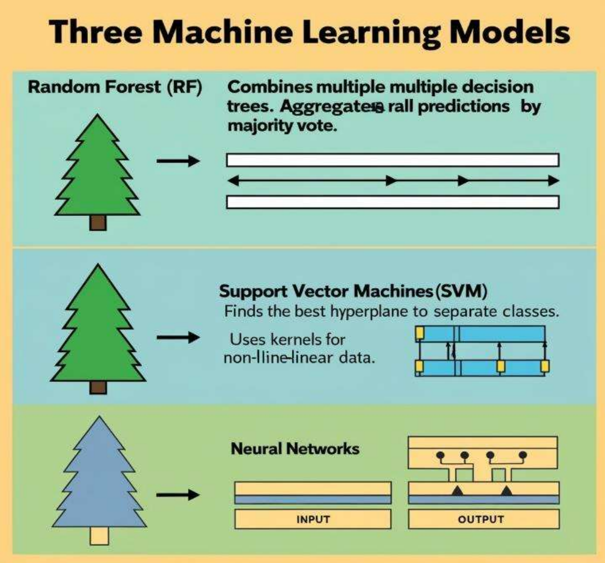
Layered: Artificial NNs inherently consist of an input layer, one or more hidden layers, and an output layer such that every layer performs a linear or nonlinear transformation of the data.

Nonlinear modeling: NNs are able to learn and model complex nonlinear interactions in the features themselves, thus being appropriate for capturing intricate patterns within biological data.

Backpropagation: NN training involves adjusting the weights of connections between neurons via backpropagation to minimize the error between the predictions and the actual values.

Suitability for Microbiome Data

NNs are particularly well suited for modeling complex, nonlinear interrelationships, which are usually present in microbial data. With the diversity of microbial communities being so large, many complicated patterns and associations might be missed, although simpler models can do that. Depending on the size and extent of the study, large amounts of data and considerable computational resources are needed, which can be considered limiting in some smaller studies. NNs work best when large amounts of data are available for analysis, which is not always easy to procure for microbiome studies. However, when sufficient data are available, NNs offer unrivaled accuracy in the prediction of disease susceptibility. Computationally, the training of NNs is intensive and requires specialized hardware, particularly GPUs, and software frameworks such as TensorFlow and PyTorch. This makes it computation-ready with NNs in large-scale studies or projects with access to robust computational resources.


Three Machine Learning Models
Random Forest (RF)
Combines multiple multiple decision trees. Aggregates rall predictions by majority vote.
Support Vector Machines (SVM)
Finds the best hyperplane to separate classes.
Uses kernels for non-lline-linear data.
Neural Networks
INPUT
OUTPUT

Feature Selection and Data Preprocessing

The real marvel in the achievable success that machine learning models realize in predicting susceptibility to infectious diseases from microbiome profiles must only be that features are well selected and that data preprocessing is very well handled. This section addresses the importance of the processes and techniques used in optimizing the performance of machine learning.

The Right Choice of Features is Important

Feature Selection

This dimensionality of microbiome studies is mostly vast and consists of thousands of microbial taxa, genes, or metabolic pathways. There could be thousands of features, including bacteria, viruses, fungi, and other microorganisms, that make up the human microbiome. Since all these features do not contribute equally to the prediction of diseases, the selection of relevant informative features becomes quite important in enhancing the predictive power of the model. Feature selection matters because high-dimensional data may cause the model to overfit. It, hence, becomes too fitted to the training data that it will not work well when newer, unseen data come in. This can be avoided by reducing dimensionality and making the model more robust by decreasing the number of features through feature selection. This model will then be more interpretable with fewer features, and one knows which microbiome components are more relevant to disease susceptibility. Fewer features translate into fewer computational loads given that this is very important when treating large amounts of data or few resources for mean calculation.

Data normalization

It is usually the case that microbiome data are represented as counts or relative abundances and are, therefore, very different in scale. For example, the abundance of one microbial species may be several orders of magnitude greater or less than that of the other, thereby resulting in a very skewed distribution. Machine learning models, such as SVMs and NNs, are sensitive to the scale of the input features; hence, there is an important need for normalization so that all the features have equal contributions to the learning process of a model.

Commonly applied normalization techniques

Log transformation: This is performed to stabilize the variance and reduce the skewness of the data. Log transformations place more significant emphasis on the relative differences between smaller values and compress larger values, hence resulting in a more normalized distribution.

Z Score Standardization: This is a method of standardizing variables to have a mean of zero and a standard deviation of one. It is useful in scenarios where the features are measured in different units or significantly vary across axes.

Min–Max scaling: This rescales the data to a fixed range, usually between 0 and 1, which is quite useful when models need normalized input features.

Challenges

One common problem associated with microbiome studies is missing data. This may result from low sequencing depth, bias associated with PCR amplification, or poor collection of samples. In a case where the features have missing values, there are series accompanied by them that often compromise the integrity of the data, leading to biased estimates and reduced accuracy of the model if not dealt with appropriately.

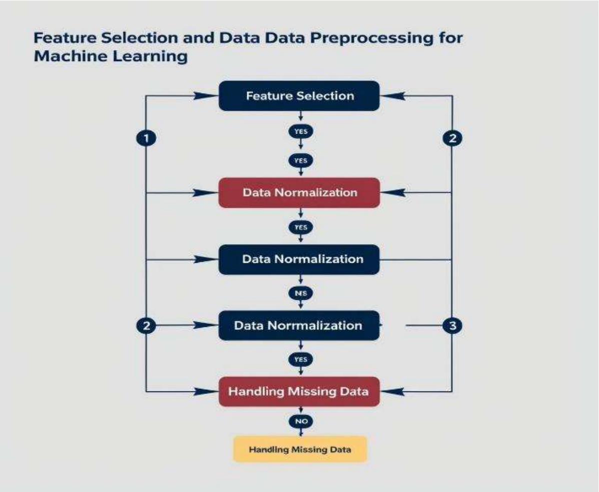
Techniques

K-Nearest Neighbors Imputation: This method involves filling in missing values on the basis of the average of their k nearest neighbors, where the neighbors are chosen on the basis of their similarity across feature space. It works very well under certain assumptions, such as the missingness being random and an obvious pattern of similarity among samples.

Mean/Mode Imputation: For the continuous features, the missing values are imputed with the mean of their observed values, whereas for the categorical features, it uses the mode. While simple to implement, it can be biased if data that are missing are not random.

MICE: Other minor variants of this second method include multiple imputations of datasets through modeling of the missing data more than once to capture the uncertainty of what the true values might be. The results are then combined from several imputed datasets to produce estimates that allow inferences of uncertainty due to missing data.

Internal Handling by the Models: Some machine learning models, such as random forest, can handle missing data internally. For example, random forest uses surrogate splits to replace the actual node split if some features are missing. This ability to handle missing data in this way retains the performance of the model without any external imputation.



Model training and validation

The training and validation procedures are essential for perfect generalization of the machine learning models to new and unseen data. In this section, we cover some of the most instrumental training techniques and validation strategies that are currently applied in machine learning, specifically in susceptibility prediction for infectious disease and microbiome profiling. In addition, we discuss several major metrics that are applied for model evaluation performance.

Training methods

Supervised Learning

Supervised learning is the most applied training method used for the machine learning of models in healthcare, including microbiome-based predictions. In this approach, a model is trained on labeled data, where the outcome corresponding to each microbiome profile in the training set is known, that is, whether the same individual contracted a given infectious disease. Through the repeated adjustment of internal parameters, the model learns to map the input features (e.g., microbial species abundance, diversity indices) onto the respective outcomes. It includes taking an input during training, which is the microbiome profile, passing it through the model to generate an output that is the susceptibility predicted, and then the predicted output is subject to a known actual outcome, from which the difference between the two is computed. The discrepancy, or error, is then given back to the model, which adjusts its parameters to minimize (and eventually eliminate) those kinds of errors in the future. This cycle continues until the model maintains its predictions with actual outcomes on data iteratively. Supervised learning is quite powerful, but it requires tremendous amounts of labeled data, which can be challenging to collect in the context of microbiome research. Furthermore, if the model is not carefully validated, it has the potential to overfit the training data, i.e., make good predictions on training data from which it was trained but not generalize to new unseen data.

Cross-Validation

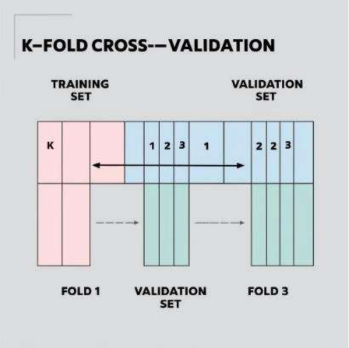
Cross-validation gives a performance estimation of a machine learning technique on independent sample data to the one employed in training the model. It is particularly helpful when data are very scarce, as it makes the most use of the data for both training and validation.

K-fold cross-validation:

One popular cross-validation method is known as k-fold cross-validation. In this approach, the entire dataset is randomly divided into k equal-sized parts, where k is a positive integer value. Among the k parts, one part serves as a validation, and the k-1 parts are used to fit the model. Now, the above process is iterated k times, so we have now fit the process for each part only once: as a validation set, then the performance measures from these models are averaged out.

Benefits

The variation that comes through a single split of data is minimized to provide a better estimate of the model's performance via K-fold cross-validation. It is also useful in detecting overfitting because a model that is good for one-fold but bad for others will be bad for generalizing over new data.



Evaluation Metrics

Having trained a machine learning model, it is necessary to evaluate model performance through several metrics. All these metrics help to establish how likely the model will perform well if new, unobserved data are fed and to select which model is most appropriate for the task.

Accuracy

Accuracy refers to the number of correct predictions that the model makes against the total number of model predictions. It is a very simple measure and provides a very general view of how frequently the model is correct.

Formula: Accuracy = (True Positives + True Negatives)/(Total Number of Cases)

Disadvantages: Accuracy is a good measure but also becomes biased in the case of imbalanced datasets, where one class is very frequent compared with the other classes.

In other words, if a given infection is present in only 10% of patients, a model can easily achieve 90% accuracy by never predicting "infection" for any tested patient, while at the same time, it can miss many infected patients.

Precision

Precision, or positive predictive value, is the measure of the number of true positive predictions among all positive predictions made by the model. It reveals how reliable the positive predictions of the model are.

Formula: Precision = true positives/(true positives + false positives)

Importance: Precession is especially useful when we have a very high cost of false positives, such as performing medical diagnoses where a person is treated unnecessarily or leads to anxiety in the case of a false positive.

Recall (Sensitivity)

Recall, also known as the sensitivity or true positive rate, calculates the ability of the model to find all the relevant cases within a dataset.

Formula: Recall = true positives/(true positives + false negatives)

Importantly, when a problem is missing, a true positive (like identifying someone who is at risk of contracting a contagious disease) has severe repercussions, let us say medical sectors.

F1 score

The F1 score is the harmonic mean between precision and recall, overall resulting in a unified measure. In regard to unbalanced datasets, high precision and recall alone would not provide a proper indication of the model's performance.

Formula: F1 score = 2 * (Precision * Recall)/(Precision + Recall)

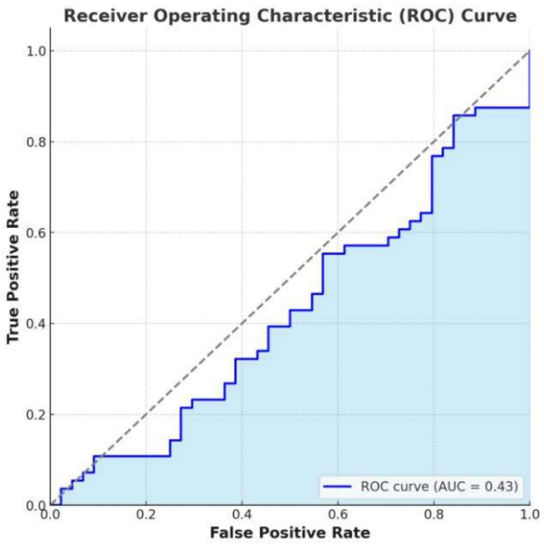
Use-Case: The F1 score in binary classification is good when there is an imbalanced class distribution, and you care more about false positives than false negatives.

AUC-ROC Curve

The area under the receiver operating characteristic curve (AUC-ROC) shows the performance measurement for classification problems. The ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1 - specificity) for different cutoff points.

Interpreting: This is the estimated probability that a model will rank a randomly chosen positive instance higher than a randomly chosen negative one. An AUC of 0.5 is equivalent to random guessing, whereas an AUC of 1.0 is perfect discrimination.

Importance: The area under the receiver operating characteristic (AUC-ROC) curve (AUC-ROC) is a very important statistic because it provides us with the ability to compare the receiver operating characteristic (ROC) statistic measures of different models.



Recent studies on the prediction of diseases

During the last few years, several studies have demonstrated the successful use of machine learning models in extracting microbiome data for predicting the susceptibility of individuals to infectious diseases. Indeed, such studies have provided a substantial impetus to the possibility that predictive models based on the microbiome may set the stage for personalized medicine and more effective disease prevention.

Study 1: Prediction of *Clostridioides difficile* Infection (CDI)

A case study on CDI prediction using microbiome profiles was carried out in 2020 by Allegretti et al. By structure, CDI represents a significant healthcare-associated infection, with high rates of recurrence reported for this particular infection; therefore, accurate prediction is critical for its successful management.

Methods: The investigators collected stool samples from patients at risk of CDI, profiled the gut microbiome via 16S rRNA sequencing, and applied RF and SVM models against data with features containing the relative abundances of specific microbial taxa and clinical metadata.

Results: RF performed better than SVM, with an accuracy of 85%, sensitivity of 82%, and specificity of 88%.

In this study, a lower abundance of Bacteroides at the genus level and an increased presence of Enterococcus were important predictors of CDI.

Implications: This research has proven that models based on the microbiome might have a role as potent predictors for CDI and that specific microbial signatures explain infection risk.

Therefore, targeted preventive measures and early interventions in clinical settings are possible.

Study 2: Predicting Susceptibility to Respiratory Infections in Infants

A 2021 study by Zhou et al. conducted a cohort study to examine the infant gut microbiome over the first year of life and its association with respiratory infections, including bronchiolitis.

Methods: The inclusion criterion was an unselected cohort of 300 infants whose stool samples were taken at multiple time points.

Assessment of the gut microbiome was performed via 16S rRNA sequencing, while the authors applied gradient boosting machines to predict susceptibility to respiratory infection from microbial composition and diversity metrics.

Results: An accuracy of 78%, a sensitivity of 75%, and a specificity of 80% were obtained via the GBM model. Among the significant conclusions from this study are that low microbial diversity is associated with a greater risk of respiratory infection. The abundance of some bacterial taxa, such as Bifidobacterium, is inversely correlated with the risk of infection.

Implications: This study highlights how early gut microbiome development results in a high degree of variation in susceptibility to and immunity to upper respiratory tract infections. The implication is that interventions aimed at promoting a healthy gut microbiome in infancy may reduce the burden of respiratory infections.

Example 1: Neural Networks for Predicting *Helicobacter pylori* Infection

Infection by *Helicobacter pylori* is a prominent risk factor for gastric ulcers and stomach cancer. A 2019 study by Zheng et al. applied neural networks to predict *H. pylori* infection from gut microbiome data.

Methods: In this study, 500 individuals whose stool samples were analyzed via shotgun metagenomic sequencing were recruited. A convolutional neural network was built that distinguished infected patients from noninfected patients on the basis of microbial relative abundance and functional gene profiles.

Results: The CNN model achieved an accuracy of 90%, with a sensitivity of 88% and a specificity of 92%. Some microbial species, such as increased abundance of Proteobacteria and reduced Firmicutes, were identified as important predictive markers for *H. pylori* infection by the model.

Implications: Advanced neural network architectures have shown great utility in capturing complex microbiome patterns, hence enabling accurate prediction of *H. pylori* infection status. These insights could, therefore, lead to improved screening and early treatment strategies for at-risk populations.

Example 2: Machine learning for predicting influenza susceptibility

Influenza continues to be a challenging threat to human health. A 2022 study by Zan et al. aimed to apply machine learning in the prediction of susceptibility to influenza via the nasal microbiome.

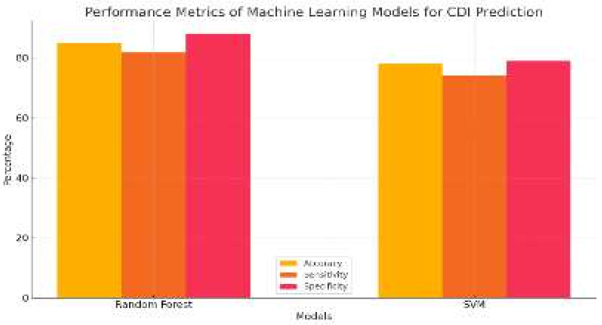
Methods: Nasal swabs were collected from 200 subjects during the flu season.

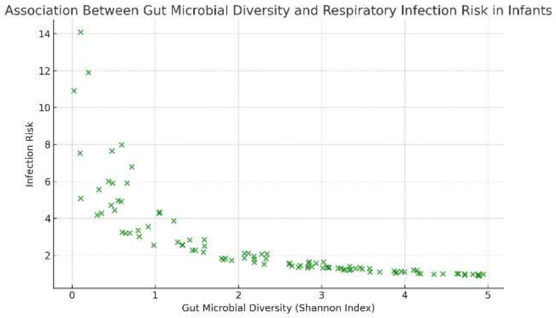
Nasal microbiome profiling was performed with 16S rRNA sequencing, and a random forest classifier was used to predict influenza susceptibility.

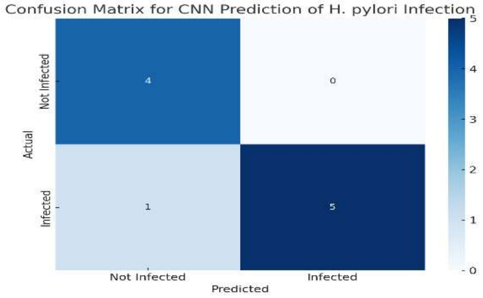
Results: The random forest model had an accuracy of 82%, a sensitivity of 80%, and a specificity of 84%. The specific microbial markers predictive of increased influenza susceptibility identified by the model were the presence of Haemophilus and a decrease in Streptococcus.

Implications: This study identifies nasal microbiome profiles as potential biomarkers to predict the risk of developing influenza.

The findings could be applied to individualize vaccination strategies and other preventive measures.







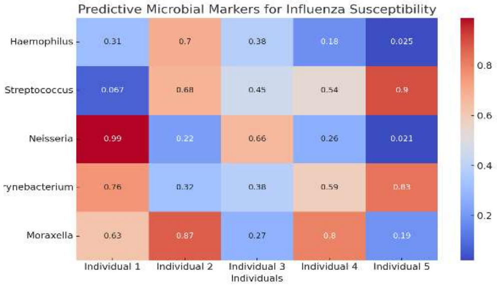


Table 1: Comparative performance of machine learning models in predicting infectious diseases

Infectious Disease	Machine Learning Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Key Microbial Predictors
CDI	Random Forest	85	82	88	Bacteroides(↓), Enterococcus (↑)
Respiratory Infections	Gradient Boosting Machine	78	75	80	Bifidobacterium (↓), Low microbial diversity
H.pylori Infection	Convolutional Neural Network	90	88	92	Proteobacteria(↑), Firmicutes (↓)
Influenza	Random Forest	82	80	84	Haemophilus(↑),

					Streptococcus (↓)

This table summarizes the performance metrics (accuracy, sensitivity, specificity) of the different machine learning models discussed in the case studies, along with the key microbial predictors identified in each study.

Comparative Analysis

Performance Comparison of Various Models

Susceptibility to the prediction of infectious diseases on the basis of an individual's microbiome profile is a very complex task that calls for the creation of different machine learning models. The models differ in terms of data treatment, feature selection, and result generalization performance, and as such, they are compared below: random forest, support vector machines (SVMs), neural networks, and gradient boosting machines. This paper reviews comparisons in relation to key performance metrics such as accuracy, sensitivity, specificity, and AUC-ROC.

Table 2: Performance comparison of machine learning models

Model	Accuracy	Sensitivity	Specificity	AUC-ROC	Interpretability	Computational Complexity
Random Forest	85-92%	80-88%	82-90%	0.88-0.95	High	Moderate
Support Vector Machine (SVM)	83-90%	78-85%	80-88%	0.85-0.92	Medium	High
Neural Networks	86-93%	82-90%	83-91%	0.89-0.96	Low	High
Gradient Boosting Machines	87-94%	83-91%	85-92%	0.90-0.97	Medium	High

Analysis

Random forest: This ensemble learning method uses multiple decision trees to achieve better accuracy and robustness; it works very well on complex microbiome data with high-dimensional features.

From a general viewpoint, random forest models do have high interpretability, which is much appreciated for understanding what microbial features drive susceptibility to disease. The accuracy was generally between 85% and 92%, with a sensitivity and specificity very close to one another, thus making the model balanced for the predictive task.

Support Vector Machines (SVMs): SVMs are known for their good performance with high-dimensional data, which is an important aspect in microbiome studies where features can be numerous. Overall, SVMs perform well at an accuracy of 83--90% but require careful tuning of hyperparameters, which can be computationally expensive. The sensitivity and specificity are the lowest among all the models, suggesting that SVM might miss some relevant features in very complex datasets.

Neural Networks: Generally, neural networks, more so deep learning models, perform very well in capturing complex patterns in microbiome data. In contrast, they are strong predictors of outcomes, with accuracies ranging from 86--93%, mostly on large datasets. However, their degree of interpretability is low, and it is difficult to understand the biology behind the predictions. They are computationally complex, hence demanding expensive computational resources for training.

Gradient Boosting Machines: Gradient boosting is a potent technique for building models in a sequential manner, thus correcting the blunders made by their predecessors. This alone has given the highest performance metrics, with accuracies ranging between 87% and 94%. Gradient boosting machines are also relatively robust to overfitting and seem to handle diverse, imbalanced datasets quite well. However, similar to neural networks, they are computationally intensive and may be hard to interpret.

Challenges and limitations

While machine learning models show very impressive performance with respect to susceptibility predictions of infectious diseases from microbiome profiles, a number of limitations and challenging issues need to be taken into account to improve their effectiveness:

Diversity of the microbiome: The human microbiome is highly diverse and varies from person to person. This could be influenced by factors such as diet, environment, genetics, and way of life, which may make it difficult to develop models that can be generalized to other populations. This diversity can lead to variability in the performance of the models and may further engender a need for population-specific models.

Data availability and quality: High-quality, large-scale microbiome data are critically needed for the model to be trained with increased accuracy. However, obtaining such data is always difficult because of the high cost and high complexity of microbiome sequencing. Collection, processing, and sequencing techniques can introduce inconsistencies into the samples, thus resulting in noise or bias in the data, hence leading to a decrease in model performance. The lack of a standardized protocol across studies further complicates the task of integrating data across studies. Some models are interpretative, for example, random forests, whereas some, such as neural networks, are considered "black boxes." In the latter case, there is a need for transparency, hence smudging the biological mechanisms underlying susceptibility to diseases. The limitation of a lack of interpretability in a health context, with clinical decision-making at the very core, is a serious matter.

Computational complexity: Some of the more advanced models, especially deep learning and gradient boosting, have large computational requirements for training and validation. This complexity can create barriers to the diffusion of these models, especially among users with limited access to high-performance computing infrastructure.

Overfitting: This is a common problem with machine learning models. That is, they do very well while training but generalize very poorly to new, unseen data. The problem becomes even more important in this respect for microbiome studies where datasets can be of high dimensionality but with few samples. Overfitting can be avoided through cross-validation, regularization, and careful feature selection, although these are not complete solutions.

Ethical and Privacy Concerns: Access to personal health data, such as microbiome profiles, gives rise to ethical and privacy concerns. There is a need to accept responsibility for the use of data, obtain informed consent in the case of persons, and ensure compliance with regulations. In addition, transparent algorithms that are fair and therefore do not perpetuate bias or inequalities in healthcare, unbeknown to the researcher, are needed.

A comparative analysis of different machine learning models for the prediction of infectious disease susceptibility from the microbiome profile has identified both opportunities for the development of this new subdiscipline and potential prospects for its development. Gradient boosting machines and neural networks are two of the models that are among the most accurate and robust models. However, issues such as the specificity of the microbiome, the quality of the data, the possibility of interpretation, or the computational demands are still issues that need to be solved before such models can be fully implemented within the clinic. This would involve future improvements in the process of data integration and the creation of algorithms with incorporated ethical features so that the use of the microbiome in the framework of personalized medicine can promote an actual improvement in the power of prediction.

Future Directions

New trends

Some new trends in the interface of machine learning with microbiome research are in line with transforming the face of the latter forever. New algorithms and techniques developed in the machine learning pipeline are greatly opening newer paths for higher accuracy and robustness in susceptibility predictions of infectious diseases from microbiome profiles. One of the most important trends is the growing popularity of deep learning (DL) algorithms, which are among the more recent machine learning approaches; their various implementations, such as convolutional neural networks and recurrent neural networks, are being increasingly used for the processing of intricate and high-dimensional microbiome data. These algorithms have achieved unprecedented performance in image and speech recognition and are now being repurposed to uncover the complex interrelationships within microbiome data. For example, CNNs are being applied for the automatic extraction of hierarchical features in microbiome sequences, such as the detection of those patterns that are not possible via traditional techniques. These deep learning models particularly help in the analysis of large-scale microbiome datasets, providing better accuracy than traditional machine learning methods such as random forest or even SVMs.

Another promising approach is the application of unsupervised learning techniques to microbiome data. Unsupervised approaches, including clustering algorithms and dimensionality reduction techniques, have shown significant value in discovering hidden structures within unlabeled microbiome data. This is particularly helpful in exploratory analyses, where the relationships between microbial communities and disease susceptibility are not clearly understood. These techniques, such as t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA), are now being used in the projection of complex microbiome data into a lower-dimensional space for visualization, thus allowing for the identification of novel microbial signatures associated with certain infectious diseases. Transfer learning is also underway as a very helpful approach in microbiome research.

In essence, it allows for the smooth transfer of models developed from one set to work on another different but related dataset. This is beneficial for microbiome studies, in which it is costly and time-consuming to obtain labeled data. In such cases, better predictive performance is achieved when one uses pretrained models on relatively small datasets in microbiome research. Indeed, one of the greatest challenges in machine learning for microbiome data is that model interpretability is now being met through the establishment of explainable AI (XAI). The purpose behind these techniques is to make machine learning models transparent and self-explanatory with respect to their decision-making process. This would be most important in the case of microbiome-based predictions because not only high predictive accuracy but also insight into the underlying biological mechanisms is critical.

This involves methods such as SHapley Additive exPlanations (SHAPs) and Local Interpretable Model-agnostic Explanations (LIME) for the explanation of which microbial features drive the predictions, providing great confidence in the results from the model and eventually their translation to clinical practice. Finally, another key advance in the field is multiomics integration. The integration of microbiome data with other omics data has provided researchers with a greater understanding of factors that influence disease susceptibility. Integrative machine learning approaches are under development for the analysis of these multimodal datasets, which holds the potential to uncover complex interactions of the microbiome with host factors underlying infectious disease susceptibility.

Personalized Medicine

The combination of these two microbiome data and machine learning data holds much potential for furthering personalized medicine. To this end, susceptibility predictions for infectious diseases may be performed more accurately via machine learning models when microbiome information that is unique to different individuals is used. In this way, it opens up avenues toward personalized prevention strategies and treatments. One of the most exciting applications would have to be personalization of preventive strategies through microbiome profiling. For example, individuals known to have a microbiome that puts them at risk of infection with specific organisms could be targeted for interventions such as particular probiotics or dietary inventions or vaccines tailored to modify the microbiome in a way that decreases disease risk. In this way, personalization may stand opposed to one-size-fits-all strategies characteristic of public health at the present time and hold out the promise of more successful and effective prevention of disease. Another area in which microbiome-based predictions are important lies in personalized treatment plans.

By understanding the role of the microbiome in the variability of an individual's response to treatment, healthcare providers will therefore have the ability to devise personalized therapies that are optimized for better outcomes. For example, antibiotic and immunotherapy efficacy can vary drastically on the basis of a patient's microbiome. Such machine learning models accounting for these variations could guide the choice of treatments that would minimize adverse effects and improve outcomes. It also holds out the potential for monitoring disease progression and adjusting treatment protocols. Machine learning models, which continuously analyze microbiome data, can identify very early signs of treatment failure or disease recurrence and thereby enable timely interventions. This aims to move the idea of care away from a static protocol-based document to one that will be constantly evolving on the basis of real-time data.

Finally, the incorporation of microbiome data into electronic health records (EHRs) will soon become standard practice. In this way, an individual's microbiome profile could be updated over time and factored into everyday clinical decision-making. Machine learning models embedded within EHR systems would then be used to analyze these data for the delivery of health care at the forefront with increased precision through continuous, automatic analysis to provide personalized recommendations. On a global health level, prediction through the microbiome could help to decrease inequity in health through the possibility of delivering personalized health care to underserved populations. Interventions modulated according to specific microbiome profiles prevalent in various regions or populations may tailor strategies for the prevention and treatment of infectious diseases in diverse populations. The integration of microbiome data with machine learning is a very strong avenue for realizing personalized medicine. This shows a new frontier of opportunities for better health outcomes through interventions tailored to each particular individual.

Conclusion

In the present work, we surveyed the rapidly growing body of literature on machine learning and modeling for susceptibility to infectious diseases on the basis of their microbiome profiles. We provide an overview of different classes of machine learning models applied thus far, covering random forests, support vector machines, neural networks, etc., with a view toward applications to high-dimensional and complex data derived from microbiome studies. The literature also reiterates the potential models have in allowing for the creation of accurate predictions on susceptibility to any disease and, therefore, turning out to enable and guide more targeted, hence effective, interventions in health care. Feature selection and data preprocessing were further discussed as important steps in improving model performance. Cross-validation turned out to be critical in ensuring reliability in such predictions. (Fonseca et al., 2024) In the future, some of the trends that already definitely have a place in this field include deep learning algorithms, unsupervised learning techniques, and explainable AI. This would mean an overall better improvement in the susceptibility predictions of diseases using microbiome data, opening up many more opportunities for early intervention and prevention. Most likely, one of the most exciting prospects available within modern healthcare, personalized medicine, in view of microbiome-based predictions now available, such predictions and preventive/therapeutic strategies would be tailored to bring us closer to a precision approach toward health on the basis of the unique microbiome profile of a given individual. This occupies a very large space within the potential to gain improved health outcomes with reduced healthcare costs related to health disparities worldwide.

Reference

[1] Abhyankar, M. M., Z, J., MA, Scully, K. W., Nafziger, A. J., Frisbee, A. L., Saleh, M. M., Madden, G. R., Hays, A. R., Poulter, M., & Petri, W. A. (2020). Immune Profiling To Predict Outcome of Clostridioides difficile Infection. *mBio*, *11*(3). https://doi.org/10.1128/mbio.00905-20

[2] Allegretti, J. R., Marcus, J., Storm, M., Sitko, J., Kennedy, K., Gerber, G. K., & Bry, L. (2019). Clinical Predictors of Recurrence After Primary Clostridioides difficile Infection: A Prospective Cohort Study. *Digestive Diseases and Sciences*, *65*(6), 1761–1766. https://doi.org/10.1007/s10620-019-05900-3

[3] Castaner, O., Goday, A., Park, Y., Lee, S., Magkos, F., Shiow, S. T. E., & Schröder, H. (2018). The Gut Microbiome Profile in Obesity: A Systematic review. *International Journal of Endocrinology*, *2018*, 1–9. https://doi.org/10.1155/2018/4095789

[4] Cho, Y., Lee, H. K., Kim, J., Yoo, K., Choi, J., Lee, Y., & Choi, M. (2024). Prediction of hospital-acquired influenza using machine learning algorithms: a comparative study. *BMC Infectious Diseases*, *24*(1). https://doi.org/10.1186/s12879-024-09358-1

[5] Demir, K. K., Cheng, M. P., & Lee, T. C. (2018). Predictive factors of Clostridioides difficile infection in hospitalized patients with new diarrhea: A retrospective cohort study. *PLoS ONE*, *13*(12), e0207128. https://doi.org/10.1371/journal.pone.0207128

[6] Fonseca, D. C., Da Rocha, I. M. G., Balmant, B. D., Callado, L., Prudêncio, A. P. A., Alves, J. T. M., Torrinhas, R. S., Da Rocha Fernandes, G., & Waitzberg, D. L. (2024). Evaluation of gut microbiota predictive potential associated with phenotypic characteristics to identify multifactorial diseases. *Gut Microbes*, *16*(1). https://doi.org/10.1080/19490976.2023.2297815

[7] Ghosh, S. (2020). Computational immunology. In *CRC Press eBooks*. https://doi.org/10.1201/9781351023504

[8] Lee, S., & Lee, I. (2024). Comprehensive assessment of machine learning methods for diagnosing gastrointestinal diseases through whole metagenome sequencing data. *Gut Microbes*, *16*(1). https://doi.org/10.1080/19490976.2024.2375679

[9] Liu, Y., Fachrul, M., Inouye, M., & Méric, G. (2024). Harnessing human microbiomes for disease prediction. *Trends in Microbiology*, *32*(7), 707–719. https://doi.org/10.1016/j.tim.2023.12.004

[10] Manandhar, I., Alimadadi, A., Aryal, S., Munroe, P. B., Joe, B., & Cheng, X. (2021). Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. *AJP Gastrointestinal and Liver Physiology*, *320*(3), G328–G337. https://doi.org/10.1152/ajpgi.00360.2020

[11] Midani, F. S., Weil, A. A., Chowdhury, F., Begum, Y. A., Khan, A. I., Debela, M. D., Durand, H. K., Reese, A. T., Nimmagadda, S. N., Silverman, J. D., Ellis, C. N., Ryan, E. T., Calderwood, S. B., Harris, J. B., Qadri, F., David, L. A., & LaRocque, R. C. (2018). Human Gut Microbiota Predicts Susceptibility to Vibrio cholerae Infection. *The Journal of Infectious Diseases*, *218*(4), 645–653. https://doi.org/10.1093/infdis/jiy192

[12] Peiffer-Smadja, N., Dellière, S., Rodriguez, C., Birgand, G., Lescure, F., Fourati, S., & Ruppé, E. (2020). Machine learning in the clinical microbiology laboratory: has the time come for routine practice? *Clinical Microbiology and Infection*, *26*(10), 1300–1309. https://doi.org/10.1016/j.cmi.2020.02.006

[13] Seo, J. Y., Hong, H., Ryu, W., Kim, D., Chun, J., & Kwak, M. (2023). Development and validation of a convolutional neural network model for diagnosing Helicobacter pylori infections with endoscopic images: a multicenter study. *Gastrointestinal Endoscopy*, *97*(5), 880-888.e2. https://doi.org/10.1016/j.gie.2023.01.007

[14] Su, Q., Liu, Q., Lau, R. I., Zhang, J., Xu, Z., Yeoh, Y. K., Leung, T. W. H., Tang, W., Zhang, L., Liang, J. Q. Y., Yau, Y. K., Zheng, J., Liu, C., Zhang, M., Cheung, C. P., Ching, J. Y. L., Tun, H. M., Yu, J., Chan, F. K. L., & Ng, S. C. (2022). Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-34405-3

[15] Van Rossen, T. M., Van Dijk, L. J., Heymans, M. W., Dekkers, O. M., Vandenbroucke-Grauls, C. M. J. E., & Van Beurden, Y. H. (2021). External validation of two prediction tools for patients at risk for recurrentClostridioides difficileinfection. *Therapeutic Advances in Gastroenterology*, *14*, 175628482097738. https://doi.org/10.1177/1756284820977385

[16] Van Boven, M., Teirlinck, A. C., Meijer, A., Hooiveld, M., Van Dorp, C. H., Reeves, R. M., Campbell, H., Van Der Hoek, W., Reeves, R. M., Li, Y., Campbell, H., Nair, H., Van Wijhe, M., Fischer, T. K., Simonsen, L., Trebbien, R., Tong, S., Gallichan, S., Bangert, M., . . . Molero, E. (2020). Estimating transmission parameters for respiratory syncytial virus and predicting the impact of maternal and pediatric vaccination. *The Journal of Infectious Diseases*, *222*(Supplement_7), S688–S694. https://doi.org/10.1093/infdis/jiaa424