Fenerbahce University
Journal of Health Sciences
Volume 5, Issue 2, 206-224, 2025

Fenerbahçe Üniversitesi
Sağlık Bilimleri Dergisi
Cilt 5, Sayı 2, 206-224, 2025

FBU-JOHS

The Comparison of Machine Learning Algorithms for Microbiome Data

Mikrobiyom Verileri için Makine Öğrenme Algoritmalarının Karşılaştırılması

Özlem AKAY[1*], Gulfer YAKICI[2]

[1] Gaziantep Islam Science and Technology University, Faculty of Medicine, Department of Biostatistics, Gaziantep, Türkiye.

[2] Gaziantep Islam Science and Technology University, Faculty of Medicine, Department of Medical Microbiology, Gaziantep, Türkiye.

Abstract

The application of next-generation sequencing (NGS) technologies has enabled the identification of both culturable and non-culturable microorganisms in blood samples, revealing their potential roles in systemic infections and immune responses. However, the complexity and high dimensionality of microbiome data present significant challenges for analysis. In this study, it was evaluated the performance of various machine learning (ML) algorithms, including logistic regression, random forest (RF), decision tree, and support vector machines (SVM), in classifying 16S rRNA gene sequencing data of blood microbiota into cultured and uncultured groups. The dataset used in this study, obtained from Kalfin and Panaiotov, consists of 16S rRNA gene sequences from a total of 18,093 OTUs and 62 observations, including control samples. After excluding the six control samples, 56 samples from target sequencing of cultured and non-cultured blood samples of healthy individuals were analyzed. Results show that the random forest (RF) algorithm exhibits the highest classification performance, successfully distinguishing between cultured and uncultured blood microbiota. In the study, the potential of ML techniques in microbiome research was evaluated and the effectiveness and accuracy of these techniques in the analysis of microbiome data were investigated.

Keywords: Blood microbiota, machine learning, metagenomics, microbiome

Özet

Yeni nesil dizileme (NGS) teknolojilerinin uygulanması, kan örneklerinde hem kültürlenebilen hem de kültürlenemeyen mikroorganizmaların tanımlanmasını sağlayarak, sistemik enfeksiyonlarda ve bağışıklık tepkilerinde potansiyel rollerini ortaya koymuştur. Ancak, mikrobiyom verilerinin karmaşıklığı ve yüksek boyutluluğu, analiz için önemli zorluklar sunmaktadır. Bu çalışmada, lojistik regresyon, rastgele orman (RF), karar ağacı ve destek vektör makineleri (SVM) dahil olmak üzere çeşitli makine öğrenimi (ML) algoritmalarının, kan mikrobiyotasının 16S rRNA gen dizileme verilerini kültürlenmiş ve kültürlenmemiş gruplara sınıflandırmadaki performansı değerlendirilmiştir. Çalışmada kullanılan veri seti, Kalfin ve Panaiotov'dan elde edilen 16S rRNA gen dizileri ile oluşturulmuş olup, toplamda 18.093 OTU ve 62 gözlem içermektedir; bunlar arasında kontrol örnekleri de bulunmaktadır. Altı kontrol örneği çalışmadan çıkarıldıktan sonra, sağlıklı bireylerden alınan kültürlü ve kültürsüz kan örneklerine ait 56 örnek üzerinde analizler yapılmıştır. Bulgular, rastgele orman (RF) algoritmasının en yüksek sınıflandırma performansını sergilediğini ve kültürlenmiş ve kültürlenmemiş kan mikrobiyotası arasında başarılı bir şekilde ayrım yaptığını göstermiştir. Çalışmada, mikrobiyom araştırmalarında ML tekniklerinin potansiyeli değerlendirilmiş ve bu tekniklerin mikrobiyom verilerinin analizindeki etkinliği ve doğruluğu, araştırılmıştır.

Anahtar Kelimeler: Kan mikrobiyatası, makine öğrenmesi, metagenom, mikrobiom

How to cite (atıf için): Akay, O., & Yakıcı, G. (2025). The comparison of machine learning algorithms for microbiome data. *Fenerbahçe University Journal of Health Sciences, 5*(2), 206-224. DOI: 10.56061/fbujohs.1636654

*Corresponding author: ozlem.akay@gibtu.edu.tr

DOI: 10.56061/fbujohs.1636654
RESEARCH ARTICLE

Subbmission Date: 10.02.2025, Acceptance Date: 07.03.2025, Publication Date: 29.08.2025

1. Introduction

The microbiota and microbiome are terms frequently used interchangeably to refer to the community of microorganisms (bacteria, viruses, fungi and parasites) present in the human body (Gotschlich et al., 2019).

Over the past decade, the importance of the human microbiome has been recognized (Porras & Brito, 2019). Recent studies on the human microbiome has significantly enhanced the understanding of the critical role that microbiota play in health throughout life.

Microbiota communities display diversity in various areas of the human body, such as the oral cavity, skin, blood, genitourinary and gastrointestinal system (Requena & Velasco, 2021). Contrary to the common perception of the gut microbiome as a passive component, it not only actively plays a role in various host functions such as circadian rhythms, dietary adaptation, metabolism, and immunity, but also assumes significant roles in other biological processes including vitamin synthesis and even neurological functions (Aggarwal et al., 2023; Zheng et al., 2020).

Microbiota is also important for human health and disease; particularly, alterations in microbial diversity are thought to contribute to the pathogenesis of various diseases (Bilgin & Hanci, 2023; Requena & Velasco, 2021).

The microbiota functions similarly to a fingerprint, with each individual has a distinct and unique composition and distribution (Ciftciler & Ciftciler, 2022).

This uniqueness underscores the complexity of our internal environments and how imbalances can influence our health. In this regard, although there has been significant research on different types of microbiota, one area of growing interest is the concept of blood microbiota, which refers to the presence of microbial DNA, RNA, or viable organisms in the bloodstream (Demirci et al., 2023; Goraya et al., 2022; Mair & Sirich, 2019).

Traditionally, the blood of healthy individuals has been presumed to be sterile. However, recent advancements in microbiome research have challenged this view. Although detecting microorganism in the blood has been interpreted as a sign of infection, recent studies highlight the blood microbiota plays a crucial role in understanding systemic infections, immune responses, and overall health (Tsafarova et al., 2022).

Blood culture remains the gold standard among methods used to detect living microorganisms in the bloodstream. But new findings have demonstrated that the presence of genetic material from beneficial and/or pathogenic microorganisms in the blood circulation with the advancements in metagenomic and molecular biology techniques such as next generation sequencing (NGS) have enabled the detection and characterization of microorganisms, including those that are difficult or impossible to culture (Cheng et al., 2023).

*Corresponding author: ozlem.akay@gibtu.edu.tr

DOI: 10.56061/fbujohs.1636654
RESEARCH ARTICLE

Studies conducted using these techniques have shown that non-culturable microorganisms contribute significantly to the diversity of blood microbiota and are also associated with chronic infections, sepsis, and systemic inflammatory conditions (Kajihara et al., 2019; Khan, Khan, Jianye, et al., 2022; Khan, Khan, Usman, et al., 2022).

Moreover, these techniques have significantly enhanced comprehension of the human microbiome and the interactions between host and microbes under both normal and pathological conditions. Notably, numerous studies analyzing blood samples have identified bacteria and their genetic material, even in individuals who are healthy (Cheng et al., 2023).

Therefore, this microbial flora has significant interest among researchers, particularly due to their potential effects on human health and/or diseases (Tsafarova et al., 2022).

Culturable microorganisms refers to microorganisms that have been grown and maintained in a controlled laboratory environment, by using specific conditions such as media that support the microbial growth (Molina-Menor et al., 2020). Successful cultivation of microorganisms provides an opportunity to study their biochemical properties and potential health effects in more detail. But culturing techniques alone have identified limited numbers of microorganisms (Emery et al., 2020).

Unlike culturable microorganisms, those non-culturable refer to microorganisms that can not be cultivated in laboratory environments using traditional culture techniques (Molina-Menor et al., 2020).

A significant portion of the normal microbial flora in humans is nonculturable (Panaiotov et al., 2021). Fortunately, the development of high-throughput technologies has made it possible to understand the genomes and functions of both culturable and non-culturable microorganisms (Lee et al., 2022).

These microorganisms have been indirectly identified by molecular techniques such as shutgun, 16s, metagenome/ whole genome sequencing methods (Panaiotov et al., 2021).

Culturable microorganisms are directly associated with bloodstream infections (BSI) and sepsis, and these pathogens can be easily identified through traditional culture methods, facilitates clinicians to initiate antimicrobial therapy (Costa & Carvalho, 2022; Timsit et al., 2020).

Meanwhile, rapid characterizing non-culturable microorganisms using alternative methods allows researchers and clinicians to investigate their pathogenic mechanisms, interactions with host immune responses, and contributions to chronic infections or inflammatory conditions (Li et al., 2014; Potgieter et al., 2015).

Discriminating between cultured and uncultured blood samples is crucial in microbiology, particularly for the rapid and accurate identification of microbial infections in clinical settings. This distinction helps determine the presence of pathogens, guiding treatment decisions, and provides insights into the microbiome's role in various disease states. It also enhances diagnostic accuracy by distinguishing between active infections, microbial contamination, and colonization. Understanding the dynamics of both culturable and non-culturable microorganisms in the blood microbiota is essential for studying the microbiota in healthy individuals, improving patient outcomes, and effectively managing infections to address public health challenges (Panaiotov et al., 2021).

The human microbiota includes various microorganisms such as bacteria, viruses, fungi, and protozoa, and its genome is over 100 times larger than the human genome. There are different types of technologies used in microbiome researches such as NGS including 16sRNA amplicon sequencing and shotgun metagenome (Loganathan & Priya Doss, 2022).

Improvements in high-throughput sequencing (HTS) have promoted rapid developments in the field of microbiome studies, leading to the generation of extensive microbiome datasets, and recently, investigations have begun exploring how these microbiome patterns can predict host traits through machine learning (ML) (Liu et al., 2021; Loganathan & Priya Doss, 2022).

ML is a subset of statistical and computational methods used to extract information from large datasets. In microbiome researches, machine learning algorithms are crucial for analyzing metagenomic datasets and extracting biological patterns and relationships from these microbiome data (D'Elia et al., 2023). For instance, data mining and deep learning techniques can assist in understanding relationships between different species and the impact of microorganisms on health in microbiota analysis (Saboo et al., 2022). Furthermore, ML models have shown promising performance in distinguishing between patients and healthy individuals (Kim et al., 2023).

2. Method

2.1. Aim of the Study

This study aimed to evaluate the performance of various machine learning methods in classifying blood microbiota as cultured or uncultured based on bacterial operational taxonomic units (OTUs) identified through next-generation sequencing technologies. Blood microbiota have recently become a popular research topic, and understanding their classification can provide valuable insights into microbial community dynamics.

2.2. Data Collection

16S rRNA gene sequence data obtained by Kalfin and Panaiotov was used as the data set (Github-1, 2024). The dataset comprises a total of 18093 OTUs and 62 observations, including controls. Control samples (n=6) were excluded from the study, resulting in a total of 56 samples for analysis. Additionally, OTUs with invalid taxonomic classification (n=11066) were removed from the dataset. Among the remaining 7027 OTUs, those with a combined abundance ≥100 reads and a taxonomic assignment cut-off value >10 reads per OTU were included (Panaiotov et al., 2021). The dataset includes two classes of 56 observations, cultured and uncultured, and 260 OTUs.

2.3. Limitation of the Study

The focus on a specific dataset and algorithms, as well as the lack of consideration for variable ordering, are limitations of the study.

2.4. Data Analysis

In the study, general information was provided on machine learning classification algorithms that have recently become prevalent in microbiome research, including logistic regression, random forest, decision trees, and support vector machines, and microbiome data were classified using these methods. Models of the algorithms were obtained in R (version 4.3.1; R Foundation for Statistical Computing, Istanbul, Turkey) using the mikropml package (Topçuoğlu et al., 2021). Performances of the obtained models have been evaluated by performance metrics (Area Under the Curve (AUC), Accuracy, Balanced Accuracy, Detection Rate, F1, Kappa, negative predictive value (NPV), positive predictive value (PPV), Precision, Recall, Sensitivity, Specificity, log Loss, cross-validation AUC (cv_metric_AUC), Precision-Recall Area Under the Curve (prAUC)).

2.5. Analysing Methods

2.5.1. Logistic Regression

Linear regression models examine the relationship between one or more independent variables and a single continuous dependent variable (Y). If the dependent variable value can only be one of two outcomes (i.e. a binary variable such as dead/alive, injured/uninjured, or accident/no accident), logistic regression is used to model binary response data. The dependent variable is usually treated as an indicator variable. The outcome being predicted is assigned a value of 1, and the other outcome is assigned a value of 0. Since it is not possible to map a linear predictor to only two values, it is mapped to a range of values between 0 and 1. Since probabilities vary between 0 and 1, the linear predictor is mapped to a probability (Bangdiwala, 2018). Consider a collection of p independent variables be denoted by the vector *X′=(X₁, X₂,…,Xₚ)*, and the conditional probability that the outcome is present be denoted by *P(Y=1|X) =π*. Then the logit of having Y=1 is modeled as a linear function of the independent variables as

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \text{ ; } 0 \leq \pi_i \leq 1 \tag{1}$$

where the function

$$\pi_i = \frac{\exp\left(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p\right)}{1+\exp\left(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p\right)} \tag{2}$$

is known as the logistic function (Rana et al., 2010).

Here, π represents the event probability, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (slopes), and $X_1, X_2, \ldots, X_p$ are the independent variables. The parameters β_0 and β are estimated using the maximum likelihood estimation (MLE) method (Abdulqader, 2017).

2.5.2. Random Forest

Random Forest (RF) is an ensemble method that constructs multiple independent decision tree classifiers from various subsets of the dataset (Kouchaki et al., 2020). Breiman (2001) proposed random forests, which add a layer of randomness to bagging. In Bagging, each classifier is built individually by working with a bootstrap sample of the input data. In a regular decision tree classifier, a decision at a node split is made based on all the feature attributes (Alam & Vuong, 2013; Breiman, 2001).

The design of a decision tree requires a pruning method and the choice of an attribute selection measure. In most approaches to attribute selection for decision tree induction, a quality measure is directly assigned to the attribute. Commonly used feature selection measures are the Gini Index, and the Information Gain Ratio Criterion. The Gini Index which measures the impurity of an attribute concerning the classes is used in the random forest classifier as an attribute selection measure. For a given training set T, selecting one case (pixel) at random and saying that it belongs to some class C_i, the Gini index can be written as

$$\sum\sum(f(C_i,T)/|T|)(f(C_j,T)/|T|) \tag{3}$$

where $f(C_i,T)/|T|$ is the probability that the selected case belongs to class C_i (Pal, 2005).

The RF generated mean decrease Gini was used as a measure of the importance of a variable (stress condition and time point) for classifying SigB from non-SigB genes. The decrease in the Gini index is determined for each variable at each node. The mean Gini decrease is the sum of all these decreases due to a given variable, normalized by the number of trees in the forest (Nannapaneni et al., 2012).

RF includes the ability to asses the importance of the predictor variable in predicting the correct answer is a scalable and robust method. This is important because it allows the elimination of statistically dependent variables, the use of computational resources, and reducing the size of the input dataset (Buckley & Harvey, 2021).

2.5.3. Decision Tree

Among machine learning algorithms, the decision tree is classified as a supervised learning method (Sathiyanarayanan et al., 2019). One widely used data mining technique is systems that create classifiers (Jijo & Abdulazeez, 2021). Since it is easier to understand and implement than other classification algorithms it is the most widely used algorithm (Sathiyanarayanan et al., 2019).

A decision tree is a flowchart-like tree structure, where each branch represents an outcome of the test, class label is represented by each leaf node, and each internal node represents a test on an attribute (Sharma & Kumar, 2016). Input is given to the decision tree based on certain criteria and the output is shown as either false or true (Sathiyanarayanan et al., 2019).

A tuple X is given and the attribute values of the tuple are tested against a decision tree. From the root, a path is followed to a leaf node that holds the class prediction for the tuple. Decision trees are easily converted into classification rules (Sharma & Kumar, 2016).

Classification rules are created by the selected path from the root node to the leaf. Since it is the most prominent attribute to separate the data root node is selected first to split each input data. The tree is created by determining the attributes and their associated values to be used to analyze the input data at each intermediate node of the tree.

Once the tree is created, it can pre-structure the new incoming data by visiting all internal nodes in the path starting from a root node and passing towards the leaf node depending on the test conditions of the attributes at each node (Rai et al., 2016).

A decision tree as a predictive model that maps observations about an item to conclusions about the item's target value is used in decision tree learning (Sharma & Kumar, 2016).

2.5.4. Support vector machine (SVMs)

SVMs, which are very effective for many applications in engineering and science, especially for classification problems constitute an important part of learning theory (Wu & Zhou, 2006). SMLs, which formulate the learning problem as a quadratic optimization problem whose error surface is free of local minima and has a global optimum, originate from Vapnik's statistical learning theory (Begg et al., 2005). Support vector machines are the two main classifiers that are attractive and more systematic for learning linear or nonlinear class boundaries. SVM involves two basic steps: training and testing, like all other machine learning techniques. Creating a finite training set by training an SVM involves feeding known data to the SVM along with previously known decision values. An SVM derives its intelligence from the training set to classify unknown data (Othman et al., 2011). The hyperplane separates the two groups of points in the training set by the largest margin when two classes of points in the training set can be separated by a linear hyperplane. This amounts to the hard margin linear support vector machine: Find $(w \in R^n, b \in R)$, to minimize $\|w\|^2$ subject to

$$(\boldsymbol{x_i}.\boldsymbol{w}) + b \geq +1 \quad for\, y_i = +1 \tag{4}$$

$$(\boldsymbol{x_i}.\boldsymbol{w}) + b \ \leq -1\, for\, y_i = -1 \tag{5}$$

Once such w and b are found, the classification rule is sign$[(\mathrm{w} \cdot \mathrm{x}) + b]$.

Constraints (4) and (5) can not be satisfied simultaneously when the points in the training data set are not linearly separable. To overcome this difficulty, non-negative slack variables ξ's can be introduced, and this results in the soft margin linear support vector machine:

Find $w \in R^n, b \in R,$ and ξ_i , $i = 1, 2, \ldots, \ell$ to minimize $\left(\frac{1}{\ell}\right)\left(\sum_{i=1}^{\ell}\xi_i\right)^q + \lambda\|w\|^2$, under the constraints

$$(\boldsymbol{x_i}.\boldsymbol{w}) + b \geq +1 - \xi_i\, for\, y_i = +1 \tag{6}$$

$$(\boldsymbol{x_i}.\boldsymbol{w}) + b \leq -1 + \xi_i\, for\, y_i = -1 \tag{7}$$

$$\xi_i \geq 0, \forall i$$

where q is a positive integer and λ is a parameter to be chosen by the user. Since this is the most common situation it is concentrated on the case q=1. Notice (6) and (7) can be combined as follows:

$$\xi_i \geq 1 - y_i\, [(\boldsymbol{x_i}.\boldsymbol{w}) + b] \tag{8}$$

(Lin et al., 2002).

The simplest way to separate two groups is with a straight line, a flat plane, or an N-dimensional hyperplane. A non-linear dividing line is needed if the points are separated by a non-linear region. In this case, SVM uses a kernel function to map the data to a different space where a hyperplane can be used to separate them. The kernel function can transform the data into a higher-dimensional space to make the separation possible (Bhavsar & Panchal, 2012). Kernel functions and their equations are given Table 1 (Chidambaram & Srinivasagan, 2019).

Table 1. Kernel functions and their equations

Kernel functions	Equations
Linear kernel	$k(x,y) = x^T y + c$
Polynomial kernel	$k(x,y) = (\alpha x^T y + c)^d$
Radial Basis Function kernel	$k(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
Sigmoid kernel	$k(x,y) = tanh(\alpha x^T y + c)$

2.5.5. Classification Metrics

When the response is binary in a machine learning model, classification models such as decision trees, logistic regression, convolutional neural networks, random forests, etc. are used. Classification metrics are used to evaluate the performance of the models. A confusion matrix contains the prediction results of any binary test that is often used to calculate classification metrics. A confusion matrix is shown in Table 2.

Table 2. A confusion matrix

	Observed		
		1 (+)	0 (-)
Predicted	1 (+)	TP	FP
	0 (-)	FN	TN

Confusion matrix and classification accuracy measure TP and TN are the true positive and true negative predictions; FP and FN are false positive and false negative predictions respectively (Table 2). Accordingly, the classification metrics used in the study are explained as follows.

Accuracy measures the overall frequency of correct predictions made by a classifier. Precision, also referred to as Positive Predictive Value (PPV), indicates the proportion of correctly predicted positive cases that are truly positive. Recall, or Sensitivity, measures the ability of the model to correctly identify actual positive cases (Huilgol, 2025). Negative Predictive Value (NPV) reflects the proportion of true negatives among all cases predicted as negative. Specificity assesses the model's effectiveness in correctly identifying actual negative cases. The F1-Score, calculated as the harmonic mean of precision and recall, is used to evaluate classification performance (Srivastava, 2024). Balanced accuracy, which is the average of sensitivity and specificity, is particularly useful when dealing with imbalanced datasets where one class significantly outnumbers the other (Olugbenga, 2024). Cohen's Kappa coefficient (κ) quantifies agreement between predicted and actual values, accounting for the likelihood of random agreement, based on the marginal distributions of each class (Kolena, 2024). Among the metrics based on cross-entropy, the log loss metric, measures the quality of predictions rather than the accuracy. Formulas for these metrics are given the Eq. 13-22. which are

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{13}$$

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

$$PPV\text{(Detection Rate)} = \frac{TP}{TP+FP} \tag{15}$$

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \tag{16}$$

$$NPV = \frac{TN}{TN+FN} \tag{17}$$

$$Specificity = \frac{TN}{TN+FP} \tag{18}$$

$$F1 - score = \frac{2*Precision*Recall}{Precision+Recall} \tag{19}$$

$$Balanced\ Accuracy\ = \frac{Sensitivity+Specificity}{2} \tag{20}$$

$$Cohen\ Kappa\ = \frac{2*(TP*TN-FN*FP)}{(TP+FN)*(FN+TN)+(TP+FP)*(FP+TN)} \tag{21}$$

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}[y_i lnp_i + (1-y_i)\ln(1-p_i] \tag{22}$$

where *y* is the actual/true value, *i* is the given observation/record, p and ln refers to the prediction probability, and the natural logarithm of a number, respectively (Dembla, 2020).

Efficiently, the log loss gages additional error coming the estimates as opposed to the true values (Log, 2024).

Cross-validation AUC (cv_metric_AUC), is the AUC calculated for the cross-validation folds for the training data (AUC, 2024).

Precision-Recall Area Under the Curve (prAUC) quantifies how well a model can distinguish between classes, considering both its ability to not mark a negative sample as positive (Precision) and its ability to find all the positive samples (Recall). A higher prAUC value signifies a better-performing model (Alon, 2023).

3. Results

In the study, for the application of four different commonly used machine learning algorithms, the arguments in the run_ml command were set to default and the dataset was first randomly divided into training and test sets (46 sample training set, 10 sample test set, training_frac=NULL). When learning a dependence from data, to avoid overfitting, it is important to divide the data into the training set and the testing set. Empirical studies show that the best results are obtained if 20-30% of the data for testing was used, and the remaining 70-80% of the data for training (Gholamy et al., 2018). For this reason, the test and training sets were determined automatically by setting the run_ml command as default. The training data was used to create and select the models and the test set was used to evaluate the model. Four different models were trained using the training data. Since the arguments in the run_ml command were set to default, hyperparameters was set as sensible defaults will be chosen automatically (hyperparameters=NULL) for model selection. For resampling, five-fold cross-validation (CV) was performed on the training set, repeated 100 times. The model was trained using the full training dataset and was applied to the test data to evaluate the test prediction performance of each model. Performance metrics were created to determine and evaluate the appropriate model selection for machine learning algorithms. 100 bootstraps (alpha=0.05) were created and the confidence interval estimate for the model performance was calculated (Topçuoğlu et al., 2020). The obtained values are given in Table 3.

Table 3. Performance metrics of machine learning algorithms

	GLM [LCI-UCI]	**RF [LCI-UCI]**	**DT-rpart2[LCI-UCI]**	**SVM [LCI-UCI]**
AUC	0.795 [0.4-1]	0.960 [0.7-1]	0.705 [0.5-1]	0.205 [0-0.6]
Accuracy	0.496 [0.2-0.8]	0.798 [0.5-1]	0.702 [0.4-1]	0.504 [0.2-0.8]
Balanced Accuracy	0.5 [0.5-0.5]	0.801 [0.5-1]	0.705 [0.5-1]	0.5 [0.5-0.5]
Detection Rate	0.266 [0-0.8]	0.428 [0.1-0.8]	0.378 [0-0.8]	0.271 [0-0.8]
F1	0.683 [0.4-0.8]	0.789 [0.4-1]	0.695 [0.2-1]	0.690 [0.4-0.8]
Kappa	0 [0-0]	0.582 [0.1-1]	0.392 [0-1]	0 [0-0]
NPV	0.458 [0.2-0.7]	0.837 [0.3-1]	0.782 [0.3-1]	0.467 [0.2-0.7]
PPV	0.533 [0.3-0.8]	0.871 [0.5-1]	0.821 [0.4-1]	0.542 [0.3-0.8]
Precision	0.533 [0-3-0.8]	0.871 [0.5-1]	0.821 [0.4-1]	0.542 [0.3-0.8]
Recall	0.5 [0-1]	0.800 [0.2-1]	0.706 [0-1]	0.5 [0-1]
Sensitivity	0.5 [0-1]	0.800 [0.2-1]	0.706 [0-1]	0.5 [0-1]
Specificity	0.5 [0-1]	0.802 [0.2-1]	0.704 [0-1]	0.5 [0-1]
cv_metric_AUC	0.865[0.86-0.87]	0.925[0.92-0.93]	0.776[0.77-0.78]	0.854[0.85-0.86]
logLoss	0.676 [0.6-0.7]	0.459 [0.2-0.6]	0.565 [0.2-0.9]	0.693 [0.6-0.7]
prAUC	0.499 [0.3-0.7]	0.629 [0.3-0.6]	0.185 [0-0.3]	0.346 [0.2-0.4]

GLM=Logistic Regression, RF=Random Forest, DT=Decision Tree, SVM=Support Vector Machines, LCI=Lower Confidence Interval, UCI=Upper Confidence Interval, AUC=Area Under the Curve, NPV=Negative Predictive Values, PPV=Positive Predictive Values, prAUC=Precision-Recall Area Under the Curve

The predictive performances of four algorithms for classifying bacteria as cultured or non-cultured were evaluated. When Table 3 is examined, the prediction performance of the random forest model is higher than other ML models for all metrics. However, the performances of logistic regression and support vector machine models are close to each other.

Additionally, to evaluate the performance of machine learning algorithms, the ROC curves obtained by using test data in the model created by training with training data for each algorithm are given in Figure 1.

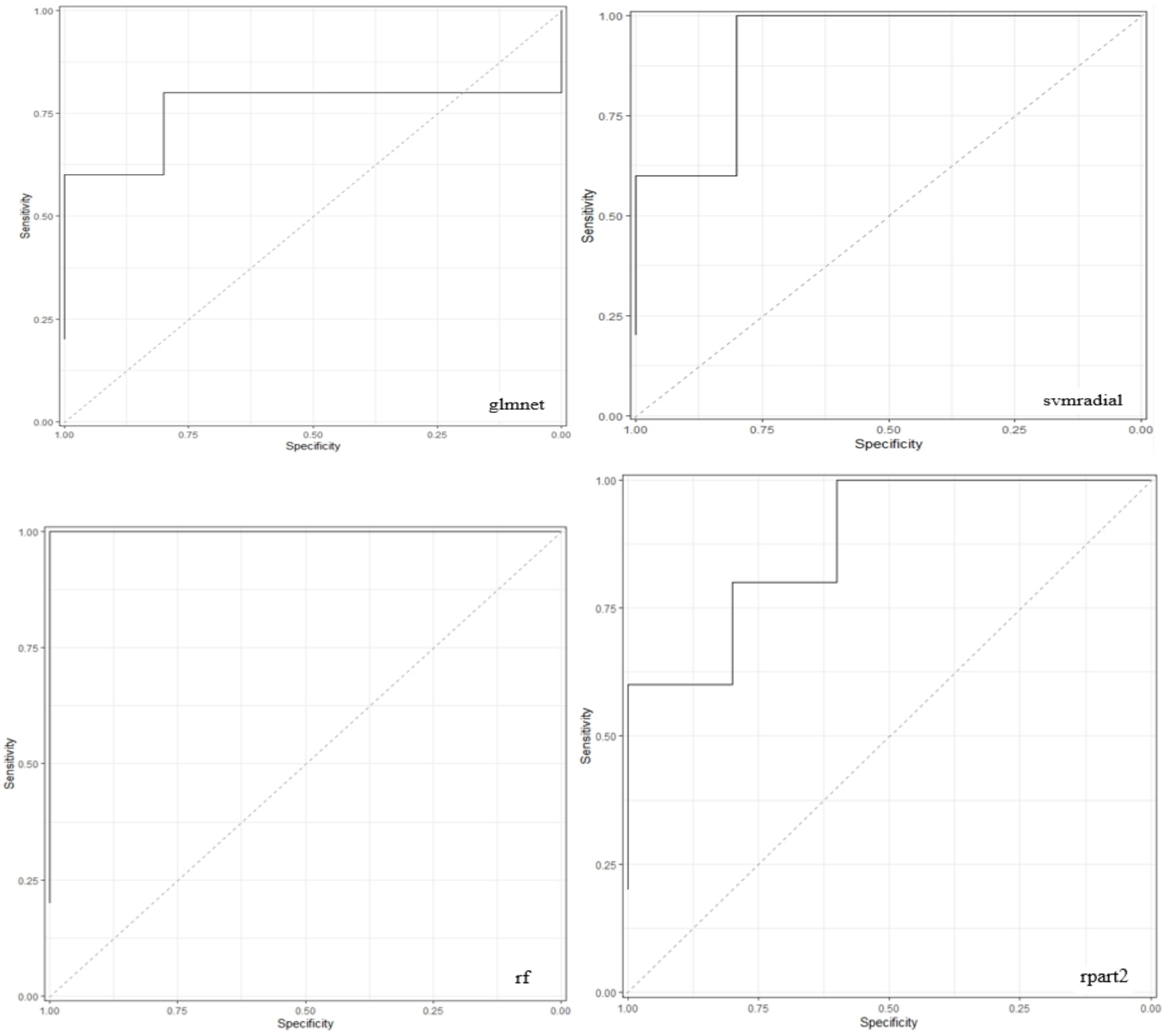

Figure 1. ROC curves for ML algorithms

When Figure 1 is examined, it can be seen that the ML algorithm with the largest area is the random forest (rf), followed by support vector machines (svmradial), decision tree (rpart2), and logistic regression (glmnet) models.

Figure 2 shows the graph of cross-validation AUROC (CV AUC) and test AUROC (test AUC) values for machine learning algorithms. The AUROC ranges from 0, where the model's predictions are completely wrong, to 1, where the model discriminates perfectly between cases and controls. An AUROC value of 0.5 indicates that the model's predictions are no different from random. To evaluate the generalizability of the models built for the algorithms, the cross-validation AUROC median is compared with the test AUROC median. If the difference between cross-validation and testing AUROCs is large, this may indicate that the models are overfitting the training data. When Figure 1 is examined, it is seen that the biggest difference in Median AUROCs is in SVM (svmradial), followed by logistic regression (glmnet), decision tree (rpart2), and random forest (rf) algorithms, respectively. These differences are relatively small and provide confidence in the models' predictions of generalization performance.

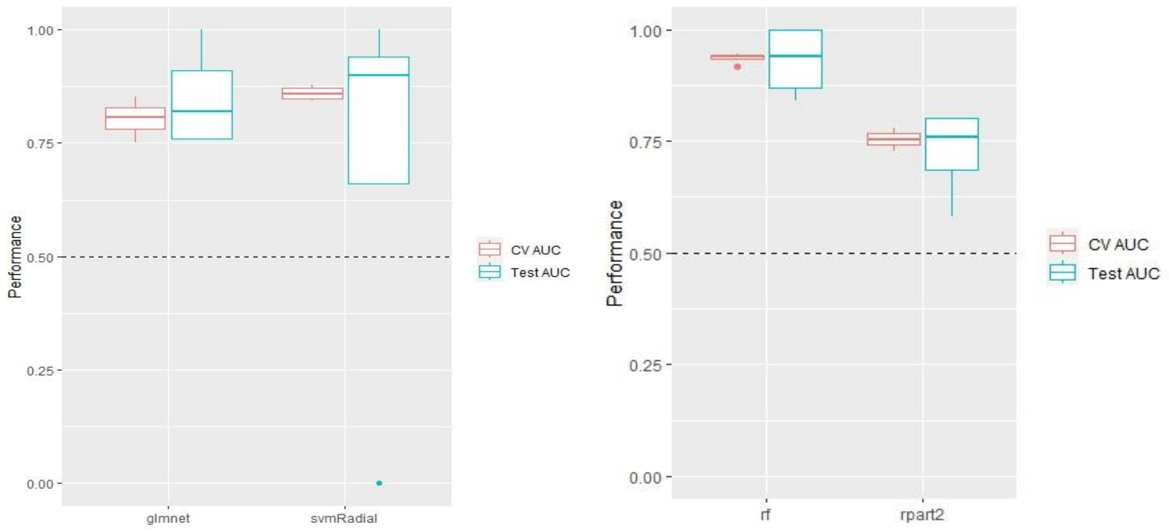

Figure 2. Generalization and classification performance of machine learning (ML) models using

ML models are used not only to predict the classification result but also to identify potential variables for classification. Figure 2 shows a graph showing feature importance, using the median ranking of absolute feature weights for each OTU of the algorithm models for classifying bacteria as cultured or non-cultured.

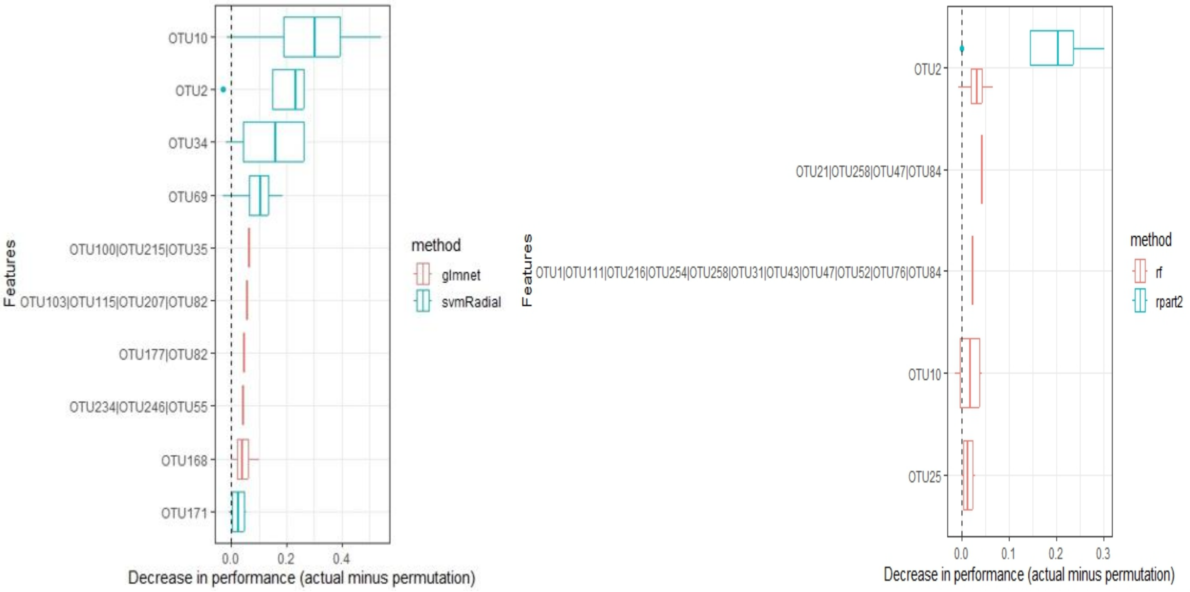

Figure 3. Median ranking of absolute feature weights for each OTU of ML algorithm models (Features which have a lower performance when permuted have a difference in performance above zero. The features with the greatest decrease are the most important for model performance)

Examining Figure 3, it can be seen that among the OTUs with the largest impact, there is only one OTU (OTU 2: FJ957443.1.1492) shared for support vector machines (svmRadial), random forest (rf), and decision tree (rpart2) algorithms. Considering Figure 3, the OTUs with the greatest impact on machine learning algorithms are given in Table 4. The microorganism names corresponding to OTU Ids in the table are long, so they could not be added to the table. However, those interested can access the open access dataset https://github.com/yhodzhev/blood_microbiota for more information (Github-1, 2024).

Table 4. The relative importance of OTUs for the Machine Learning Algorithms

Machine Learning Algorithms	**OTU Name**
Logistic Regression	OTU 35: GQ129964.1.1443 OTU 55: GU561370.1.1451 OTU 82: HM333626.1.1345 OTU 100: JF087988.1.1339 OTU 103: JF111953.1.1335 OTU 115: JF231450.1.1376 OTU 168: JX222876.1.1353 OTU 177: KF063018.1.1342 OTU 207: KF078047.1.1392 OTU 215: AY167836.1.1338 OTU 234: JVEO01000014.1806.3342 OTU 246: FJ558013.1.1408
Support Vector Machines	OTU 2: FJ957443.1.1492 OTU 10: FN421949.1.1367 OTU 34: GQ129886.1.1458 OTU 69: HM072285.1.1438 OTU 171: KC122696.1.1416
Decision Tree	OTU 2: FJ957443.1.1492
Random Forest	OTU 1: FJ948822.1.1302 OTU 2: FJ957443.1.1492, OTU 10: FN421949.1.1367 OTU 21: GDLT01000297.39.1633 OTU 25: GQ012876.1.1370 OTU 31: GQ067833.1.1345 OTU 43: GQ338714.1.1521 OTU 47: GQ487987.1.1471 OTU 52: JF142842.1.1376 OTU 76: HM445303.1.1360 OTU 84: HQ697417.1.1457 OTU 111: JF179224.1.1335 OTU 216: AY328673.1.1474 OTU 254: M01113:65:000000000-BG2CF:1:1103:26342:13806 1:N:0:9/M01113:65:000000000-BG2CF:1:1103:26342:13806 2:N:0:9 OTU 258: M01113:65:000000000BG2CF:1:1101:10166:89021:N:0:41/M01113:65:000000000-BG2CF:1:1101:10166:8902 2:N:0:41

OTU=An Operational Taxonomic Unit

4. Discussion

It is known that a significant part of the microbial flora in healthy individuals is not able to be cultured. This has been demonstrated by next-generation sequencing methods, which are used to better understand the existence and diversity of microorganisms, especially those that are difficult or impossible to culture. Blood microbiota has also been examined in this context, and while some of the microorganisms found in the blood could not be detected by culturing techniques, while others can be successfully cultured. Machine learning plays a critical role in making this distinction, as it can differentiate between cultured and uncultured blood samples with high accuracy, owing to its pattern recognition and classification capabilities across large data sets and complex microbiota profiles. This ability enables more reliable and effective results in diagnosing infections, monitoring health, and conducting microbiota research.

In this study, the classification performance of microbiome data was evaluated using four common machine learning algorithms: logistic regression, random forest, decision tree, and support vector machines. The results demonstrated that the random forest algorithm has the highest classification performance compared to the other algorithms. Logistic regression and support vector machines showed similar performances but fell behind the random forest model. Additionally, when model performances were compared based on AUC (ROC curve) values, it was observed that the random forest algorithm covered the largest area, followed by support vector machines, decision tree, and logistic regression models.

The findings obtained in this study show that the random forest algorithm is the most effective method for classifying microbiome data. The superior performance of the random forest algorithm could be because this algorithm captures more variation by using multiple tree structures and can better manage data complexity. In particular, the high dimensionality and complexity of microbiome data reveal the advantage of such algorithms. Logistic regression and support vector machines gave similar results in terms of performance, but they fell behind the random forest algorithm. This shows that these algorithms may be limited in modeling nonlinear relationships and complex data structures.

When the differences between the cross-validation and test AUROC values of the model performances were examined, it was seen that the largest difference was in SVM, followed by logistic regression, decision tree, and random forest algorithms. The relatively small differences indicate that the generalization performance of the models is reliable. Similar to the results of the study, Teixeira et al. (2022) discussed Machine Learning methods for cancer characterization from microbiome data in their review, where it was stated that the proposed methods, generally based on Random Forests, showed promising results but were insufficient for widespread clinical use (Teixeira et al., 2024). Topcuoğlu et al. (2020) trained seven models that used fecal 16S rRNA sequence data to predict the presence of colonic screen relevant neoplasias (SRNs). To show the effect of model selection, the predictive performance, interpretability, and training time of L2-regularized logistic regression, L1- and L2-regularized support vector machines (SVM) with linear and radial basis function kernels, a decision tree, random forest, and gradient boosted trees (XGBoost) were assessed. The random forest model performed best at detecting SRNs (Topçuoğlu et al., 2020). Beck and Foster (2015) used random forests and logistic regression classifiers to model the relationship between the microbial community and Bacterial vaginosis. Models generated performed nearly identically and identify largely similar important features (Beck & Foster, 2015). Wilhelm et al. (2022) evaluated the accuracy of random forest (RF) and support vector machine (SVM) regression and classification models in predicting 12 measures of soil health, tillage status, and soil texture from 16S rRNA gene amplicon data with an operationally relevant sample set. Acording to results, the efficacy and performance of the ML algorithms differed by task with SVM outperforming RF in classifying health categories while RF surpassed SVM in regression-based prediction of ratings (Wilhelm et al., 2022). Freitas et al. (2023) aimed to distinguish cancer type based on the analysis of tissue-specific microbial information using Random Forest algorithms. They were trained to classify five cancer types, namely head and neck, esophageal, stomach, colon, and rectal cancers, with examples provided by The Cancer Microbiome Atlas database. Random Forest models achieved promising performances when predicting head and neck, stomach, and colon cancer cases (Freitas et al., 2023). However, Wang and Liu (2020) systematically compared Random Forests (RF), eXtreme Gradient Boosting decision trees (XGBoost), elastic network (ENET), and Support Vector Machine (SVM) in the classification analysis of 29 benchmark human microbiome datasets and found that XGBoost outperformed all other methods only on a few benchmark datasets (Wang & Liu, 2020).

This study demonstrates the effectiveness of machine learning algorithms in analyzing and interpreting microbiome data. The superior performance of the random forest algorithm suggests that this method should be preferred for future microbiome studies. However, it should be noted that the performance of other algorithms under different data sets and conditions should also be evaluated. It should also be taken into account that the order of the attributes in the classification has a significant impact on the model performance (Tallón-Ballesteros et al., 2019). Otherwise, it may cause bias in the model results. Therefore, applying appropriate methods for variable ordering when using different algorithms is critical to obtain more reliable and accurate results.

5. Conclusion

Distinguish between cultured and non-cultured blood samples with machine learning algorthims based on microbiota composition, is important for better understanding of blood microbiota dynamics in clinical microbiology. Furthermore, it is believed that these techniques could contribute to the development of more precise diagnostic tools for microbiome analysis in clinical settings. By integrating machine learning, computational methods can enhance microbiome analysis, aid in biomarker identification, and ultimately improve clinical decision-making. These advancements can be considered a promising path for future research into the complex dynamics of microbial communities in health and disease. Furthermore, they constitute an important step for a deeper understanding of microbiome data and the development of more sensitive prediction models.

Future research should focus on expanding the dataset by incorporating a larger and more diverse sample population to enhance the model's generalizability and robustness. Additionally, investigating alternative and hybrid machine learning approaches, such as deep learning models and ensemble techniques, may further improve classification performance. The optimization of feature selection and data preprocessing techniques, including normalization and class balancing methods, could also contribute to increased model accuracy and reliability.

Authors Contributions

Topic selection: OA, GY; Design: OA, GY; Planning: OA, GY; Data collection and analysis: OA, GY; Writing of the article: OA, GY; Critical revision: OA, GY.

Conflict of Interest

All authors report no conflict of interest.

References

Abdulqader, Q. M. (2017). Applying the Binary Logistic Regression Analysis on The Medical Data. Science Journal of University of Zakho, 5, 330-334. https://doi.org/10.25271/2017.5.4.388

Aggarwal, N., Kitano, S., Puah, G. R. Y., Kittelmann, S., Hwang, I. Y., & Chang, M. W. (2023). Microbiome and Human Health: Current Understanding, Engineering, and Enabling Technologies. Chem Rev, 123(1), 31-72. https://doi.org/10.1021/acs.chemrev.2c00431

Alam, M. S., & Vuong, S. T. (2013). Random forest classification for detecting android malware. 2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing, Beijing, China, 2013, pp. 663-669. https://doi.org/10.1109/GreenCom-iThings-CPSCom.2013.122

Alon, T. (2023). Ultimate Guide to PR-AUC: Calculations, uses, and limitations. Retrieved 05.06.2024 from https://www.aporia.com/learn/ultimate-guide-to-precision-recall-auc-understanding-calculating-using-pr-auc-in-ml/

AUC, C. (2024). Retrieved 05.06.2024 from https://cran.r-project.org/web/packages/mikropml/vignettes/introduction.html

Bangdiwala, S. I. (2018). Regression: binary logistic. Int J Inj Contr Saf Promot, 25(3), 336-338. https://doi.org/10.1080/17457300.2018.1486503

Beck, D., & Foster, J. A. (2015). Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. BioData mining, 8, 1-9. https://doi.org/10.1186/s13040-015-0055-3

Begg, R. K., Palaniswami, M., & Owen, B. (2005). Support vector machines for automated gait classification. IEEE transactions on Biomedical Engineering, 52(5), 828-838. https://doi.org/10.1109/TBME.2005.845241

Bhavsar, H., & Panchal, M. H. (2012). A review on support vector machine for data classification. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(10), 185-189.

Bilgin, B., & Hanci, H. (2023). Gut Microbiota and Its Importance for Our Health. Pharmata, 3(3), 71-73. https://doi.org/10.5152/Pharmata.2023.1318972

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Buckley, S. J., & Harvey, R. J. (2021). Lessons Learnt From Using the Machine Learning Random Forest Algorithm to Predict Virulence in Streptococcus pyogenes [Mini Review]. Frontiers in Cellular and Infection Microbiology, 11. https://doi.org/10.3389/fcimb.2021.809560

Cheng, H. S., Tan, S. P., Wong, D. M. K., Koo, W. L. Y., Wong, S. H., & Tan, N. S. (2023). The Blood Microbiome and Health: Current Evidence, Controversies, and Challenges. Int J Mol Sci, 24(6). https://doi.org/10.3390/ijms24065633

Chidambaram, S., & Srinivasagan, K. (2019). Performance evaluation of support vector machine classification approaches in data mining. Cluster Computing, 22, 189-196. https://doi.org/10.1007/s10586-018-2036-z

Ciftciler, R., & Ciftciler, A. E. (2022). The importance of microbiota in hematology. Transfus Apher Sci, 61(2), 103320. https://doi.org/10.1016/j.transci.2021.103320

Costa, S. P., & Carvalho, C. M. (2022). Burden of bacterial bloodstream infections and recent advances for diagnosis. Pathog Dis, 80(1). https://doi.org/10.1093/femspd/ftac027

D'Elia, D., Truu, J., Lahti, L., Berland, M., Papoutsoglou, G., Ceci, M., Zomer, A., Lopes, M. B., Ibrahimi, E., Gruca, A., Nechyporenko, A., Frohme, M., Klammsteiner, T., Pau, E. C. S., Marcos-Zambrano, L. J., Hron, K., Pio, G., Simeon, A., Suharoschi, R., . . . Claesson, M. J. (2023). Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. Front Microbiol, 14, 1257002. https://doi.org/10.3389/fmicb.2023.1257002

Dembla, G. (2020). Intuition behind Log-loss score. Towards Data Science.

Demirci, M., Saribas, A. S., Siadat, S. D., & Kocazeybek, B. S. (2023). Editorial: Blood microbiota in health and disease. Front Cell Infect Microbiol, 13, 1187247. https://doi.org/10.3389/fcimb.2023.1187247

Emery, D. C., Cerajewska, T. L., Seong, J., Davies, M., Paterson, A., Allen-Birt, S. J., & West, N. X. (2020). Comparison of Blood Bacterial Communities in Periodontal Health and Periodontal Disease. Front Cell Infect Microbiol, 10, 577485. https://doi.org/10.3389/fcimb.2020.577485

Freitas, P., Silva, F., Sousa, J. V., Ferreira, R. M., Figueiredo, C., Pereira, T., & Oliveira, H. P. (2023). Machine learning-based approaches for cancer prediction using microbiome data. Scientific reports, 13(1), 11821. https://doi.org/10.1038/s41598-023-38670-0

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. Int. J. Intell. Technol. Appl. Stat, 11(2), 105-111. https://doi.org/10.6148/IJITAS.201806_11(2).0003

Github-1. (2024). Retrieved 06.12.2024 from https://github.com/yhodzhev/blood_microbiota

Goraya, M. U., Li, R., Mannan, A., Gu, L., Deng, H., & Wang, G. (2022). Human circulating bacteria and dysbiosis in non-infectious diseases. Front Cell Infect Microbiol, 12, 932702. https://doi.org/10.3389/fcimb.2022.932702

Gotschlich, E. C., Colbert, R. A., & Gill, T. (2019). Methods in microbiome research: Past, present, and future. Best Pract Res Clin Rheumatol, 33(6), 101498. https://doi.org/10.1016/j.berh.2020.101498

Jijo, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2, 20-28. https://doi.org/10.38094/jastt20165

Kajihara, M., Koido, S., Kanai, T., Ito, Z., Matsumoto, Y., Takakura, K., Saruta, M., Kato, K., Odamaki, T., Xiao, J. Z., Sato, N., & Ohkusa, T. (2019). Characterisation of blood microbiota in patients with liver cirrhosis. Eur J Gastroenterol Hepatol, 31(12), 1577-1583. https://doi.org/10.1097/MEG.0000000000001494

Khan, I., Khan, I., Jianye, Z., Xiaohua, Z., Khan, M., Hilal, M. G., Kakakhel, M. A., Mehmood, A., Lizhe, A., & Zhiqiang, L. (2022). Exploring blood microbial communities and their influence on human cardiovascular disease. J Clin Lab Anal, 36(4), e24354. https://doi.org/10.1002/jcla.24354

Khan, I., Khan, I., Usman, M., Xiao Wei, Z., Ping, X., Khan, S., Khan, F., Jianye, Z., Zhiqiang, L., & Lizhe, A. (2022). Circulating microbiota and metabolites: Insights into cardiovascular diseases. J Clin Lab Anal, 36(12), e24779. https://doi.org/10.1002/jcla.24779

Kim, H., Na, J. E., Kim, S., Kim, T. O., Park, S. K., Lee, C. W., Kim, K. O., Seo, G. S., Kim, M. S., Cha, J. M., Koo, J. S., & Park, D. I. (2023). A Machine Learning-Based Diagnostic Model for Crohn's Disease and Ulcerative Colitis Utilizing Fecal Microbiome Analysis. Microorganisms, 12(1). https://doi.org/10.3390/microorganisms12010036

Kolena, T. w. (2024). Retrieved 06.12.2024 from https://docs.kolena.com/metrics/cohens-kappa/

Kouchaki, S., Yang, Y., Lachapelle, A., Walker, T. M., Walker, A. S., , C. C., Peto, T. E. A., Crook, D. W., & Clifton, D. A. (2020). Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking [Original Research]. Frontiers in Microbiology, 11. https://doi.org/10.3389/fmicb.2020.00667

Huilgol, P. (2025). Retrieved 18.04.2025 from https://www.analyticsvidhya.com/articles/precision-and-recall-in-machine-learning/

Lee, E. J., Sung, J., Kim, H. L., & Kim, H. N. (2022). Whole-Genome Sequencing Reveals Age-Specific Changes in the Human Blood Microbiota. J Pers Med, 12(6). https://doi.org/10.3390/jpm12060939

Li, L., Mendis, N., Trigui, H., Oliver, J. D., & Faucher, S. P. (2014). The importance of the viable but non-culturable state in human bacterial pathogens. Front Microbiol, 5, 258. https://doi.org/10.3389/fmicb.2014.00258

Lin, Y., Lee, Y., & Wahba, G. (2002). Support vector machines for classification in nonstandard situations. Machine Learning, 46, 191-202. https://doi.org/10.1023/A:1012406528296

Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell, 12(5), 315-330. https://doi.org/10.1007/s13238-020-00724-8

Log, G. (2024). Retrieved 05.06.2024 from https://rdrr.io/github/jeffreyevans/rfUtilities/man/logLoss.html

Loganathan, T., & Priya Doss, C. G. (2022). The influence of machine learning technologies in gut microbiome research and cancer studies- A review. Life Sci, 311(Pt A), 121118. https://doi.org/10.1016/j.lfs.2022.121118

Mair, R. D., & Sirich, T. L. (2019). Blood Microbiome in CKD: Should We Care? Clin J Am Soc Nephrol, 14(5), 648-649. https://doi.org/10.2215/CJN.03420319

Molina-Menor, E., Gimeno-Valero, H., Pascual, J., Pereto, J., & Porcar, M. (2020). High Culturable Bacterial Diversity From a European Desert: The Tabernas Desert. Front Microbiol, 11, 583120. https://doi.org/10.3389/fmicb.2020.583120

Nannapaneni, P., Hertwig, F., Depke, M., Hecker, M., Mäder, U., Völker, U., Steil, L., & van Hijum, S. (2012). Defining the structure of the general stress regulon of Bacillus subtilis using targeted microarray analysis and random forest classification. Microbiology (Reading), 158(Pt 3), 696-707. https://doi.org/10.1099/mic.0.055434-0

Olugbenga, M. (2024). Retrieved 06.12.2024 from https://neptune.ai/blog/balanced-accuracy

Othman, M. F. B., Abdullah, N. B., & Kamal, N. F. B. (2011). MRI brain classification using support vector machine. 2011 fourth international conference on modeling, simulation and applied optimization, Kuala Lumpur, Malaysia, 2011, pp. 1-4. https://doi.org/10.1109/ICMSAO.2011.5775605

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217-222. https://doi.org/10.1080/01431160412331269698

Panaiotov, S., Hodzhev, Y., Tsafarova, B., Tolchkov, V., & Kalfin, R. (2021). Culturable and Non-Culturable Blood Microbiota of Healthy Individuals. Microorganisms, 9(7). https://doi.org/10.3390/microorganisms9071464

Porras, A. M., & Brito, I. L. (2019). The internationalization of human microbiome research. Curr Opin Microbiol, 50, 50-55. https://doi.org/10.1016/j.mib.2019.09.012

Potgieter, M., Bester, J., Kell, D. B., & Pretorius, E. (2015). The dormant blood microbiome in chronic, inflammatory diseases. FEMS Microbiology Reviews, 39(4), 567-591. https://doi.org/10.1093/femsre/fuv013

Rai, K., Devi, M. S., & Guleria, A. (2016). Decision tree based algorithm for intrusion detection. International Journal of Advanced Networking and Applications, 7(4), 2828.

Rana, S., Midi, H. B., & Sarkar, S. K. (2010). Validation and Performance Analysis of Binary Logistic Regression Model.

Requena, T., & Velasco, M. (2021). The human microbiome in sickness and in health. Rev Clin Esp (Barc), 221(4), 233-240. https://doi.org/10.1016/j.rceng.2019.07.018

Saboo, K., Petrakov, N. V., Shamsaddini, A., Fagan, A., Gavis, E. A., Sikaroodi, M., McGeorge, S., Gillevet, P. M., Iyer, R. K., & Bajaj, J. S. (2022). Stool microbiota are superior to saliva in distinguishing cirrhosis and hepatic encephalopathy using machine learning. J Hepatol, 76(3), 600-607. https://doi.org/10.1016/j.jhep.2021.11.011

Sathiyanarayanan, P., Pavithra, S., M. Sai, S., & Makeswari, M. (2019, 29-30 March 2019). Identification of Breast Cancer Using The Decision Tree Algorithm. 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN). https://doi.org/10.1109/ICSCAN.2019.8878757

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097.

Srivastava, T. (2024). Retrieved 06.12.2024 from https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

Tallón-Ballesteros, A. J., Fong, S., & Leal-Díaz, R. (2019). Does the order of attributes play an important role in classification? Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain, September 4–6, 2019, Proceedings 14. https://doi.org/10.1007/978-3-030-29859-3_32

Teixeira, M., Silva, F., Ferreira, R. M., Pereira, T., Figueiredo, C., & Oliveira, H. P. (2024). A review of machine learning methods for cancer characterization from microbiome data. NPJ Precision Oncology, 8(1), 123. https://doi.org/10.1038/s41698-024-00617-7

Timsit, J. F., Ruppe, E., Barbier, F., Tabah, A., & Bassetti, M. (2020). Bloodstream infections in critically ill patients: an expert statement. Intensive Care Med, 46(2), 266-284. https://doi.org/10.1007/s00134-020-05950-6

Topçuoğlu, B. D., Lapp, Z., Sovacool, K. L., Snitkin, E., Wiens, J., & Schloss, P. D. (2021). mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines. J Open Source Softw, 6(61). https://doi.org/10.21105/joss.03073

Topçuoğlu, B. D., Lesniak, N. A., Ruffin IV, M. T., Wiens, J., & Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. MBio, 11(3). https://doi.org/10.1128/mbio.00434-20

Tsafarova, B., Hodzhev, Y., Yordanov, G., Tolchkov, V., Kalfin, R., & Panaiotov, S. (2022). Morphology of blood microbiota in healthy individuals assessed by light and electron microscopy. Front Cell Infect Microbiol, 12, 1091341. https://doi.org/10.3389/fcimb.2022.1091341

Wang, X.-W., & Liu, Y.-Y. (2020). Comparative study of classifiers for human microbiome data. Medicine in microecology, 4, 100013. https://doi.org/10.1016/j.medmic.2020.100013

Wilhelm, R. C., van Es, H. M., & Buckley, D. H. (2022). Predicting measures of soil health using the microbiome and supervised machine learning. Soil Biology and Biochemistry, 164, 108472. https://doi.org/10.1016/j.soilbio.2021.108472

Wu, Q., & Zhou, D.-X. (2006). Analysis of support vector machine classification. Journal of Computational Analysis & Applications, 8(2).

Zheng, D., Liwinski, T., & Elinav, E. (2020). Interaction between microbiota and immunity in health and disease. Cell Res, 30(6), 492-506. https://doi.org/10.1038/s41422-020-0332-7