UCLouvain

École polytechnique de Louvain

Leveraging Machine Learning for Optimal Gut Health

Author: **Simon FRANCOIS**
Supervisor: **Benoît MACQ**
Readers: **John LEE, Mathias MARQUES**
Academic year 2024–2025
Master [120] in Biomedical Engineering

Acknowledgements

First and foremost, I would like to express my sincere gratitude to everyone who supported me throughout this journey.

I am especially thankful to my supervisor, Professor Benoit Macq, for his guidance, support, and expertise throughout this thesis.

I also want to sincerely thank Mathias Marques and Manon Dausort for the time they dedicated, their constant availability, and their invaluable support during the whole process. Without your help and motivation, this work would probably not have been finished yet.

I would like to acknowledge Julie Vandekerckhove for her support and all her ideas.

Of course, I cannot finish this section without a special thanks to my friends and family, for all the encouragement and support. Specifically, I extend my thanks to Anne Frenay and Rudy Francois, my parents, for their careful reading of the last version of this thesis.

Abstract

Context: The equine gut microbiome plays an important role in horse health, with an influence on digestion, immunity, and the onset of diseases such as colitis, colic, and laminitis. However, leveraging this data for disease prediction is unexplored in veterinary contexts.

Objective: This thesis aims to evaluate whether machine learning algorithms can predict gastrointestinal diseases in horses based on their gut microbiome and how different preprocessing pipelines (QIIME2 and Kraken2) impact the classification performance across models.

Methods: Publicly available 16S rRNA sequencing datasets were processed using two pipelines: QIIME2, which focuses on precision through denoising, and Kraken2, offering rapid taxonomic classification using multiple reference databases. Features were extracted and analyzed using dimensionality reduction (PCA, t-SNE), and classification was performed using classical models (Logistic Regression, SVM, Random Forest, MLP), as well as deep learning models (1D CNN, Graph Neural Network).

Results: The best multiclass accuracy (78.4%) was obtained with an SVM on QIIME2 data. Random Forest achieved the best performance in binary classification (84.8% accuracy, AUC = 0.95). CNN performed poorly, while the best GNN achieved up to 71.8% in multiclass and 82.0% in binary classification.

Conclusion: SVM and Random Forest outperformed deep learning models. QIIME2 preprocessing consistently led to better results. Key microbial taxa linked to disease were identified, confirming the potential of microbiome-based prediction despite current limitations in dataset size, class imbalance, and biological variability.

Contents

List of Figures

1 Introduction

The health of horses, but also of other mammals, is closely related to their gut health. Numerous microorganisms, including bacteria, fungi, viruses, and archaea, play a significant role in several biological processes such as digestion, immune regulation but also resistance to infection. In recent years, with the help of constantly improving high-throughput sequencing technologies, researchers have been able to better characterize and understand this complex microbial ecosystem. This gut microbiome is composed of trillions of microorganisms living in the digestive tract.

Moreover, research has revealed the important role of gut health in humans. As of now it is known that gut microbiome influences metabolic functions, neurodevelopment, and immune homeostasis. A disruption in its composition (known as dysbiosis) is associated with a large range of pathologies, including obesity, type 2 diabetes, inflammatory bowel disease, allergies, and even neuropsychiatric conditions such as depression and anxiety [1, 2, 3].
One of the most unexpected finding is the communication axis between the gut and the brain, named brain-gut axis, showing that gut microorganisms can modulate behavior and neural functions [3]. These numerous discoveries have created an explosion of the interest in therapeutic strategies aiming to modulate the microbiota.

With regards to horses, the microbiota of the gastrointestinal tract may play an even more critical role in horses. As large non-ruminant herbivores, horses depend heavily on microbial fermentation for their energy needs. Dysbiosis can cause disorders, from local gastrointestinal disturbances such as colitis, colic, or diarrhea to diseases such as laminitis or obesity that affect multiple organ systems and have consequences throughout the body. More importantly, these conditions often begin with small changes in the microbial composition that can go unnoticed without precise microbiome analysis. If left untreated, the consequences for the animal can be serious, including chronic pain, loss of function, and even death [4].

Following the diagnosis, multiple options are then available to improve the gut health of the horses. The first option is dietary changes, such as increasing fiber content or supplementing with prebiotics and probiotics, that can help rebalance the microbiota. Furthermore, fecal microbiota transplantation shows promising results, but further research is needed to assess its safety and efficacy, especially in equines [5].

Therefore, early diagnosis of intestinal dysbiosis is essential. Thanks to techniques such as 16S rRNA sequencing and metagenomic analysis, powerful tools are now available to analyze equine microbiomes to identify and detect alterations in the abundance of key bacterial taxa. The primary goal of this thesis is hence to identify bacterial taxa playing an important role in equine pathologies and to be able to classify between diseases using the gut metagenomic information. The term "classify" is used to refer to the ability to predict whether a horse is healthy or affected by a specific pathology (e.g., colic, colitis, laminitis).

The use of machine learning and bioinformatics tools allows us to even go one step further, by analyzing complex microbial data and uncovering disease-specific patterns that may be unidentifiable to traditional statistical methods. For example, the TaxoNN model and graph neural network-based classifiers have demonstrated potential in predicting disease states based on human gut microbiota profiles [6, 7]. Similar techniques as well as classical machine learning algorithms will be explored in a dataset containing data from the equine microbiome found in multiple studies.

By improving the ability to diagnose and understand equine diseases through gut microbiota analysis, broader efforts in veterinary precision medicine are also contributed to. Ultimately, this work may aid in the development of early-warning tools and more personalized management strategies for equine health.

2 State of the Art

Some parts of this thesis have been reformulated using ChatGPT [8].

2.1 Key Microbiological Concepts and Definitions

Before diving deeper into the subject, it is needed to define a few terms that will be used throughout this work. Firstly, it is important to explain what what a bacterium and a microorganism are.

Microorganism encompasses all the organisms that can only be seen under a microscope. This includes bacteria, protozoa, algae, fungi and sometimes viruses even though they are not considered as living organisms [9].

Bacteria are microorganisms composed of a single cell, that belong to the group of prokaryotes, which means that they lack a membrane-bound nucleus. A standard bacterium is represented at Figure 2.1. Their DNA floats inside the cell. Bacteria can have many shapes such as spheres, rods, or spirals, and they can live alone or form clusters, chains, or pairs. Some species have flagella for movement while others have capsules for protection. Bacteria are really small (0.5–5 micrometers), but are diverse and perform a lot of essential roles in environments ranging from soil to animal intestines [10].

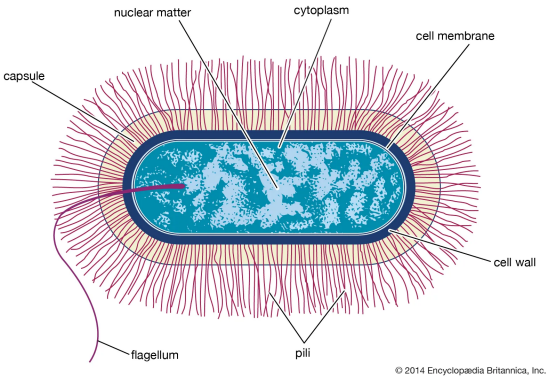


Figure 2.1: Structure of a Bacillus-type Bacterial Cell [10].

Bacteria are organized using a hierarchical system that allows scientists to study species in a systematic and structured manner. This system is known as taxonomy, and the hierarchy it follows is called taxonomic classification, which ranges from broad groups to more specific categories. Additionally, it helps to identify bacteria and understand their relationships, as illustrated in Figure 2.2.

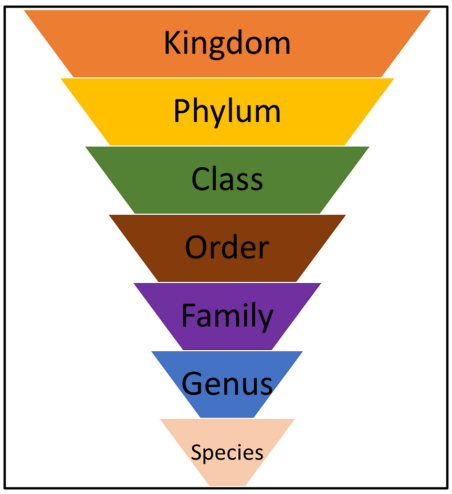


Figure 2.2: Hierarchical taxonomy of bacteria [11].

Archaea will be used in the analyses as well. Although they look like bacteria, they form a separate domain. They are also prokaryotic and single-celled, but their biochemistry and genetics differ significantly from those of bacteria. For example, cell walls of archaea do not contain peptidoglycan, which is a key component of bacterial cell walls [10].

The second confusion that can exist is between the words **Microbiota** and **Microbiome**. These words are often interchangeable but there is a difference that needs to be made.

- **Microbiota**: This term can be defined as a collection of microorganisms residing in a defined environment, such as the gut, skin, or even the respiratory tract. It includes thousands of bacterial species and other microorganisms that often form complex networks. This work will mainly focus on the microbiota of the horses [12].
- **Microbiome**: The microbiome refers not only to the microorganisms themselves but also their collective genomes (the complete set of genetic material of an organism) and the surrounding environmental conditions. It therefore has a broader definition [12].

2.2 Microbiota

2.2.1 Review of studies about mammalian microbiota

In recent years much research over mammals has shown that mammalian microbiota plays a significant role in their overall health. Here below is a small review of the different studies existing on the subject, showing the underestimated role of the microbiota.

Ridaura et al. (2013) showed that the gut microbiota of obese humans could induce fat gain in mice. Furthermore, when mixed with lean microbiota the process was reversed. The study also showed that diet strongly influenced this effect [13].

Hsiao et al. (2013) demonstrated that disruptions in the gut microbiota could modify behaviors associated with autism spectrum disorder in a mouse. The symptoms of autism were diminished when *Bacteroides fragilis* was administered orally [14].

Belkaid et al. (2014) explain that the microbiota plays an important role in shaping the host immune system, maintaining a balance between defense and tolerance. However, modern changes like the overuse of antibiotics and dietary shifts have altered this equilibrium, which causes more autoimmune and inflammatory diseases [15].

Cani et al. (2007) highlighted that gut barrier dysfunction leads to metabolic endotoxemia (a leak of bacterial toxins in the blood), contributing to obesity and insulin resistance [16].

Sampson et al. (2016) demonstrated that gut microorganisms are able to modulate neuroinflammation and motor symptoms in a Parkinson's disease mouse model, showing a microbiota–brain connection [17].

2.2.2 Overview of horse microbiota

The gastrointestinal tract of horses contains a complex microbial ecosystem for their digestion and immune modulation with up to a quadrillion bacterial cells [18]. Horses are non-ruminant herbivores, which means that they mostly rely on microbial fermentation to extract energy from the plant fibers.

The composition of their microbiota varies throughout the gastrointestinal tract. In the foregut (stomach and small intestine), the transit is rapid, favoring microbes that are able to metabolize starches, sugars, and proteins. In that part, the microbiota is dominated by Firmicutes (e.g., Lactobacillaceae, Streptococcaceae) and Proteobacteria (e.g., Pasteurellaceae). Contrarily, the hindgut (composed of caecum and colon), where the fermentation of fibers happens, contains more Firmicutes and Bacteroidetes. The most important bacterial families include Ruminococcaceae, Lachnospiraceae, and Prevotellaceae, specializing mostly in fiber degradation [4].

Many factors can influence the composition and function of the equine gut microbiota. The composition of the diet is an important factor: horses naturally adapted to high-fiber diets often experience dysbiosis when fed with high-starch diets, which promote an overgrowth of lactic acid-producing bacteria, increasing the risk of diseases. Even small shifts of diet can induce a dysbiosis. Transportation stress is critical as well, because of how it can lead to a decreased microbial diversity but also a reduced abundance in bacteria that degrade fibers. In the same way, intense physical activity in performance horses alters the gut microbiota. Additionally, heat stress affects microbial communities by reducing fibrolytic populations and promoting species adapted to stressful metabolic conditions. All these findings highlight how sensitive the microbiota is to environmental conditions [4].

Although sensitive, the microbiota can be modulated using various strategies:

- Probiotics, like inulin or fructooligosaccharides can help promote the growth of beneficial bacteria.
- Prebiotics are non-digestible compounds, typically fibers, that selectively stimulate the growth and activity of beneficial gut bacteria, contributing to host health.
- Fecal microbiota transplantation, consisting in transferring fecal material from a healthy donor, is a promising approach to restore microbiota after a dysbiosis.

However, more research is needed to standardize these interventions and confirm their long-term benefits. Moreover, the diet is obviously an important approach to keep a good balance, with a need to keep high fiber intakes [19].

2.2.3 Analysis of microbiota

One of the most adopted technologies in current sequencing and in the data used for this thesis, is Illumina sequencing, which uses the sequencing-by-synthesis method. Firstly, the genetic material needs to be isolated from the samples, ensuring high quality of the nucleic acids. Then, the DNA is prepared for sequencing by fragmenting, and adapters (special sequences) are ligated to the end to allow sequencing. These adapters ensure a good attachment to the flow cell and enable sample identification. The DNA fragments are first attached to a solid surface called a flow cell. Through a process known as bridge amplification, each fragment bends and binds to nearby primers on the surface, forming a bridge. These bridges are then amplified through repeated cycles, creating dense clusters of identical DNA sequences in localized spots on the flow cell. The sequencing-by-synthesis can begin, with each DNA strand extended one base at a time using fluorescently labeled nucleotides. The emitted signal is captured by a camera to determine the DNA sequence. Finally, bioinformatic tools are used to analyze the light emitted and produce the sequences. The key steps of this process are illustrated in Figure 2.3 [20].

Illumina sequencing is highly accurate, cost-effective and particularly fast. However, the reads produced are relatively small, making it harder to assemble. Furthermore, it produces large volumes of data making it hard to process and the initial investment needed is important due to the high cost of sequencing machines and supporting laboratory infrastructure [20].

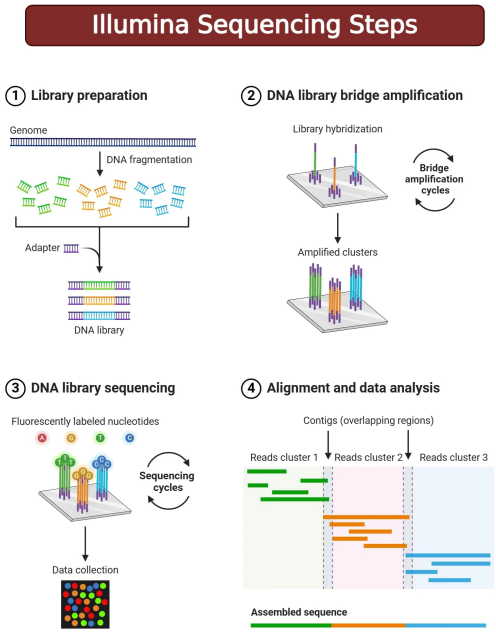


Figure 2.3: Overview of the Illumina sequencing process [20].

In the context of this work, Illumina sequencing is employed specifically for 16S rRNA gene sequencing, a technique commonly used to study microbial diversity. The 16S rRNA gene encodes the 16S ribosomal RNA, which is a structural component of the 30S subunit of the bacterial ribosome. This gene is interesting because it can be found in all bacteria and archaea, and plays an essential role in protein synthesis. Its function has also remained stable through evolution which is an important feature to assess evolutionary divergence and therefore taxonomy. Furthermore, its size is about 1,500 base pairs long, which provides enough variability for meaningful comparisons but is still short enough to be easily sequenced. However, although useful, this technique has limitations because in some cases, even different species can share up to 97% similarity in their sequences. Additionally, some bacteria have multiple copies of the 16S rRNA gene that are not identical, which might complicate the analysis [21].

2.3 Equine Diseases

In this work 3 different diseases will be analyzed and used for classification.

2.3.1 Laminitis

The first pathology that will be explored is equine laminitis. It is a common and multifactorial disease that is extremely painful and can even be life-threatening. Over 7% of equine deaths are attributed to laminitis, with euthanasia frequently being the outcome [22]. Recent studies have established links between the gut microbiota and its pathogenesis [4].

Laminitis is characterized by the inflammation and degeneration of the laminae, which is the tissue between the hoof wall and the coffin bone as shown in Figure 2.4. This inflammation then causes the laminae of the hoof to separate causing severe damage to the bones and soft tissue [23]. Horses affected by laminitis often stand awkwardly to ease pain and may be very reluctant to take steps.

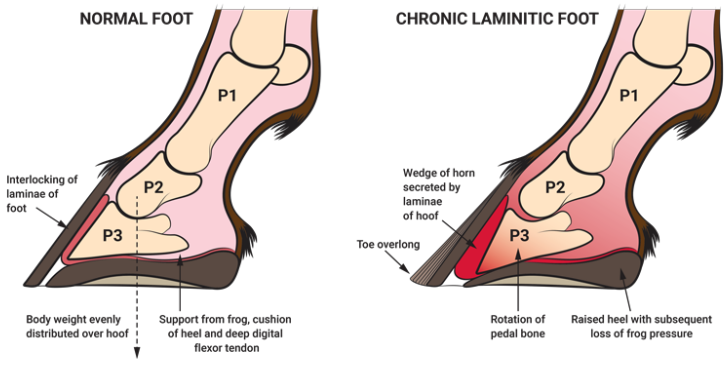


Figure 2.4: Comparison between a normal equine foot and one affected by chronic laminitis [24].

Laminitis may arise from three primary pathways: mechanical overload, systemic inflammation, and metabolic dysfunction. Overload laminitis, which is less common, happens when the horse is unable to bear weight on one limb and therefore overloads the other limb. Inflammatory laminitis occurs secondary to systemic illness, often triggered by high-starch grain overload. When large amounts of starch bypass digestion they may ferment in the hindgut, which causes a dysbiosis, and leads to the release of toxins in the bloodstream, ultimately causing inflammation

within the hoof. Finally, metabolic laminitis is linked to conditions like pituitary pars intermedia dysfunction and equine metabolic syndrome, where horses show abnormal insulin regulation. These disorders cause an excessive insulin response to starches and sugars, which ends up increasing the risk of laminitis [25]. The risk of laminitis is higher for older and obese horses and those previously affected by this condition [26].

This condition is not directly related to the microbiota. However, high-starch diets can result in lactic acidosis which causes a disruption in the hindgut microbiota. Such an acid environment can cause the death of the bacteria that digest the fiber and also causes the proliferation of bacteria that produce lactic acid like Streptococcus and Lactobacillus species. In contrast, a decrease in other taxa such as Akkermansia, Ruminococcaceae, and Phascolarctobacterium occurs in laminitis, suggesting a possible protective role of those bacteria [4].

The diagnosis of laminitis can be made using clinical signs and history. However, diagnosis can be challenging when only the hindlimbs are affected, as symptoms may mimic neurological disorders. Radiographic signs are often subtle in the early phases, showing only mild laminar swelling. This assesses the needs of other diagnosis methods such as microbiome biomarkers detection. In the chronic cases, changes in the hoof capsule and bone displacement become more visible [27].

There exists no treatment working in all cases for laminitis. However, early treatment can improve the prognosis. Firstly, dietary changes include reducing sugars and starch (non-structural carbohydrates), avoiding rich spring grass, feeding low-sugar hay while ensuring gradual dietary transitions, and maintaining high-fiber intake. Drugs can be given to control the pain and improve blood flow. Foot support is important to reduce the load and reduce the pain too. Finally, cryotherapy is especially effective to reduce laminar change [27].

2.3.2 Colitis

The second condition explored in this work is equine colitis. Colitis refers to an inflammation of the colon and can lead to severe diarrhea, dehydration, systemic inflammation, and even death. It is therefore a medical emergency even though the severity and prognosis depend on the cause and how quickly a treatment is administered [28]. Colitis, acute or chronic, compromises the mucosal integrity of the large intestine, affecting not only fluid and electrolyte absorption but increasing the permeability of the gut wall to bacteria and toxins as well.

The symptoms often include important diarrhea, fever, anorexia, depression, and signs of abdominal pain. Its diagnosis is not an easy task and tests can go from physical examinations and blood tests to fecal analysis, and even imaging [29].

The etiology of colitis is multifactorial, and the causes are both infectious and non-infectious. Infectious agents can be bacteria, viruses or parasites. Infectious bacteria, that are interesting for this work, are Salmonella, Clostridium difficile and Escherichia Coli. Non-infectious causes are broad and include antibiotic induced dysbiosis, dietary imbalances or sand irritation of the large bowel [28].

Horses affected by colitis exhibit significant alterations in their fecal microbiome, including decreased diversity and shifts in bacterial populations. Healthy horses have a microbiome rich in Firmicutes, especially Clostridia and Lachnospiraceae, while colitic horses show increased Bacteroidetes, Proteobacteria, and Fusobacteria. All these changes reflect a loss of the microbial balance rather than the overgrowth of a specific pathogen [30].

Most often, it is not needed to be aware of the cause of the colitis to start a treatment. The only requirement is to know if the horse is contagious or not. Until proven otherwise, it must be treated as such, with relevant biosecurity measures [29]. Supportive care can begin immediately and focuses on stabilizing the horse while limiting further damage. This includes fluid and electrolyte therapy, inflammation control, and nutritional support. In more severe cases, additional interventions such as plasma transfusions, analgesics, anti-diarrheal agents, and microbiome support may be required. With fast and appropriate care, clinical improvement is often seen within a few days, provided that complications are avoided [28].

2.3.3 Colic

The third and last condition used in this thesis is colic. This term refers more to a clinical syndrome, with different origins and severity than to an actual disease. The common symptom is the abdominal pain of the horses. It is one of the most common emergencies in equine medicine and a leading cause of death in horses if not treated fast enough [31].

Horses with colic may show a variety of clinical signs such as pawing, flank watching, lip curling, rolling, sweating, and loss of appetite. In some cases, symptoms can escalate to profuse sweating, rapid heart rate, and signs of shock. These signs indicate abdominal pain but do not reveal the specific location or severity of the issue, nor whether surgery will be required [31].

The assessment of colic includes rectal examination, ultrasound scan but also palpation of the area concerned. The use of samples of fluid collected in the horses' stomach can help the diagnosis too [32].

Because of the broad definition of the colic the causes are numerous. One of the common causes is the obstruction of the colon by dehydrated ingesta and occasionally with sand. Gas colic is a type of colic in which gas builds up in the colon causing discomfort. Mechanical causes may also involve displacement of the colon. Other factors can occur such as parasites, sudden feed changes and even ulcers [33].

Some recent studies suggest that colic in horses is linked to changes in the gut microbiota, even though findings are inconsistent. The horses with colic show an increased abundance of Proteobacteria, including potential pathogens such as Escherichia coli and Acinetobacter, and lactate-producing bacteria like Streptococcus and Lactobacillus, as seen before with the colitis. This microbial shift, showing increased lactate and lower pH, may inhibit Fibrobacter, Ruminococcus, and Methanobrevibacter, causing problems for the fiber fermentation and the overall gut stability [4].

Treatment of colic often depends on the cause but also on the severity of pain. In mild cases of colic, the veterinarian may give medication to relieve the pain and help the horse to relax, which ultimately can help restore normal gut function. Some fluids and treatments are sometimes delivered directly to the stomach using a nasogastric tube. If the symptoms persist despite the initial treatment, some surgery may be advised, which requires the horse to be taken to an equine hospital [34].

Some prevention must be taken to avoid colic and promote a good gastrointestinal health. The prevention focuses on good management: provide fresh water, regular pasture access, avoid feeding on sand, and make any dietary change slowly. Simple strategies such as a good dental care routine and effective parasite control are essential steps to reduce the colic risk. The horses that have a history with colic must be monitored even more closely [33].

2.4 Preprocessing and visualization

2.4.1 Scaling, Splitting, and Cross-Validation

In most studies involving supervised machine learning, it is common to partition the dataset into two distinct subsets: a training set and a test set. A widely adopted strategy is to allocate 80% of the data to the training set and the remaining 20%

to the test set, which remains untouched throughout model tuning to ensure an unbiased final evaluation [35].

To mitigate overfitting and to ensure robust hyperparameter tuning, stratified k-fold cross-validation with $k = 5$ is frequently applied to the training set. In this procedure, the training set is divided into k equal parts (folds), and the model is trained and validated k separate times, each time using a different fold for validation and the others for training. This technique provides a reliable estimate of model performance and assists in selecting optimal hyperparameters. An illustration of this process is shown in Figure 2.5 [35].

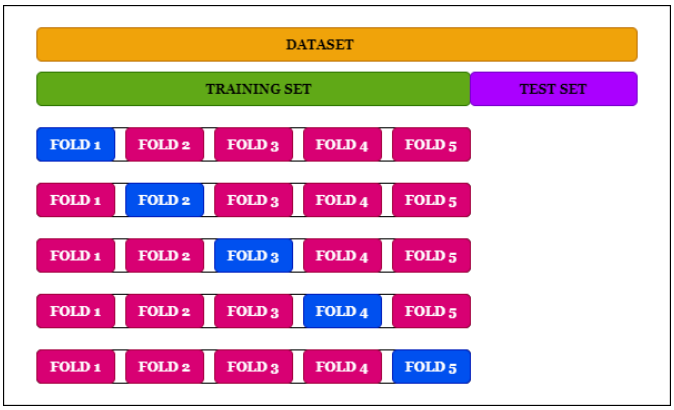


Figure 2.5: Illustration of a common data splitting strategy: the test set is held out for final evaluation, while the training set undergoes k-fold cross-validation for hyperparameter tuning [36].

Prior to training, feature standardization is generally applied to ensure that all input features contribute equally to the learning process. This preprocessing step is crucial for gradient-based algorithms, as it improves convergence behavior and model stability. The most widely used approach is z-score standardization, which transforms each feature to have zero mean and unit variance:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{2.1}$$

here, x_i is the value of a given feature for sample i, μ is the feature's mean, and σ its standard deviation. To prevent data leakage, μ and σ are computed solely on the training set, and the same transformation is applied to the test set [37].

2.4.2 Principal Component Analysis (PCA)

Principal Component Analysis helps with visualization and dimensionality reduction of data. PCA is a linear method used to reduce the dimensionality of a dataset while preserving the most variance possible. The first principal component captures the direction along which the data varies the most. The following components are chosen to capture as much of the remaining variance as possible, with each being orthogonal to the previous ones as shown in Figure 2.6.

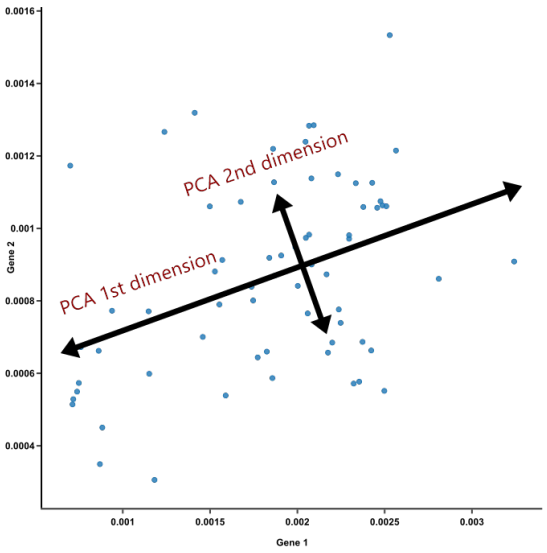


Figure 2.6: Graphical representation of PCA [38].

If the dataset is centered $X \in \mathbb{R}^{M \times N}$ (where M is the number of features and N the number of observations), PCA aims to find an orthogonal transformation $Y = WX$ such that the covariance matrix of Y is diagonal, implying that all features are linearly uncorrelated [39].

The covariance matrix Σ measures how much pairs of features in the dataset vary together. Diagonalizing this matrix means that the new features (principal components) are statistically uncorrelated.

To achieve this goal, the eigenvalue decomposition of the covariance matrix is computed:

$$\Sigma = \frac{1}{N} X X^\top = V \Lambda V^\top$$

where V is an orthogonal matrix, whose columns are the eigenvectors (principal components), and Λ is a diagonal matrix containing the eigenvalues which represent the amount of variance explained by each principal component.

To reduce dimensionality, only the $P < M$ leading eigenvectors are selected:

$$W = V_P^\top \quad \text{with} \quad V_P = [v_1, \ldots, v_P]$$

and project the data as:

$$Y = V_P^\top X.$$

This projection captures the directions of maximal variance [39].

2.4.3 t-distributed Stochastic Neighbor Embedding (t-SNE)

T-distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique [40]. It is well-suited for the visualization of high-dimensional datasets in two or three dimensions. The t-SNE algorithm converts the similarities between data points $\{x_1, \ldots, x_n\}$ in the high-dimensional space into joint probability distributions and then attempts to keep these similarities for their low-dimensional counterparts $\{y_1, \ldots, y_n\}$. In the high-dimensional space, the conditional probability that x_j is a neighbor of x_i is given by:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

where σ_i is the variance of the gaussian that is centered on the datapoint x_i.

In the low-dimensional map, the similarities are modeled with a heavy-tailed Student-t distribution, having one degree of freedom, giving the following joint probabilities q_{ij}:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}.$$

t-SNE minimizes the following Kullback-Leibler divergence between $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$ using a gradient descent method:

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

This minimization is typically performed using gradient descent [40].

To further analyze the latent structures obtained through t-SNE, it is possible to use clustering algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [41]. DBSCAN groups together samples that are closely packed in the reduced space and labels isolated points as noise, allowing for flexible detection of heterogeneous or irregular cluster shapes.

2.4.4 VarianceThreshold

VarianceThreshold is a simple and commonly used feature selection method that removes all features whose variance does not exceed a given threshold. The underlying assumption is that features with very low variance do not carry significant discriminative information and may only add noise to the model. This method is useful as a baseline filter in high-dimensional datasets such as microbiome profiles, where many features appear in very few samples and can be discarded without loss of relevant signal.

In practice, the variance of each feature is computed independently, and features below a user-defined threshold are excluded. While this approach does not consider the relationship with the target variable, it offers a fast and interpretable way to reduce dimensionality before applying more complex preprocessing [42].

2.5 Classification Algorithms

As stated earlier, the final goal of this thesis is to classify the horses between the different diseases using different machine learning algorithms. Here is an overview of the different algorithms that will be used.

2.5.1 Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for classification. When given an input vector $x = [x_1, x_2, \ldots, x_n]$, Logistic Regression computes a linear combination:

$$z = w \cdot x + b$$

where $w = [w_1, w_2, \ldots, w_n]$ are the weights and b is the bias.

The final goal being to map z to a probability between 0 and 1, the sigmoid function is applied, its definition being:

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

Knowing that the predicted probability $\sigma(z)$ represents $P(y = 1 \mid x)$, the decision rule is therefore:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y = 1 \mid x) > 0.5 \\ \\ 0 & \text{otherwise} \end{cases}$$

The parameters w and b are learned by minimizing the binary cross-entropy loss function:

$$L_{CE}(\hat{y}, y) = -\left(y \log(\hat{y}) + (1-y) \log(1-\hat{y})\right)$$

where $\hat{y} = \sigma(w \cdot x + b)$ is the model's prediction [43].

The optimization is performed using gradient descent:

$$\theta \leftarrow \theta - \eta \nabla L(\theta)$$

where η is the learning rate and $\theta = \{w, b\}$ are the model parameters.

Logistic Regression can be generalized to more than two classes, which is useful for this work. In that case, the sigmoid is replaced by the softmax function:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

for $i = 1, \ldots, K$, where K is the number of classes. The model then outputs a probability distribution over all possible classes.

The loss function is also adapted to the multiclass setting using the categorical cross-entropy:

$$L_{CE}(\hat{y}, y) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

where y_i is a one-hot encoded vector representing the true class and $\hat{y}_i$ is the predicted probability for class i [43].

2.5.2 Support Vector Machines

Support Vector Machines (SVM) are other supervised learning models used for classification tasks. They find the optimal separating hyperplane between classes by maximizing the margin, which is the distance between the hyperplane and the nearest data points from each class, that are called the support vectors. The whole model is represented in Figure 2.7.

Given labeled training data (x_i, y_i) with $y_i \in \{-1, 1\}$, SVM solve the following optimization problem for linearly separable data:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 \quad \forall i \end{aligned}$$

here, $w \in \mathbb{R}^n$ is the weight vector orthogonal to the separating hyperplane, and $b \in \mathbb{R}$ is the bias term that shifts the hyperplane. The decision function is given by $\hat{y} = \text{sign}(w^T x + b)$.

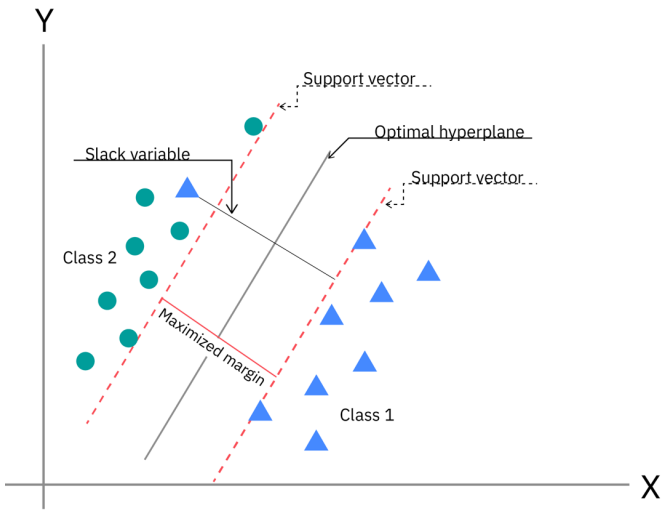


Figure 2.7: Graphical representation of a SVM [44].

This problem is convex and guarantees a unique global minimum, it is called a large-margin classifier in opposition to the next type.

When data are not linearly separable, slack variables $\xi_i \geq 0$ are introduced to allow some misclassifications, those are called SVM, soft margin classifiers:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \end{aligned}$$

where C is a parameter that controls the trade-off between maximizing the margin and minimizing the classification error [45].

To handle non-linearly separable data, SVMs can be extended using the kernel trick, which maps the input data into a higher-dimensional feature space where a linear separation is possible. Instead of explicitly computing this transformation, a kernel function $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is used to compute inner products in the feature space efficiently.

Commonly used kernels include:

- **Linear kernel:** $K(x_i, x_j) = x_i^T x_j$.
- **Polynomial kernel:** $K(x_i, x_j) = (x_i^T x_j + c)^d$, where c and d are constants.
- **Radial Basis Function (RBF) or Gaussian kernel:** $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$, with $\gamma > 0$ [46].

Standard SVMs are binary classifiers. However, for the multi-class classification needed, several strategies are possible:

- **One-versus-All (OvA)**: Train one classifier per class that will discriminate that class from all the others. The class with the highest confidence is then chosen.
- **One-versus-One (OvO)**: Train one classifier for every pair of classes. A relevant strategy is used to determine the final prediction.
- **True Multiclass SVMs**: Instead of decomposing the problem into binary tasks, the optimization formulation is extended to directly handle multiple classes. In practice, libraries such as scikit-learn implement this by solving a single optimization problem that incorporates all classes simultaneously. This formulation learns a separate weight vector w_k and bias b_k for each class $k \in \{1, \ldots, K\}$, and predicts the class with the highest score:

  $$\hat{y} = \arg\max_k \left(w_k^T x + b_k \right).$$

  The optimization problem minimizes the sum of hinge losses over all training examples and all incorrect classes, enforcing that the correct class' score is sufficiently higher than the others [46].

2.5.3 Random Forest Classifier

The Random Forest is a supervised machine learning algorithm that combines multiple decision trees to improve classification accuracy. Each tree is built using a random subset of the training data and a random subset of features. The final prediction is then made by computing the predictions from all trees, usually using majority voting. They have three hyperparameters including node size, number of trees, and the number of features sampled [47].

Each tree is trained on a different dataset chosen by sampling from the original dataset with replacement to introduce diversity. Each decision tree is grown until its full depth without using pruning, which allows the tree to capture the complex patterns specific to each sample. The advantages are numerous. First, it is really robust to overfitting because of the number of trees used. Moreover, it handles well missing data. Finally, it provides an estimate of the importance of the features [48, 47].

2.5.4 Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron is a type of artificial neural network used for regression and classification tasks, its graphical representation is available in Figure 2.8.

It consists of multiple layers of neurons: an input layer, one or more hidden layers, and an output layer. Each neuron computes a weighted sum of its inputs, applies a nonlinear activation function, and propagates the result to the next layer.

An MLP can approximate any continuous function with only two layers of weights, considering that the number of hidden units is large enough [49].

The general form of an MLP output is:

$$y = g\left(\sum_{i=0}^{M} w_i^{(2)} h\left(\sum_{j=0}^{D} w_{ij}^{(1)} x_j\right)\right)$$

where:

- x_j are the input features,
- $w_{ij}^{(1)}$ and $w_i^{(2)}$ are weights of the first and second layers,
- $h(\cdot)$ is the hidden layer activation function (commonly sigmoid or tanh),
- $g(\cdot)$ is the output activation.

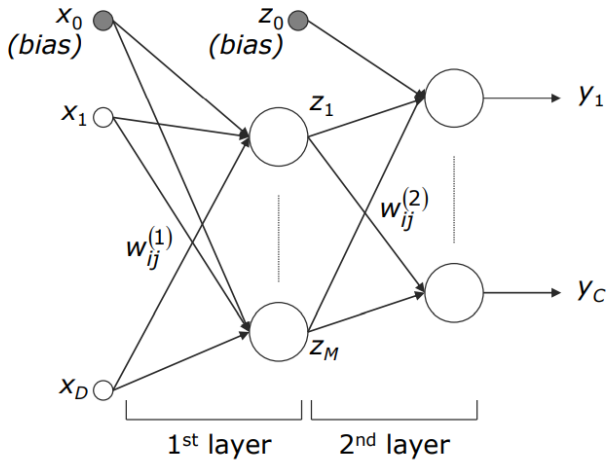


Figure 2.8: Graphical representation of an MLP [50].

Learning in MLPs involves minimizing an error function (for example cross-entropy for classification) by adjusting the weights via gradient descent. This is performed using the backpropagation algorithm. It computes the gradient of the error for each weight by applying the chain rule from the output layer back to the input recursively [49].

2.5.5 Convolutional Neural Network (CNN)

Convolutional Neural Networks are a class of deep neural networks designed to process data with a grid-like topology, such as images or temporal sequences. They use convolutional layers, where filters (called kernels) slide over the input to detect local patterns. This weight sharing mechanism helps reduce the number of parameters compared to fully connected layers.

CNNs usually consist of a sequence of convolutional layers, non-linear activations (e.g., ReLU), pooling layers (such as max or average pooling), and fully connected layers at the end for classification. The early layers capture low-level features like edges or textures, while deeper layers learn more abstract representations [51].

Figure 2.9 shows an example of a CNN for classification. After the feature learning, the feature maps are flattened and passed through fully connected layers that act as a classifier. The last layer is a softmax layer that outputs a probability distribution over the possible classes.

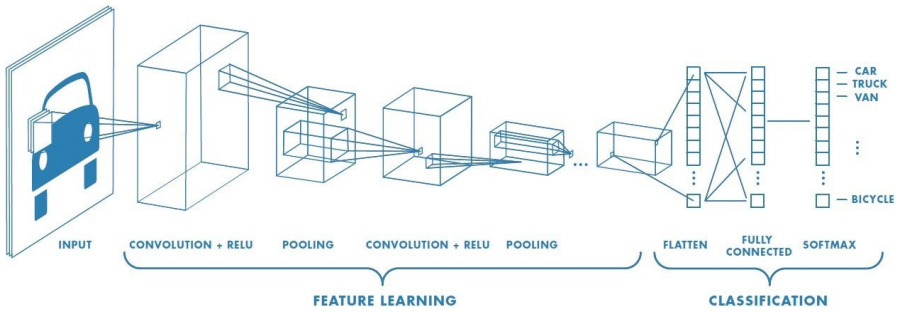


Figure 2.9: Architecture of a typical Convolutional Neural Network [52].

The training of a CNN involves feeding labeled data through the network and then compares the predictions to the true labels using a loss function (typically cross-entropy). The parameters are adjusted by backpropagation and the weights are optimized using algorithms like stochastic gradient descent. Overfitting is avoided by balancing the model complexity and using validation data [53].

Another type of CNN exists for data that are in one dimension, this type of CNN is called 1D CNN, with a general architecture shown in Figure 2.10 . They use 1D filters to capture temporal or sequential patterns along a single axis. These networks apply convolution, activation, and pooling operations to enable an efficient feature extraction directly from raw 1D data. Unlike traditional 2D CNNs designed for image processing, 1D CNNs are optimized for signals such as time series, ECG, or textual sequences. They have shown excellent performance in applications where the data are inherently sequential and limited in dimension, while also being computationally efficient and suitable for real-time applications [54].

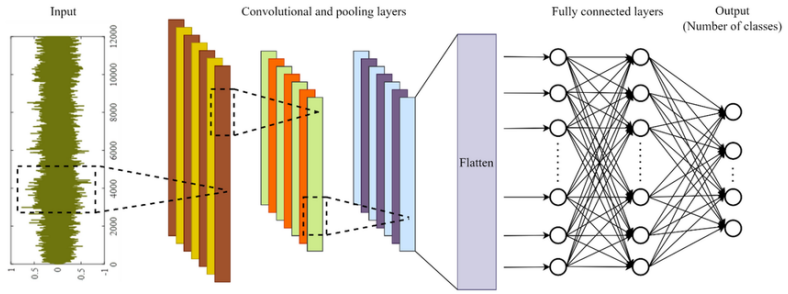


Figure 2.10: Graphical representation of a one dimensional CNN [55].

2.5.6 Graph Neural Network (GNN)

Graph Neural Networks (GNNs) are a class of deep learning models that work on graph-structured data. They use relational information, unlike convolutional neural network, working on grid-like structures, specific to graphs to perform tasks such as node classification, link prediction, and graph classification. A graph is composed of nodes and edges, with three types of attributes: node attributes, edge attributes, and global attributes. A standard GNN architecture is shown in Figure 2.11.

At the core of GNNs is the message passing, where each node receives messages from its neighbors and updates its own representation. A basic propagation rule can be expressed as:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right)$$

where $\tilde{A}$ is the adjacency matrix with self-loops, $\tilde{D}$ the degree matrix, $W^{(l)}$ the trainable weight matrix, and σ an activation function [56].

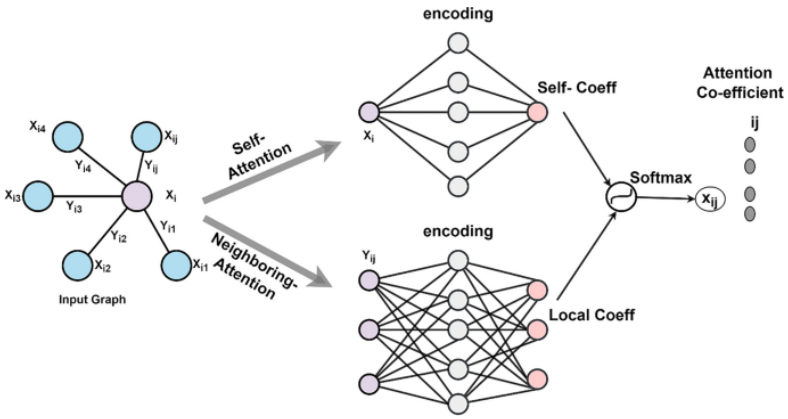


Figure 2.11: Graphical representation of a GNN [55].

Using this kind of data representation allows to perform multiple types of prediction tasks. The goals can be to predict whole graph, node or edge properties. To make predictions at the graph level, the GNNs aggregate node embeddings into a global embedding through a pooling function such as sum, mean, or max pooling:

$$h_G = \text{READOUT}\left(\{h_v : v \in V\}\right).$$

The READOUT function aggregates all node embeddings into a single vector that summarizes the entire graph for downstream prediction tasks.

Despite being relatively new, GNNs have applications in many domains such as graph clustering, molecular fingerprint prediction, traffic state estimation, and even text classification [56].

However, GNNs still face challenges that limit their large adoption. Over-smoothing in deep architectures, high computational costs for large-scale graphs, and limited expressivity in understanding complex structures are active areas of research [56].

2.6 Existing research

Research until now has largely recognized and highlighted the role of microbiota in equine diseases. Biomarkers like the Firmicutes/Proteobacteria and Proteobacteria/Fibrobacteres ratios have been proposed as potential indicators of gut health [4].

The use of machine learning to leverage this microbial information for predictive or diagnostic purposes in horses remains unexplored.

However, some studies exist for human patients, which is informative for this thesis. One example is the TaxoNN, a machine learning approach designed to perform microbiome-based disease prediction in humans. It creates clusters of operational taxonomic units (label for microbes with similar genetic material) and then uses convolutional neural networks within each cluster. Finally, they aggregate the results for each cluster to output a final prediction. When comparing two studies on datasets related to the liver cirrhosis and the type 2 diabetes for humans, this model outperformed conventional techniques, achieving area under curve (AUC) scores up to 0.92 [6].

A second example is Irwin et al., that proposed a method leveraging GNNs to model gut microbiome data, with the goal of predicting human phenotypes such as Inflammatory Bowel Disease. Their model included both metagenomic (abundance of microbial genes) and metatranscriptomic (active expression of these genes) data. Using GNN, they learn node embeddings using different strategies (including Laplacian Eigenvector Positional Encoding (LPE), Random Walk Positional Encoding (RWPE), and Node2Vec (N2V)), showing AUC score up to 0.929 using a simple classifier like SVM. This human-centered research demonstrates the feasibility and potential of graph-based representation learning in microbiome analysis [7].

3 Methods and Material

3.1 Materials

In this Section, the data and its origin, used for the rest of this thesis, are described, as well as how they have been preprocessed for use by the machine learning algorithms.

3.1.1 Dataset

The dataset used for the analysis has been created using multiple independent studies for which the data was made available, through the NCBI Sequence Read Archive (SRA) [57]. It is important to note that collecting data to create a dataset is not part of the scope of this thesis and relies only on publicly available data. These studies include several BioProjects and cover a wide range of conditions, such as colic, colitis, laminitis, and healthy controls, which are the only conditions that will be overviewed. These conditions were chosen because they were frequent, have a link with microbiota and are potentially treatable. In total, 1227 samples were retained. The data available included sequencing data and diagnosis but also sometimes metadata, which was not used for the study for two main reasons. First, it was not always available and could introduce too much sparsity, limiting their utility in such a large dataset. Furthermore, the objective of this work was to focus exclusively on the microbial composition, in order to isolate the role of the gut microbiome in the diagnosis of equine pathologies, without the influence of external factors.

The sequencing data only consists of `FASTQ` files. `FASTQ` are text files specific to biological sequence data (DNA or RNA). Each read in a `FASTQ` file is stored in four consecutive lines: (i) a label line starting with @ that uniquely identifies the read, (ii) a sequence line containing the nucleotide bases, (iii) a *plus line* serving as a separator (optionally repeating the identifier), and (iv) a quality line with ASCII-encoded quality scores for each base. Figure 3.1 illustrates the structure of a typical `FASTQ` entry. In this thesis, there is one `FASTQ` file per sample, with an average size of approximately 170 MB per file. The raw data was downloaded from

the NCBI Sequence Read Archive using the `fastq-dump` command from the SRA Toolkit [57, 58]. This command converts sequencing data from SRA format into standard `FASTQ` format. Only single-end reads were considered, and all files were stored in a local directory.

Figure 3.1: Structure of a `FASTQ` entry: each read is composed of four lines — an identifier (label), the nucleotide sequence, a separator ('+'), and the quality scores encoded as ASCII characters [59].

To build a comprehensive and diverse dataset, raw sequencing data from several publicly available studies listed in the NCBI Sequence Read Archive (SRA) were aggregated. These studies were selected based on their focus on the equine gut microbiome and their availability in `FASTQ` format. They include a wide variety of clinical conditions—such as colic, colitis, laminitis, and healthy controls—as well as different sequencing technologies and experimental designs. Table 3.1 summarizes the main characteristics of each BioProject included in the dataset, including the number of samples, average `FASTQ` file size, read length, and sequencing platform.

- **PRJNA562547**: "Clostridioides difficile carriage in animals and the associated changes in the host fecal microbiota."
  This study investigated the Clostridium difficile strains in various animals and their association with gut microbiota alterations. The paper revealed its significant impact on microbial communities in both dogs and horses, highlighting the potential protective role of Clostridia in blocking infections in the gut [60].

- **PRJNA1032075**: "A novel dataset of 2,362 equine fecal microbiomes from eight veterinary teaching hospitals on three continents reveals dominant effects of geography, breed, and disease."
  This study presents the EGG database, which includes 2,362 fecal samples from 1,190 horses, aiming to analyze how geography, breed, and various

Table 3.1: **Summary of BioProjects**: Number of `FASTQ` files, average file size and read length, and sequencing instrument used.

BioProject	#FASTQ	Avg Size (MB)	Read Length	Instrument
PRJNA562547	31	27.6	501.0	Illumina MiSeq
PRJNA1032075	853	162.1	501.9	Illumina MiSeq
PRJEB47719	36	31.7	373.1	Illumina MiSeq
PRJEB32490	5	61.9	301.0	Illumina MiSeq
PRJNA279335	20	5.3	512.5	454 GS Junior
PRJNA580257	106	79.5	289.5	Illumina MiSeq
PRJNA728793	48	211.1	289.0	Illumina MiSeq
PRJDB17872	55	678.0	3307.3	MinION
PRJNA177883	18	5.4	539.4	454 GS FLX
PRJNA662009	44	34.7	500.0	Illumina MiSeq

diseases impact the equine gut microbiome. Although the dataset originally included a wide range of conditions, only the samples corresponding to the disease categories relevant to this thesis were retained. It is the largest source used to build the dataset, representing approximately two thirds of the final samples [61].

- **PRJEB47719**: "Changes in the gut microbiome and colic in horses: Are they causes or consequences?"
  This review investigates whether alterations in the equine gut microbiome observed during colic are a cause or a consequence, concluding with the need for longitudinal studies [62].

- **PRJEB32490**: "The fecal microbiota of healthy donor horses and geriatric recipients undergoing fecal microbial transplantation for the treatment of diarrhea."
  This study explores Fecal Microbial Transplantation (FMT) in geriatric horses with diarrhea and shows that successful FMT can restore microbiota diversity, becoming more similar to that of the donor horse [63].

- **PRJNA279335**: "Longitudinal Changes in Fecal Microbiota During Hospitalization in Horses With Different Types of Colic."
  This study tracks fecal microbiota changes in hospitalized horses with different types of colic and found microbial differences were more linked to the type of colic than to the hospitalization duration [64].

- **PRJNA580257 and PRJNA728793**: "Alterations in the Fecal Microbiome and Metabolome of Horses with Antimicrobial-Associated Diarrhea Compared to Antibiotic-Treated and Non-Treated Healthy Case Controls."
  This study shows that horses affected with antimicrobial-associated diarrhea have distinct microbial profiles compared to control horses [65].

- **PRJDB17872**: "Minimal disruption of equine gut microbiota by intravenous cephalothin treatment."
  This study shows that cephalothin has a low impact on the gut microbiota compared to other antibiotics, showing that it could be a safer antimicrobial option [66].

- **PRJNA177883**: "Pyrosequencing of 16S rRNA genes in fecal samples reveals high diversity of hindgut microflora in horses and potential links to chronic laminitis."
  This study showed that horses with chronic laminitis had a greater bacterial diversity and distinct Clostridiales taxa when compared to healthy controls [67].

- **PRJNA662009**: "Gut microbiota resilience in horse athletes following holidays out to pasture."
  This study shows that letting elite horse athletes spend time out to pasture improves their well-being and induces long term positive shifts in their gut microbiota composition [68].

The data has been separated in four classes. A numeric label has been assigned for each condition as follows:

- Healthy: 0,
- Colitis: 1,
- Colic: 2,
- Laminitis: 3.

The dataset is imbalanced, as you can see in the Figure 3.2, with most samples labeled as either colic or healthy, and relatively few for laminitis and colitis. It must be taken into account for the machine learning algorithms that tend to be biased toward majority classes without corrective measures.

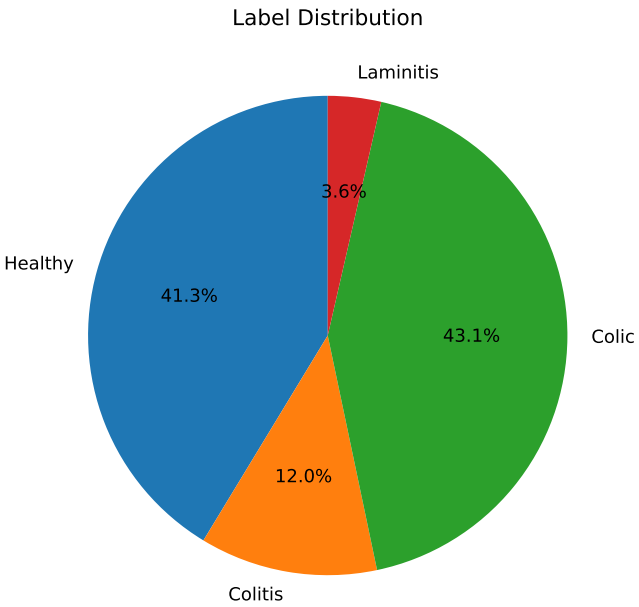


Figure 3.2: Proportion of each label in the dataset.

3.2 Methods

3.2.1 Preprocessing using `QIIME2`

The first way to preprocess the raw sequencing data is by using `QIIME2` (Quantitative Insights Into Microbial Ecology) to build a complete bioinformatics pipeline. `QIIME2` is a free, open-source data science platform developed for microbiome analysis. All samples are processed individually [69]. This way of processing the data aims to maximize precision in the analysis of the microbiota. The final objective is to obtain a taxonomy table, which provides an estimated proportion of each type of microorganism.

The whole process includes multiple steps, described below:

1. **Importing `FASTQ` files**
   The first step involves transforming the single end with quality sequencing files to the `QIIME2` standard format (`.qza`). This step also ensures that all files are in the correct format.

2. **Denoising and filtering**
   Once imported, the DADA2 algorithm is used to denoise the sequences [70]. This is an important step that helps removing the sequencing errors and separate the noise introduced during sequencing from the true biological signal. Sequencing errors include incorrect, deletion and insertion of bases as well as low-quality end of reads. Indeed, the beginning of the reads is often high quality and then the furthest on the read, the poorer the quality. In this case, all sequences were truncated to 250 base pairs to ensure consistent quality across samples. This value is often used across studies [71, 72]. After truncation, DADA2 applies a statistical model to identify and correct all errors. Then, identical reads are grouped to reduce the computations for the end of the pipeline. This step thus produces two files: a table listing the abundance of each representative sequence across the sample (called a "feature table") and a set of those representative sequences (error-corrected).

3. **Taxonomic classification**
   Once representative sequences were obtained through the denoising step, it was necessary to determine the organism from which each sequence originated. To perform taxonomic classification, a Naive Bayes classifier is used, it is a simple but effective algorithm that applies Bayes' theorem to estimate the most likely taxonomic assignment based on the observed sequence features. This classifier was trained on a reference database called SILVA (version 138), containing millions of known 16S rRNA gene sequences from bacteria and archaea [73, 74]. For each representative sequence, the classifier predicts the most probable taxonomic assignment by comparing with the sequences present in the SILVA database. The output is a table that maps a representative sequence to a taxonomic label up to the genus level (as can be seen in Figure 2.2), as the species-level is unreliable using SILVA. If the classifier is not able to confidently assign the feature to a known genus, it may stop at a higher level. For instance, as shown in Figure 2.2, if there is not enough discriminative information to reach the Genus level, the classifier might stop at the Family or even Order level, providing only a partial but still informative classification.

4. **Taxonomic collapse**
   Once the classification is done, the feature table is collapsed, meaning that the occurrences of all features belonging to the same assignment are counted. Finally, all tables are merged together to transform them into a single tab-separated value (.tsv) file on the QIIME2 website. The final output is a lightweight file of only 3.7 megabytes, especially small compared to the original 200 gigabytes of FASTQ data. In this table, each row corresponds to

a sample, and each column represents a detected bacterial taxon. The values indicate how many times each feature was identified in the corresponding sample.

3.2.2 Preprocessing using `Kraken2`

Another approach was used to preprocess the data, using `Kraken2`, a taxonomic sequence classifier based on exact k-mer matches [75]. For each raw read of DNA in the `FASTQ` files, `Kraken2` will break it into fixed length subsequences (k-mers). From those k-mers, they will store a subset called minimizers, which are a representative subset of the k-mers and then checks to which taxa they are linked to infer the most likely taxonomic assignment. The goal here is to process the fastq files without intermediate transformations. The choice was made to use three different databases to maximize the information and completeness of the taxonomic assignments. However, performing this type of analysis with `QIIME2` would have been prohibitively time-consuming and computationally intensive, as `QIIME2` workflows are not optimized for handling multiple large reference databases in parallel. `Kraken2` offered a much faster alternative, allowing us to classify all reads efficiently across all three databases without extensive preprocessing. An effort was also made to reduce the bias introduced by using a single database.

- **SILVA**, as presented earlier is very comprehensive and regularly updated but sometimes includes ambiguous entries due to its broad inclusion criteria [73].
- **Greengenes**, is more compact but is not updated since 2013. It is based on a tree taxonomy approach [76].
- **RDP (Ribosomal Database Project)** is highly curated and offers a robust taxonomy. It also works with fungal species. Its resolution can however be poor at the genus level [77].

For each sample and each reference database, `Kraken2` produces a standard report file summarizing the taxonomic classification results. Each line in the report corresponds to a taxon and includes the percentage of reads assigned to it (or its descendants), the exact number of reads, the taxonomic rank, and the scientific name, all organized hierarchically. The structure of a `Kraken2` report is illustrated in Figure 3.3, and includes the following columns:

- %: Proportion of total reads assigned to this taxon or its descendants.
- **Reads**: Total number of reads assigned to this taxon or its descendants.
- **Reads (taxon)**: Number of reads assigned directly to this taxon.
- **Rank**: Taxonomic rank (e.g., genus, species), as shown in Figure 2.2.
- **Taxon ID**: NCBI taxonomic identifier.
- **Taxon name**: Scientific name of the taxon.

Since each sample produced three separate reports (one per database), a custom transformation was developed to merge them into a single, unified file. During this process, the hierarchical structure, which can be difficult to manipulate programmatically, was also flattened, keeping only the taxon-specific lines for ease of analysis.

```
 8.38  11963   11963   U       0       unclassified
91.62  130712  3       R       1       root
87.11  124286  231     D       3         Bacteria
61.53  87784   86      P       34          Firmicutes
47.16  67289   2       C       226           Clostridia
47.16  67283   3188    O       527             Clostridiales
14.28  20381   6580    F       1017              Lachnospiraceae
 2.42  3459    2277    G       1797                Blautia
 0.63  898     898     S       2777                  Blautia producta
 0.20  284     284     S       2776                  Blautia obeum
 2.00  2849    2633    G       1800                Coprococcus
 0.14  201     201     S       2779                  Coprococcus eutactus
 0.01  15      15      S       2778                  Coprococcus catus
```

Figure 3.3: Example of a `Kraken2` report. Each line represents a taxon with its associated classification metrics.

3.2.3 Comparison of QIIME2 and Kraken2 Pipelines

To better understand the impact of preprocessing on downstream disease classification, two widely used taxonomic classification pipelines were compared: QIIME2 and Kraken2. Although both aim to generate a representation of the microbial composition in each sample, their underlying methodologies and outputs differ substantially, as summarized in Table 3.2.

QIIME2 relies on a multi-step statistical approach, including quality filtering, denoising (via DADA2), and taxonomic assignment based on curated databases. This results in a relatively smaller set of high-confidence features with exact sequence resolution. In contrast, Kraken2 uses exact k-mer matches and provides a rapid read-by-read classification, supporting the integration of multiple reference databases. While this yields broader coverage, it may introduce more noise due to the lack of error correction.

This comparison highlights the trade-off between precision and scalability, and raises the core research question of this thesis:

> *How does the performance of various machine learning algorithms applied to the equine microbiome vary depending on the preprocessing pipeline used (QIIME2 or Kraken2) for diagnosing gastrointestinal diseases?*

Table 3.2: Comparison between QIIME2 and Kraken2 pipelines for taxonomic analysis.

Criteria	**QIIME2**	**Kraken2**
Approach	Denoising and sequence inference with statistical models (DADA2)	Exact k-mer matching against reference databases
Speed	Slower, due to multi-step pipeline and denoising (Multiple days for the whole dataset)	Very fast, optimized for high-throughput data (A couple of hours for the whole dataset)
Output	Feature table with taxonomic assignment	Read-by-read taxonomic classification
Granularity	High-resolution, exact sequences	Read-level classification, no sequence clustering
Error handling	Actively denoises and corrects sequencing errors	No denoising, relies on raw reads
Database used	SILVA	SILVA, Greengenes and RDP
Features created	Structured features	Higher coverage, but noisier features
Amount of features	1371	3394
Files processed	1223	1227
Sparsity of matrix	89%	85%
Use case	Ideal for in-depth microbiome analysis with robust statistics	Ideal for rapid, large-scale taxonomic profiling

3.2.4 Visualization

To investigate the structure of the microbiome data and assess potential patterns between clinical conditions, dimensionality reduction techniques followed by clustering were applied .

First, Principal Component Analysis (PCA) is performed on the scaled feature matrix to obtain a linear projection of the data that preserves the directions of maximum variance. The t-distributed Stochastic Neighbor Embedding (t-SNE) is also applied, it is a nonlinear technique particularly suited for high-dimensional data, to better visualize local similarities between samples.

Both PCA and t-SNE were applied to the `Kraken2`-based feature matrix after scaling. Each sample was then visualized in a two-dimensional space, colored by clinical class.

In addition, the DBSCAN clustering algorithm is applied on the resulting 2D t-SNE projection to detect dense groupings of samples. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters based on local density, without requiring a predefined number of clusters. The parameters `eps=2.2` and `min_samples=15` were used to detect core clusters and outliers (noise points) in the t-SNE space.

These visualizations were designed to provide qualitative insight into the separability of classes and to explore whether natural clusters exist among the samples prior to classification.

3.2.5 Classical Machine Learning Models

For the SVM, MLP, Logistic Regression and Random Forest, further preprocessing steps have been applied. The goal is to improve the model performance and reduce the overfitting due to the high-dimensional feature space. The strategies were tested during the hyperparameter tuning.

Three types of feature selection or dimensionality reduction techniques were considered:

- **Principal component analysis:**
  PCA is used as a preprocessing step to reduce the number of features. By retaining only the principal components that capture the highest variance, noise is reduced and less informative features are eliminated. The drawback is that less interpretability of the results is achieved, which could be useful in this case. Another drawback is that important features with lower variance may be discarded. The configurations tested contained 10, 50, and 300 components (explaining 18, 41 and 87% of the variance), reducing drastically the feature space.

- **Variance Threshold:**
  This method consists in removing the features with low variance across all the samples, with the assumption that those features are uninformative, as they exhibit minimal variation. Once again, a too aggressive thresholding might reduce model performance if relevant rare taxa are filtered out. Thresholds of 0.001, 0.01 and 0.1 were considered (removing 6, 21 and 37% of taxa). The idea of using variance as a criterion assumes that features with higher variance may carry more informative signal, which is not always true.

- **Top presence selector:**
  This custom method selects a fixed proportion (e.g., 1%, 20%, or 50%) of the taxa that are most frequently present in the samples. The frequency of the presence was computed as the proportion of samples for which the abundance of a taxon is greater than zero. The goal is to only keep taxa often present. The matrix with all features is particularly sparse (see Table 3.2) with a lot of taxa being specific to few samples, inducing sometimes non-informative variability. This method ensures that models are trained on taxa with sufficient representation. The problem can obviously be that pathogens associated with specific diseases may be excluded by this approach if the selection ratio is too low.

- **No preprocessing:**
  A control group with no feature reduction is also kept, but standard preprocessing steps such as normalization was still used. This setting retains all available information but increases the risk of overfitting.

All those preprocessing options were evaluated using a grid search, a systematic method for hyperparameter tuning that tests all possible combinations of specified parameters to identify the configuration yielding the best performance. This ensures that the most appropriate strategy is selected for each model. Grid search is widely used in machine learning workflows for model optimization [42].

Each of the four classical machine learning algorithms was tested with different hyperparameters in a grid search, along with the different preprocessing steps already presented. Below are described the key parameters tested for each model and their potential influence on classification performance.

The **SVM** was tested with multiple parameters, beginning with different kernels (`linear`, Radial Basis Function (`rbf`), and `polynomial`). They control how the model maps the input features into higher-dimensional spaces in order to better separate the classes. For `rbf` and `polynomial` kernels, the `gamma` parameter controls the influence of each individual sample; a small `gamma` leads to smoother decision boundaries. The two options are `auto` and `scale`, with `gamma='auto'` depending on the number of features, and `'scale'` also incorporating the variance. The `degree` parameter controls the complexity of the polynomial kernel; in this case, degrees 2 and 3 were tested. Finally, the regularization parameter `C` defines the trade-off between margin and training error. Values between 0.01 and 10 were tested [78, 79].

The **Random Forest** method combines multiple decision trees. The key parameters include the number of trees (`n_estimators`), the maximum depth of each tree (`max_depth`), the minimum number of samples required to split a node (`min_samples_split`), and the maximum number of features used at each split (`max_features`). Reducing tree depth and increasing `min_samples_split` can help reduce overfitting. A smaller `max_features` value increases diversity among trees, potentially reducing overfitting, but may also degrade individual tree performance [79, 80].

For **Logistic Regression**, the default `l2` penalty is used, which is the sum of squared coefficients multiplied by a regularization parameter. Two solvers were tested: `lbfgs` and `saga`, both compatible with multiclass classification using `l2` regularization. The main parameter to tune is the regularization strength `C` (inverse of regularization), where smaller values impose stronger regularization—reducing variance but possibly increasing bias [79, 81].

The **MLP** being a feedforward neural network, several architectural and training hyperparameters were tested. These included different hidden layer sizes, activation functions (`relu`, `tanh`), regularization strength (`alpha`), and learning rate (`learning_rate_init`). Configurations with one or two hidden layers were tested, with either 50 or 100 neurons in the first hidden layer and 50 in the second. More neurons allow the model to learn more complex patterns but increase the risk of overfitting. The learning rate controls how quickly the model adapts to errors—too large, and training may diverge; too small, and it may never converge. Regularization penalizes large weights to improve generalization and prevent overfitting [79, 82].

Several design choices were made to mitigate the impact of class imbalance. All classifiers (except MLP, which does not support it) were configured with `class_weight='balanced'` to adjust the loss function so that each class contributes proportionally. Additionally, `StratifiedKFold` is used to preserve class distributions across cross-validation folds. Finally, model selection was based on `f1-weighted` scoring. This metric is preferred over accuracy in imbalanced settings, as accuracy can be dominated by majority classes. It also outperforms plain F1-score by avoiding excessive influence from minority classes.

Here is the formula of this scoring method:

$$\text{F1}_{\text{weighted}} = \sum_{i=1}^{K} \frac{n_i}{N} \cdot \text{F1}_i = \sum_{i=1}^{K} \frac{n_i}{N} \cdot \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

- K is the number of classes,
- n_i is the number of true samples in class i,
- $N = \sum_{i=1}^{K} n_i$ is the total number of samples,
- F1_i is the F1-score of class i,
- $\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$ is the precision for class i,
- $\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$ is the recall for class i,
- TP_i, FP_i, FN_i are the true positives, false positives, and false negatives for class i.

3.2.6 Convolutional Neural Network

As explained in the state of the art a one-dimensional CNN can be well-suited for a list of taxa, being able to capture local patterns in a list of abundances of microbial taxa. As in the last part, multiple parameters were tested using a grid search.

Before training, several preprocessing strategies were evaluated as part of the grid search, aiming to reduce noise and improve learning efficiency:

- **Full feature set:** The first configuration includes all available microbial features without filtering.
- **Top-abundance filtering:** The second setup retains only the top 10% most abundant taxa across all samples, based on total read counts. This approach reduces noise by discarding rare taxa that are unlikely to contribute meaningfully to classification.

In addition to feature selection, reordering the columns of the input matrix was explored too to enhance the effectiveness of convolutional filters. Specifically, the pairwise Pearson correlation between taxa based on their abundance profiles across samples was computed. Pearson correlation quantifies the linear relationship between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). By clustering and reordering features with high average Pearson correlation, the goal was to position similar taxa closer together in the input sequence. This spatial proximity can help convolutional kernels better detect local patterns of co-abundance that may be relevant for classification.

After this preprocessing step (with or without reordering), the input data was passed through a convolutional neural network composed of two convolutional layers. It was experimented with several kernel size combinations, selected from $(k_1, k_2) \in (7, 3), (9, 5), (5, 3), (15, 4)$:

- **Conv1:** from 1 to 32 channels, with kernel size k_1,
- **Conv2:** from 32 to 64 channels, with kernel size k_2.

Kernel size variations allow learning shorter or longer range dependencies. Each convolutional layer is followed by a batch normalization that will normalize the output of the layer, acting as a regularizer. A ReLu activation function is applied to introduce non-linearity. Then, each block ends with a maximum pooling to reduce the resolution by half to focus on the most important features and reduce overfitting.

Those two convolutional blocks are followed by fully connected layers, it was tested with one and two layers, from those configurations: ((32),(64), (32,64)). These layers combine the new high-level features in a smaller discriminative representation right before classification. A dropout of 0.5 is applied, randomly deactivating half of the neurons to avoid overfitting. The final layer outputs the last representation into the class labels. A focal loss is used to reduce the impact on class imbalance. The whole network is trained using the Adam optimizer with a learning rate of 10^{-3} and again with 5-fold cross-validation. An early stopping is implemented to reduce overfitting and computational time. The best model, based on the weighted f1-score, across the folds is saved and evaluated on the test set.

3.2.7 Graph Neural Network

To classify the obtained data, a Graph Neural Network is also implemented. Each sample is represented as a node in a graph and the edges encode the similarity between two nodes. The final goal will be to classify each node into one of the four health conditions. Just as in previous classifiers, a grid search is performed to find the best parameters.

The model implemented consists of two GraphSAGE (Graph Sample and Aggregate) convolutional layers followed by a fully connected linear classifier [83]. GraphSAGE is a graph neural network architecture that learns node embeddings by information from each node's local neighborhood, enabling a better learning on unseen graph data, which is integrated in pytorch.

Again, two different ways to preprocess the data are tried. One by keeping all the data and the other one by only keeping a certain percentage of the most abundant taxa (10, 20 and 50). A StandardScaler and a StratifiedKFold are also used with 5 folds as before.

Then, two different strategies are used to build the graph:

- Bray-Curtis: An edge is drawn between two nodes if their Bray-Curtis dissimilarity is below a given threshold. Thresholds of 0.1, 0.3 and 0.5 are tried. A Bray-Curtis score of zero means that two samples are exactly the same.

- K-nearest neighbours: Each sample is connected to its k (5,10,15) nearest neighbors, based on Euclidean distance.

A first SAGEConv layer maps the obtained input data to a latent space with dimension chosen between 32 and 256 in the grid search. A dropout layer with a rate of 0.5 is applied after this first GraphSAGE layer to prevent overfitting.

A second SAGEConv layer maps into a latent space with dimension of half of the previous one. Finally, a linear layer maps this node embedding into the final classes. As for the CNN, the Focal Loss is used to address the class imbalance and the model is optimized using the Adam optimizer with a learning rate of 10^{-3}. Early stopping based on validation accuracy was applied to avoid overfitting.

4 Results

4.1 Visualization

To assess the underlying structure of the microbiome profiles and investigate potential patterns between clinical conditions, unsupervised dimensionality reduction techniques are first applied followed by clustering. Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are both applied to the scaled feature matrix produced using `Kraken2`. One scatter plot was produced for each projection, with one color per condition and one point represents a sample.

The first algorithm used for visualization is PCA, Figure 4.1 shows the projection of the samples onto the first two principal components. This distribution reveals a dense cluster near the origin, with a visible separation of some samples along PC1, particularly labeled as Healthy and others along PC2 labeled mostly as Colic and as Colitis.

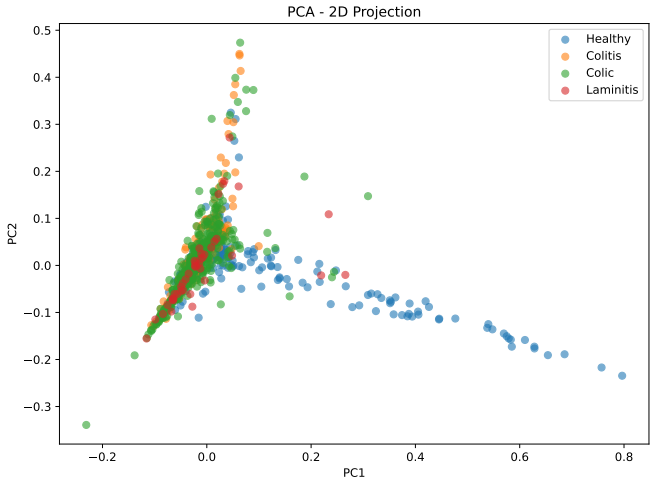


Figure 4.1: 2D PCA projection of `Kraken2`-based features. Each point represents a sample colored by its clinical class (Healthy, Colic, Colitis, Laminitis).

Figure 4.2 presents the two-dimensional t-SNE embedding. Compared to PCA, this representation shows a more dispersed structure, with several small clusters. Some groups of Healthy and Colic samples appear spatially separated from the rest, while other classes are more intermingled. No clear separation of the data can be seen.

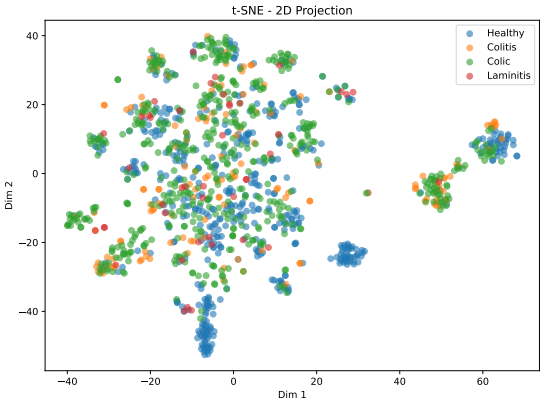


Figure 4.2: 2D t-SNE projection of `Kraken2`-based features, colored by clinical class.

Figure 4.3 shows the t-SNE 2D projection of the microbiome samples, colored according to clusters detected by DBSCAN (with eps=2.2 and min_samples=15). The algorithm identifies 12 distinct clusters among the samples (Cluster 0 to Cluster 11). Points that do not belong to any cluster are shown in gray and labeled as "Noise". These likely correspond to more isolated or ambiguous samples that do not form dense enough neighborhoods.

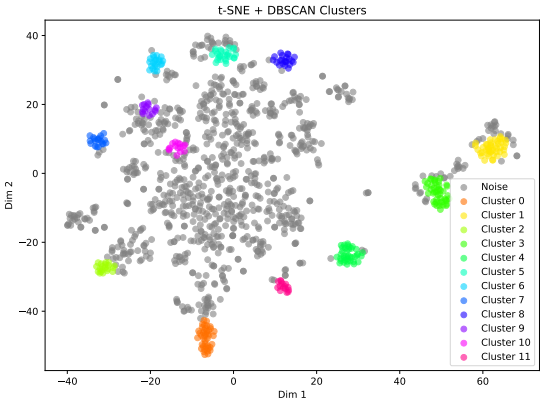


Figure 4.3: t-SNE projection with DBSCAN clustering (`Kraken2` data). Twelve clusters were identified, with gray points labeled as noise.

4.2 Predictive machine learning algorithms

4.2.1 Classical Machine Learning Models

In this first part, the ability of classical machine learning models is assessed to predict the clinical condition of a horse based on its microbiota profile. The task consists of a multiclass classification among four conditions: Healthy, Colic, Colitis, and Laminitis. The goal is to evaluate how different algorithms and preprocessing pipelines (`QIIME2` vs `Kraken2`) influence predictive performance. The classification of the data was first tested on the "classical" machine learning algorithms, including SVM, Random Forest, MLP and Logistic Regression. All those algorithms were tested using both the `QIIME2` and the `Kraken2` preprocessing, with accuracy values represented in a bar plot in Figure 4.4 for the best hyperparameters found using the Grid Search. It can be seen that for all algorithms except Logistic Regression the `QIIME2` pipeline offered a higher accuracy. Table 4.1 presents a detailed summary of the results, including the best preprocessing strategy applied for each model, the resulting accuracy, and the corresponding F1-score. For each model, the best-performing configuration is reported. The best model using `QIIME2` is the SVM, regarding both accuracy and F1-score. Using `Kraken2` the best model is the Random Forest algorithm. For six of the eight combinations, the best preprocessing is to keep the most abundant data.

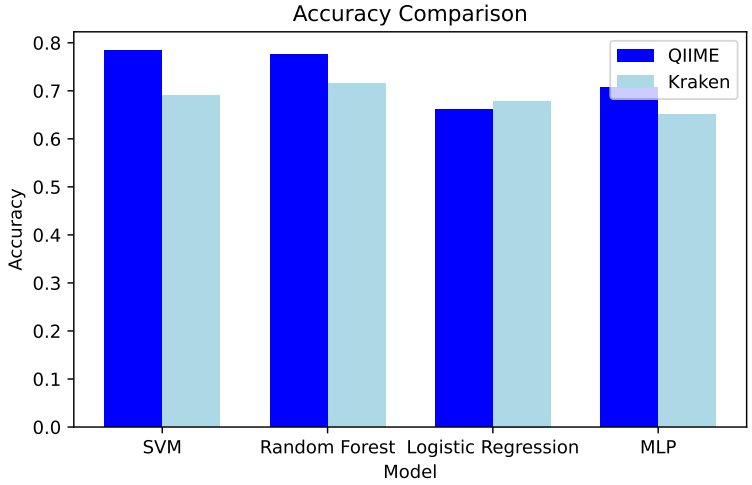


Figure 4.4: Comparison of classification accuracies for `QIIME2` and `Kraken2` pipelines across four models (SVM, Random Forest, MLP and Logistic Regression). For each model, the accuracy shown corresponds to the best configuration obtained from grid search.

Table 4.1: Best model configurations for each pipeline, with associated preprocessing strategy, accuracy, and F1-score.

Model	**QIIME**			**Kraken2**		
	Preproc	Acc	F1	Preproc	Acc	F1
SVM	TopPres. 0.2	0.784	0.776	TopPres. 0.2	0.691	0.682
RF	TopPres. 0.5	0.776	0.758	VarThresh 0.01	0.715	0.691
LogReg	VarThresh 0.01	0.661	0.667	TopPres. 0.5	0.679	0.669
MLP	TopPres. 0.1	0.706	0.702	TopPres. 0.5	0.650	0.646

To better evaluate the classification performances, the confusion matrices are computed for each model. Each matrix is a 4x4 grid with the rows representing the true labels and the columns, the predicted labels. The diagonal elements indicate the correct classifications, other elements being misclassifications.

The Support Vector Machine (SVM) model shows strong performance on the Healthy class (label 0), correctly classifying 83 out of 99 samples. However, there is moderate confusion with the Colic class, and some misclassifications occur for Laminitis and Colitis. This is illustrated in the confusion matrix shown in Figure 4.5.

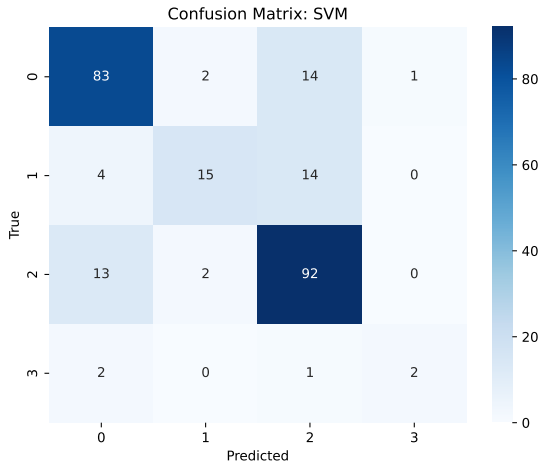


Figure 4.5: Confusion matrix of the SVM classifier trained on `QIIME2` data (multi-class setting). Rows correspond to true clinical classes and columns to predicted classes. Classes: Healthy, Colic, Colitis, Laminitis.

The Random Forest classifier achieves the most accurate predictions for both Healthy and Colic classes. It misclassifies fewer samples overall and shows the clearest diagonal pattern in its confusion matrix, reflecting stronger generalization. However, Laminitis remains a challenge. These results are visualized in Figure 4.6.

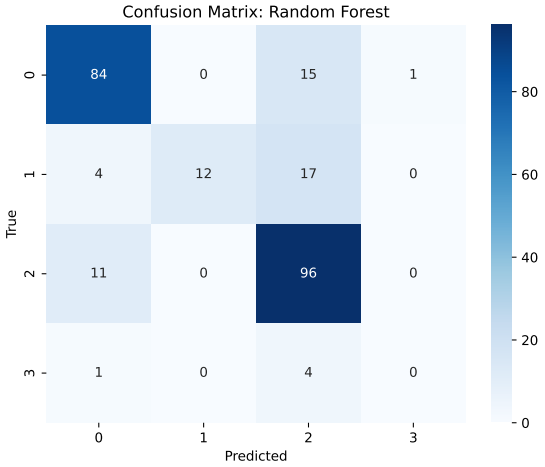


Figure 4.6: Confusion matrix of the Random Forest classifier trained on QIIME2 data (multiclass setting). Rows correspond to true clinical classes and columns to predicted classes. Classes: Healthy, Colic, Colitis, Laminitis.

The Logistic Regression model exhibits more dispersion in its predictions, with lower diagonal counts, particularly for the Colic class where 25 samples are misclassified. This model appears less capable of separating overlapping classes. Figure 4.7 displays the corresponding confusion matrix.

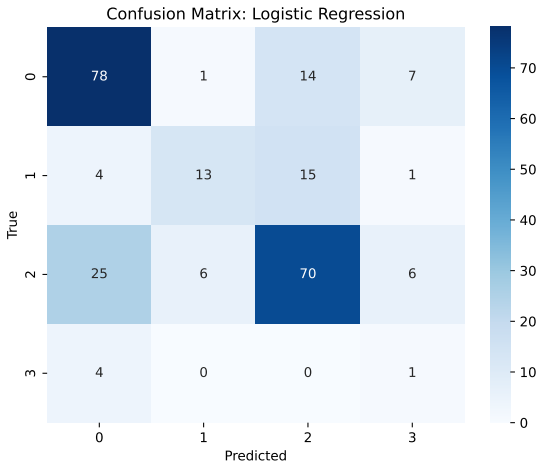


Figure 4.7: Confusion matrix of the Logistic Regression classifier using QIIME2 features. Rows correspond to true clinical classes and columns to predicted classes. Classes: Healthy, Colic, Colitis, Laminitis.

Across all models, the most frequent misclassifications are between Colic and Colitis. Conversely, Healthy and Colitis are rarely confused. The Laminitis class continues

to be the most poorly predicted. These trends are further reflected in the confusion matrix of the MLP classifier (Figure 4.8).

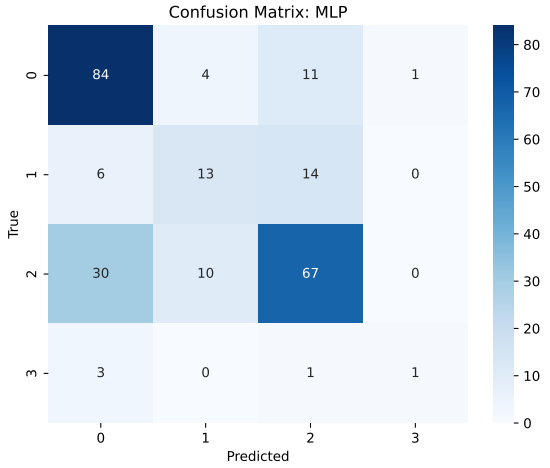


Figure 4.8: Confusion matrix of the MLP classifier using QIIME2 features (multiclass classification). Rows correspond to true clinical classes and columns to predicted classes. Classes: Healthy, Colic, Colitis, Laminitis.

For further analyses, a feature importance plot was computed on the QIIME2 dataset to identify which taxa contributed most to the classification performance. The Random Forest provides a natural mechanism to assess the importance of the features using the impurity criterion. Figure 4.9 shows the ten most important taxa ranked by importance score but without giving any biological interpretation.

The most important taxon is the genus Escherichia shigella with an importance score of 0.0182 and the second one being a genus from the Rikenellaceae RC9 gut group, with a score slightly above 0.01. The remaining features exhibit similar importance values, ranging approximately between 0.008 and 0.005. All features are present at the genus level except for two of them: the overall bacteria abundance and the abundance in the phylum Firmicutes. This is explained by the way QIIME2 performs taxonomic classification. The classifier used assigns each representative sequence to the most specific taxonomic level for which it has enough confidence. In some cases, the classifier cannot confidently assign a sequence to a genus due to insufficient resolution in the 16S rRNA sequence. As a result, the assignment is truncated at a higher level in the taxonomy. These higher-level assignments, still capture meaningful variation across samples and can appear as important features in the classification model.

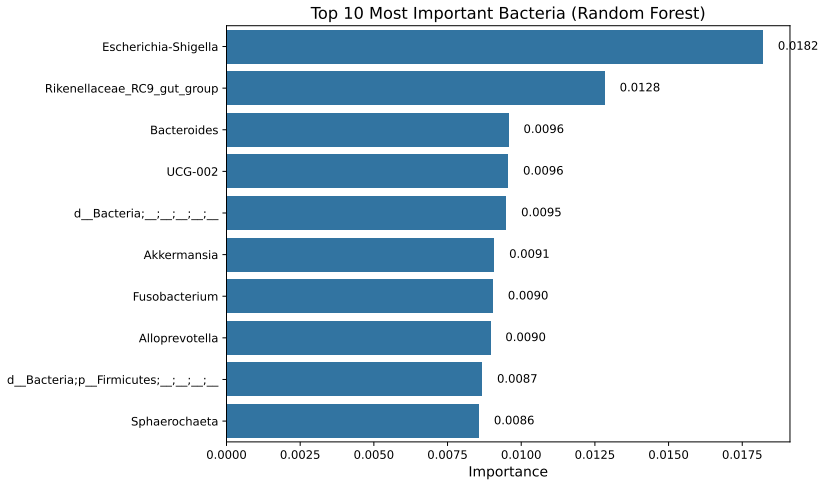


Figure 4.9: Top 10 most important taxa based on Random Forest feature importance scores (QIIME2 data). The features are assigned at the genus level, except for two (overall bacteria and phylum Firmicutes), which appear at higher levels due to QIIME2's confidence-based classification mechanism.

From a clinical and epidemiological standpoint, the ability to simply detect whether a horse is diseased could serve as an efficient early screening tool. This setup also reflects the approach used in several prior microbiome studies, which aim to detect a general dysbiosis rather than a specific disease label. Such a formulation may be more robust for early detection or screening use cases.

The same algorithms were used to discriminate between sick and healthy horses. Among the four classifiers, the SVM and Random Forest models performed best, each reaching an overall accuracy of 0.85.

The SVM classifier showed balanced performance, correctly identifying 83 Healthy horses and 125 Diseased ones. It made 17 false positives and 20 false negatives, indicating a relatively low misclassification rate in both directions. These results are detailed in the confusion matrix shown in Figure 4.10.

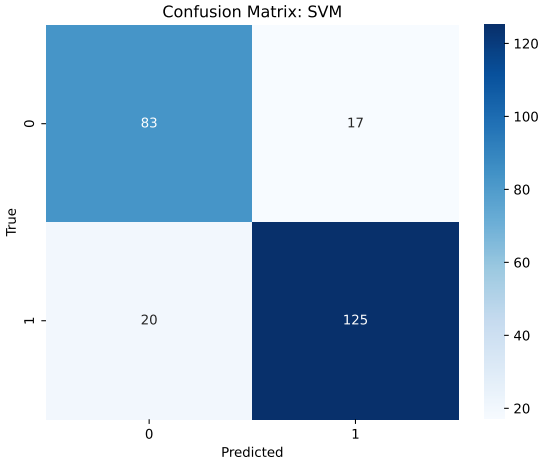


Figure 4.10: Confusion matrix of the SVM classifier for binary classification (Healthy vs Diseased) using QIIME2 data.

In contrast, the Random Forest model correctly classified 133 Diseased horses but only 75 Healthy ones. It achieved the same global accuracy (0.85) as the SVM but leaned more heavily toward detecting positive (Diseased) cases, with 25 false positives and 12 false negatives. This bias is visible in the confusion matrix displayed in Figure 4.11.

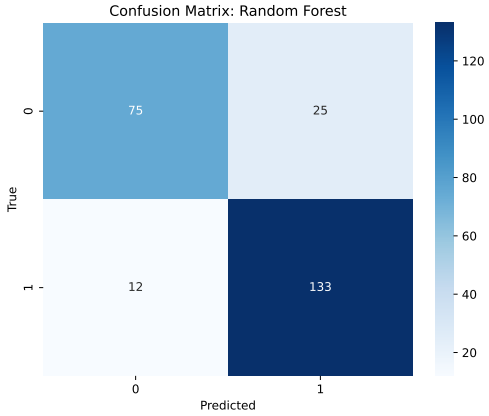


Figure 4.11: Confusion matrix of the Random Forest classifier for binary classification (Healthy vs Diseased) using QIIME2 data.

The Logistic Regression model showed reduced accuracy, especially on the Diseased class. It correctly identified 111 Diseased horses but misclassified 34 of them. Despite slightly better detection of Healthy samples (85 correct), the increased number of false negatives highlights weaker sensitivity to diseased individuals. These outcomes are shown in Figure 4.12.

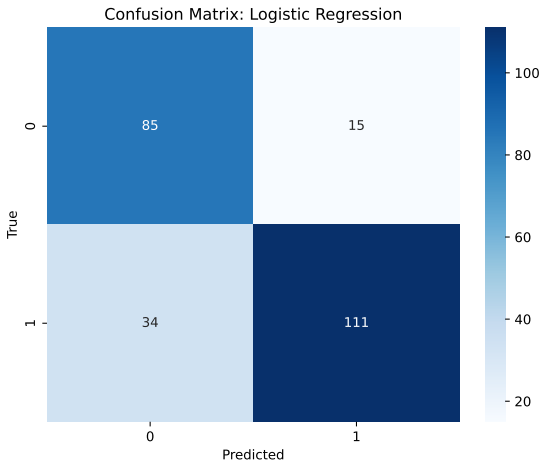


Figure 4.12: Confusion matrix of the Logistic Regression classifier for binary classification (Healthy vs Diseased) using `QIIME2` data.

Lastly, the MLP classifier performed similarly to Logistic Regression, correctly classifying 118 Diseased and 75 Healthy horses. It struggled particularly with false positives (25) and had a moderate number of false negatives (27), resulting in the lowest precision among the four models. These results are shown in Figure 4.13.

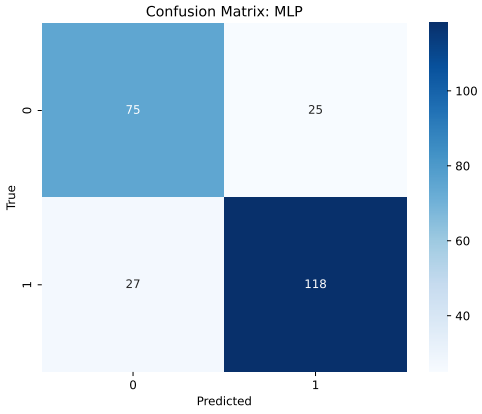


Figure 4.13: Confusion matrix of the MLP classifier for binary classification (Healthy vs Diseased) using `QIIME2` data.

To evaluate the discriminative power of each classifier, the Receiver Operating Characteristic (ROC) curves are computed and plotted using the test set. The Area Under the Curve (AUC) was used as a summary metric of performance. As shown in Figure 4.14, the Random Forest achieved the highest AUC of 0.95, followed by the SVM with an AUC of 0.91. The MLP and Logistic Regression models obtained slightly lower AUC values of 0.87 and 0.85, respectively.

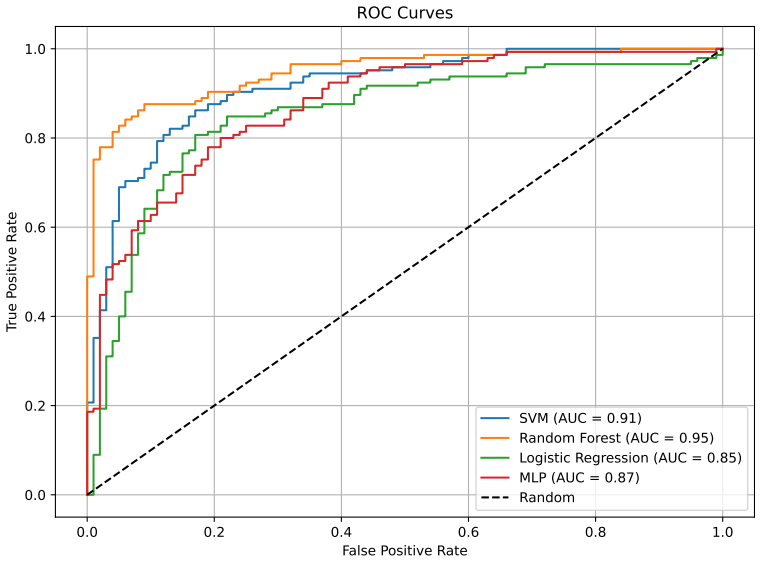


Figure 4.14: Receiver Operating Characteristic (ROC) curves for the binary classification task (Healthy vs Diseased) using four classifiers trained on `QIIME2` features. The Area Under the Curve (AUC) is computed for each model to quantify overall classification performance.

4.2.2 Convolutional Neural Network

After evaluating classical models, it was tested whether a 1D convolutional neural network could improve classification by capturing local patterns in microbial abundance profiles. As QIIME-based preprocessing yielded better performance in previous experiments, only `QIIME2` data was used for this analysis.

The one-dimensional Convolutional Neural Network (CNN1D) was trained on the QIIME-preprocessed dataset. The best-performing configuration on the validation tests used a feature selection step that retained the top 10% most abundant bacterial taxa. The convolutional architecture included two kernels of size 15 and 4. This was followed by a single fully connected layer with 64 units. The order of the bacteria list chosen is the original order found in the `QIIME2` report. This configuration achieved a validation accuracy of 0.699.

The model was then evaluated on the independent test set, and the resulting confusion matrix is shown in Figure 4.15. The overall test accuracy was 0.600, a result significantly lower than all previous models.

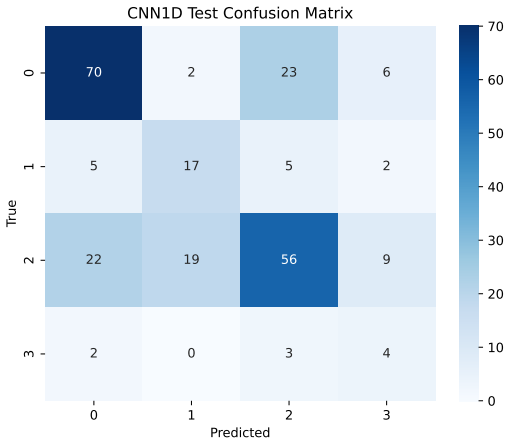


Figure 4.15: Confusion matrix of the 1D CNN model trained on QIIME2 data (test set). Architecture: two convolutional layers (kernel sizes 15 and 4), one dense layer (64 units), top 10% taxa retained.

4.2.3 Graph Neural Network

Finally, the last model is a graph neural network trained on the QIIME2 dataset for both multiclass and binary classification tasks as shown in Figure 4.16. It shows a steady decrease in training loss over epochs, converging to a near-zero value after 300 epochs. The training accuracy increased rapidly during the first 100 epochs and stabilized above 0.95 after approximately 200 epochs. The best configuration used a Bray-Curtis distance below a threshold of 0.5, a hidden size of 64, and a learning rate of 0.001, with feature selection limited to the top 20% most present taxa. Early stopping was triggered at epoch 187. The final test accuracy for this configuration was 0.718, as shown in Figure 4.17a.

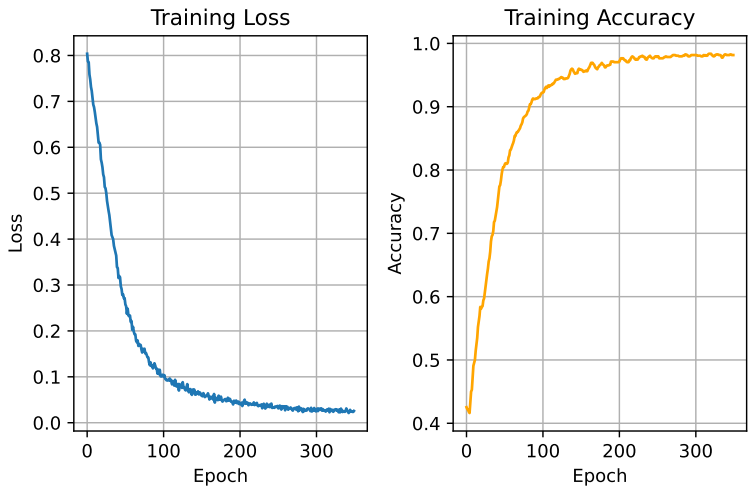
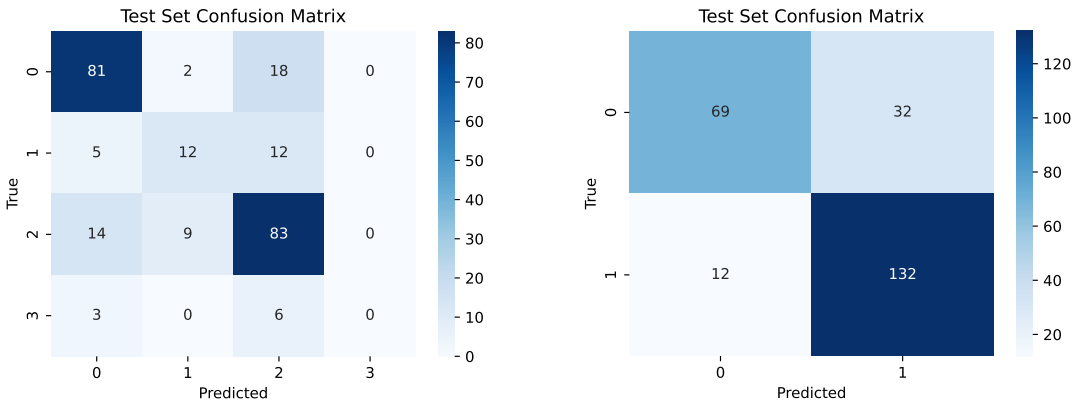


Figure 4.16: Training and validation accuracy of the best Graph Neural Network model on QIIME2 data. Model: GraphSAGE with 64 hidden units, learning rate 0.001, early stopping at epoch 187.

The best binary classification configuration used the same graph and preprocessing settings but a larger hidden layer of 128 units. The model reached a final test accuracy of 0.820 with an early stopping triggered at epoch 165.



(a) Multiclass GNN (4 classes), 64 hidden units, 20% taxa, Bray-Curtis 0.5.

(b) Binary GNN (Healthy vs Diseased), 128 hidden units, same graph.

Figure 4.17: Confusion matrices for the best GNN configurations trained on QIIME2 data. GNN outperforms CNN in both multiclass and binary settings.

5 Discussions

5.1 Visualization

The PCA and t-SNE plots were obtained using the `kraken2` preprocessing. Visualizations based on `QIIME2` preprocessing yielded less interpretable patterns and are therefore included in the Appendix (Figure 7.2 and Figure 7.1).

The PCA and t-SNE visualizations provide significant insights on why it may be difficult to distinguish the four clinical conditions (Healthy, Colitis, Colic, Laminitis). Most of the samples being close to the center in the PCA in Figure 4.1 indicates that the variance across all samples exhibit similar patterns. There are only subtle separations observable: some Healthy samples tend to spread along PC1, while some other Colic and Colitis samples diverge along PC2. An overlap is seen in other microbiome studies: PCA plots often show a mixing of sample groups, reflecting the high-dimensional nature of microbial communities [84].

The t-SNE plot (Figure 4.2) does not separate the classes into distinct clusters either. Instead, it produces a more dispersed cloud of points with small groupings; partial grouping of some Healthy and Colic samples can be observed, but overall, the four classes are mixed. To further investigate whether hidden structure exists in the data, DBSCAN clustering is applied on the t-SNE coordinates (Figure 4.3). While t-SNE alone did not reveal well-separated class-specific regions, DBSCAN identified several dense subgroups of samples that form distinct clusters, along with a large proportion of samples labeled as noise. It is important to note that some clusters form depending on the origin of the data. For example, the Cluster 0 (containing 41 samples) contains data from the Bioproject PRJDB17872 (Table 3.1) exclusively, while Clusters 1 through 4 are also each composed of samples from BioProject PRJNA580257. These groupings may not be driven by pathology but possibly by technical or environmental consistency, possibly differences in DNA extraction methods, sequencing platform, or even the geographic and dietary background of the horses involved in the original studies. These results might also indicate that classifiers might learn to recognize study-specific artifacts instead of disease-specific patterns.

PCA captures linear directions of variance in the data, which means it is well-suited for detecting broad patterns but may miss smaller non-linear structures. On the other hand, t-SNE is designed to preserve local relationships between points and can reveal complex structures that PCA cannot. This explains why t-SNE often appears more clustered even if global class separability is still poor. Those two plots suggest that the data does not form distinct cluster in two dimensions, corresponding to each disease. Both PCA and t-SNE reduce thousands of microbial features to only two dimensions, which causes unavoidable information loss and can mask separations that might exist in higher dimensions.

Both these methods are unsupervised and therefore do not incorporate class labels, which limits their ability to capture the discriminative patterns needed for classification. The implications for the machine learning classification are direct: overlapping clusters and unclear boundaries between classes mean that simple decision rules will struggle. This encourages the use of more complex models such as the GNNs or CNNs being able to learn more complex patterns across the entire feature space. However, linear classifiers like Logistic Regression are unlikely to perform well in such poorly separated spaces.

5.2 Predictive machine learning algorithms

Several machine learning algorithms were trained to classify equine diseases based on gut microbiota profiles, including Logistic Regression, Support Vector Machines (SVM), Random Forest, Multi-Layer Perceptron (MLP), a Convolutional Neural Network (CNN), and a Graph Neural Network (GNN).

Overall, SVM and Random Forest emerged as the best classifiers in this study. On data preprocessed with `QIIME2`, SVM achieved the best performance (78.4% accuracy), slightly outperforming Random Forest (77.6% accuracy). On the other hand, on data preprocessed using `Kraken2` (which was noisier) Random Forest led with 71.5% accuracy, outperforming SVM (69.1%). The MLP achieved intermediate results (around 70% accuracy), better than Logistic Regression but worse than the leading classifiers (SVM and Random Forest). The Logistic Regression was the least effective model, reflecting the complex patterns lying in the data.

Those observations are consistent with previous research that highlight the performance of tree-based methods in microbiology compared to linear or shallow models. For example, one study on human disease classification using microbiota data reported that Random Forest achieved 94% accuracy, compared to 85% for SVM and 82% for Logistic Regression, showing the superiority of ensemble models in this context [85].

A closer look at the classification metrics highlights the ability of the models to maintain a good balance between precision and recall, despite the large class imbalance in the dataset. For instance, although some classes, such as laminitis, were largely underrepresented, the SVM and Random Forest models achieved a well-balanced trade-off between detecting rare diseases and maintaining precision. The SVM achieved the highest accuracy (78.4%) and F1-score (77.6%), along with identical precision (78.4%) and recall (78.4%), showing a balanced performance across all classes. Similarly, Random Forest maintained a high precision of 79.0% while still achieving a recall of 77.6%, showing that the model was sensitive to minority classes. The MLP and the Logistic Regression also performed consistently. These results confirm that the evaluation based on F1-score, which integrates both recall and precision, was essential for accurately assessing model performance in this context.

Both deep learning models were only trained using `QIIME2` data, which consistently produced better results in classical models. `Kraken2` was excluded from these experiments due to its higher noise and lower classification performance in earlier tests, which would likely hinder the convergence and generalization of complex architectures like CNNs and GNNs.

The use of deep learning models did not result in significant performance improvement. The 1D Convolutional Neural Network (CNN) performed poorly, with a multi-class test accuracy reaching only 60%. The CNN clearly failed to learn useful representations from the microbiota data, when compared to other models. Additionally, despite the use of cross-validation and dropout, the CNN developed showed overfitting with almost 10% difference between the validation and test accuracy. It can be concluded that the CNN did not generalize and was unsuited for this classification task. In contrast, the Graph Neural Network (GNN) delivered more promising results. In its best configuration (graphs built from `QIIME2` data with a Bray-Curtis threshold of 0.5 and retaining the top 20% most abundant taxa), the GNN reached 71.8% accuracy on the 4-class classification task and 82% on the binary classification. These scores are slightly below those of the SVM, though they indicate that the GNN was able to leverage some structural similarity between samples.

To understand better the behavior of the best-performing GNN, the learning curves were plotted across 300 epochs (Figure 4.16). The training loss steadily decreases while the accuracy improves, indicating proper convergence. However, the training accuracy reaches a plateau around 98%, which is inconsistent with the 71.8% test accuracy reported earlier. This large gap suggests that the model is overfitting the training data. While it learns to predict well on the training set, its performance does not generalize as effectively to unseen data. Despite exploring strategies such as early stopping and dropout tuning to improve the generalization of the model, it still exhibits important overfitting, indicating that additional regularization techniques or even a reduction in model complexity might be necessary to achieve even better test performance. A limitation of the GNN approach is that the graph is constructed using all samples at once, making it difficult to add a new individual sample for diagnostic purposes without rebuilding the entire graph structure. This hinders its applicability in real-time or clinical settings where fast individual predictions are needed.

5.3 Comparison to other studies

To contextualize model performance of this study, two recent approaches applied to human microbiome datasets were used for comparison: TaxoNN [6] and the graph-based model of Irwin et al. [7] in the Table 5.1. Those two studies aim for a binary disease prediction, but they differ in models and data used. The model presented in this work, based on 16S rRNA sequencing data from 1,223 equine samples processed via `QIIME2`, achieved an AUC of 0.95 using a Random Forest classifier (Figure 4.14), outperforming both TaxoNN (AUC = 0.92 for cirrhosis; 0.75 for type 2 diabetes) and the graph-based method (AUC = 0.93 for IBD prediction). TaxoNN utilized stratified CNNs trained by phylum to incorporate a relevant taxonomic structure. Irwin et al. employed graph embeddings over both metagenomic and metatranscriptomic features, followed by a classification using a Support Vector Machine. Despite the higher dimensionality of the human datasets, our model's performance highlights the importance of good preprocessing. Moreover, the previously obtained results suggest that simpler models such as Random Forests can be highly effective when paired with rigorous pipeline design, even in comparison to more complex deep learning frameworks. Furthermore, in contrast to CNNs or graph embeddings, the Random Forests offer interpretable outputs, enabling the identification of taxa most predictive of disease. However, since these comparative studies were conducted on human datasets, a direct conclusion remains tentative. Applying our preprocessing pipeline and modeling choices to their human data would be necessary to confirm whether the observed performance gain holds in that context.

Table 5.1: Comparison of the best model in this study with two other microbiome based models.

Aspect	**Thesis (Equine)**	**TaxoNN** [6]	**GNN** [7]
Goal	Binary classification (Healthy vs Disease)	Binary disease classification (Cirrhosis / T2D)	IBD prediction (human)
Sample Size	1223	Cirrhosis: 232; T2D: 344	1594
Data Type	16S rRNA (QIIME2)	16S rRNA (OTU table)	Metagenomics + Metatranscriptomics
Model Type	Random Forest	Ensemble CNNs by phylum	Graph embedding (Node2Vec, LPE) + SVM
Area Under Curve	0.95	0.92 (Cirrhosis), 0.75 (T2D)	0.93

5.4 Biological Relevance of the prediction

Figure 4.9 shows how the Random Forest algorithm allows to identify which taxa are most relevant for the classification. To assess the biological relevance of these taxa, the best predictors were cross-referenced with the literature on the equine gut microbiota, particularly the comprehensive review by Chaucheyras-Durand et al. [4]. This reveals clear links between microbial community and health status.

For example, Escherichia-Shigella is the top-ranked taxa and is known to proliferate in horses with colitis or diarrhea. It includes Escherichia Coli, a well-known bacteria, its overgrowth of Escherichia coli is a sign of dysbiosis, as they exploit inflamed environments to invade epithelial cells and disrupt the intestinal homeostasis, which contributes to conditions like Inflammatory Bowel Disease in the human being and potentially colitis in horses as well [86]. Fusobacterium, another key taxon, was frequently elevated in colitis and linked to gut inflammation. In the opposite way, some taxa such as Akkermansia and members of the Ruminococcaceae and Lachnospiraceae families were more abundant in healthy horses and are associated with mucosal health. These findings confirm that the most important taxa from machine learning not only capture statistical patterns, but reflect known biological signatures of equine gastrointestinal health as well.

Among the top predictors identified by the Random Forest model, several entries are labeled as unclassified or simply as "Bacteria". These represent sequences that were statistically informative for classification, but that could not be taxonomically resolved beyond the domain or phylum level by QIIME2's conservative annotation pipeline. This may reflect technical differences between datasets, as highlighted in Section 5.1: the clusters often aligned with specific BioProjects rather than diseases. These unclassified features may capture study-specific artifacts introduced by different experimental protocols. Additionally, this can suggest that useful diagnostic signals can be in parts of the microbiome that are still taxonomically unresolved.

5.5 Influence of Data Preprocessing (QIIME2 vs Kraken2)

An important aspect of this study was the comparison between two microbiome sequence preprocessing pipelines: QIIME2 and Kraken2. These approaches differ in how they filter noise and assign taxonomic labels, which has implications for the resulting datasets used in classification tasks and the corresponding results. QIIME2 produces more reliable feature tables by aggressively denoising reads and grouping the corresponding exact sequences, but limits taxonomic resolution. Kraken2, in contrast, offers higher taxonomic granularity, often down to species level, but lacks a denoising step, making it more prone to including erroneous or low-abundance taxa. As a result, QIIME2 tends to enhance classifier robustness by reducing noise, while Kraken2 can introduce variability that may degrade classification performance. In addition to taxonomic and methodological differences, another key motivation for comparing QIIME2 and Kraken2 lies in their computational processing times, which can influence the choice of pipeline in large-scale or time-constrained studies.

These pipeline divergences had a direct impact on classification performance. The matrix produced when the dataset was preprocessed with QIIME2 generally led to better predictive performance than the one from Kraken2. Concretely, the SVM trained on QIIME2 features achieved 78% accuracy, compared to 69% on Kraken2 features. Similarly, Random Forest scored 77.6% accuracy with QIIME2 and 71.5% with Kraken2. The gap was even more pronounced for MLP (70.6% vs 65.0%). The only model where Kraken2 performed in a similar way was Logistic Regression (slightly better with Kraken2: 67.9% vs 66.1%), likely because the added variables helped compensate for the linear model's simplicity.

These results indicate that the additional noise introduced by Kraken2 had a bigger importance than the benefits of the finer granularity. In this pipeline, several feature selection strategies (variance filtering, or keeping only the most abundant taxa) were tested. The best results for both QIIME2 and Kraken2 were consistently achieved by removing a portion of the rarest taxa—typically keeping only the top 10%–20% most abundant taxa. This preprocessing choice was optimal in 6 out of 8 configurations, confirming that very low-abundance taxa mostly add statistical noise.

5.6 Limitations of the Study

Despite the insights gained from this work, several limitations must be analyzed as they may affect the scope and generalizability of the results.

First, as already mentioned before, the study relies exclusively on publicly available datasets collected from multiple independent studies. This approach provided a larger sample, but it also introduced heterogeneity in sample populations, sequencing protocols, and inevitably in data quality across the different sources. Such variability can lead to dataset bias with the models possibly learning study-specific artifacts rather than true microbial patterns related to the diseases. Therefore, the classifier performance observed might not fully generalize to new datasets and has to be treated with great attention if used in a diagnosis tool. The current models should be considered exploratory and absolutely not definitive diagnostic tools.

Then there is an uneven representation of the different classes in the dataset obtained in this study. Certain disease categories (typically laminitis) are severely underrepresented compared to others like colic or healthy controls. This class imbalance can bias not only the training process but also the final evaluation metrics. Laminitis cases comprised only a small fraction of the samples, which makes it challenging for the models to learn stable patterns for this condition, leading to poor results for this condition. Despite using mitigating strategies, the low abundance of samples for certain outcomes reduces confidence in those predictions. Furthermore, it makes laminitis more prone to overfitting.

Third, the findings are influenced by bioinformatics pipelines used. The reliance on 16S rRNA gene data imposes constraints on taxonomic resolution and accuracy. Very high sequence similarity (e.g. >97%) can occur among distinct species complicating a lot the abundance estimates and taxonomic assignments. Moreover, most reference taxonomic databases were developed using human-associated microbiota.

Thus, they may lack adequate representation of microbial taxa specific to horses, which ultimately reduces the accuracy of classification.

This study did not incorporate additional host and environmental metadata (e.g. horse breed, age, diet, management, or health history) into the predictive models, which is another important limitation. The results of the study by Irwin et al. achieved better results by incorporating other information related to the patient [7]. Without host metadata, potential confounding factors that might influence the presence of a disease cannot be accounted for.

There are also limitations in the biological interpretability of the machine learning models used for classification. Complex classifiers that offer a strong predictive capability were employed, but their decision boundaries are not interpretable in biological terms. As shown before, feature importance rankings can be extracted and it can be examined which taxa strongly influence the model outputs, but these models still operate as "black boxes". The relationships they learn between the features and the disease status do not directly reveal causal explanations. It remains particularly challenging to translate the model outcomes to biological interpretations. This means that further analysis is necessary to understand why the models make certain predictions. Without such interpretative steps, the utility of the models in guiding interventions but also in helping the biological understanding of equine gut health is limited, marking this as an important limitation of the study.

5.7 Future perspectives

Multiple interesting insights emerge from this thesis leading to potential future work, both to enhance the classification performance but also to make the understanding of the equine microbiome broader.

A first natural extension of this work would be the integration of host and environmental metadata into the predictive models. Variables such as age, sex, breed, diet, antibiotic history, or management practices can significantly influence microbial composition and give important clues on which disease affects a horse. High-starch diets have been directly linked to laminitis and colic episodes due to microbial dysbiosis. Transportation stress, even after a 2-hour truck ride, can lead to colitis and diarrhea [4]. It is also interesting to consider the integration of functional data (e.g. fecal metabolites, microbial gene profiles) in addition to taxonomic composition, to determine if predicted functional traits of the microbiome improve disease discrimination.

Another perspective is to go toward longitudinal and temporal data. The datasets used in this thesis capture a single snapshot per individual. However, equine diseases such as colitis and laminitis evolve over time, and early microbial shifts may precede any other clinical symptom. Time-series microbiome data would allow the development of models that can detect disease onset or predict progression of the microbiome, enabling the development of early-warning tools and predictive systems. This would require collaboration with veterinary clinics to design prospective studies, but the benefits for equine healthcare could be substantial. However, a particularly large amount of data would be needed for such work.

The modest performance of the deep learning models in this study suggests that there is room for improvement for example with a better architectural design. The performance of the GNN model in this study could be improved in several ways. First, the graph construction process was based on simple similarity measures (Bray-Curtis and Euclidean distance), which probably does not capture all the biological relationships between samples; more biologically informed metrics or learned edge weights could provide a more meaningful structure. More advanced graph neural network variants could be tested, for example a Graph Attention Network that weights neighbors differently. Hyperparameter tuning was limited to a few options, a larger search might produce better configurations. Another perspective would be to train the graph on more diverse datasets, perhaps including some human microbiome data, could help the GNN learn more general patterns of microbial interactions that may not be specific to the equine gut. Moreover, combining both the `QIIME2` and `Kraken2` outputs into a merged representation could make use of the strengths of both pipelines.

6 Conclusion

This thesis explored the use of machine learning algorithms to classify equine diseases (laminitis, colitis, and colic) related to the gastrointestinal health, based uniquely on the gut microbiome profiles. These diseases can have serious consequences, sometimes leading to death, for horses and are associated in multiple studies with changes in intestinal microbial communities. The objective was to investigate whether these changes could be detected and perhaps used as biomarkers to help with the diagnosis, performed by veterinarians.

To do this, over 1,200 samples were preprocessed, collected from publicly available data, using two different pipelines, `QIIME2` and `Kraken2`, designed to extract relevant features from sequencing data. Several classification models were tested and compared, from classical approaches like Logistic Regression, Support Vector Machines and Random Forests, to more complex architectures such as CNNs and GNNs. A few additional steps like feature selection, dimensionality reduction, and cross-validation were applied to achieve the most reliable results possible.

The best performance was achieved with two classical models: the Support Vector Machine and the Random Forest classifier. Both achieved an accuracy of around 78% in multi-class classification, and nearly 85% in the binary task of separating healthy horses from the sick ones. Deep learning models like CNN and GNN performed slightly worse, with with the best GNN configuration reaching around 72% accuracy despite extensive tuning. The models based on `QIIME2` outperformed those based on `Kraken2` in almost all settings.

From a more biological perspective, the Random Forest model was able to identify microbial taxa already known in the literature to be associated with the analyzed disorders, such as Escherichia/Shigella, Fusobacterium, and Akkermansia. These findings provide the idea that the equine microbiome carries useful information about the animal's health, and that it could be used to assist in early diagnosis.

However, the study has non-negligible limitations. The dataset used was imbalanced, with very few samples for laminitis and few for colitis. This probably reduced the ability of the models to learn correct patterns for those two conditions. Also, no further metadata was used, even though other variables are known to influence the microbiota. The use of 16S rRNA sequencing limits the taxonomic resolution, making it harder to distinguish between similar bacterial species. Lastly, the heterogeneity of the data (coming from different studies and sequencing platforms) may have introduced unwanted biases.

Future work could focus on integrating more metadata to improve predictions, or using longitudinal data to detect disease onset earlier than when the symptoms are already there. Another direction could be to test other architectures, like graph attention networks or methods that combine several pipelines. Improving interpretability is also a goal, to help veterinarians understand the predictions made by the models and make a diagnosis based on tangible information.

In conclusion, this thesis shows the potential of machine learning applied to gut microbiome data for disease classification in horses. While more work is needed before such models could be used in clinical practice, the results show that this approach is promising. With better data and improved models, diagnostics made based on the microbiota could become a valuable tool in equine veterinary medicine.

Bibliography

[1] Jason Lloyd-Price et al. "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases". en. In: *Nature* 569.7758 (May 2019), pp. 655–662.

[2] Jack A Gilbert et al. "Current understanding of the human microbiome". en. In: *Nat Med* 24.4 (Apr. 2018), pp. 392–400.

[3] John F. Cryan et al. "The Microbiota-Gut-Brain Axis". In: *Physiological Reviews* 99.4 (Oct. 2019), pp. 1877–2013. DOI: 10.1152/physrev.00018.2018.

[4] Frédérique Chaucheyras-Durand et al. "Gastro-Intestinal Microbiota in Equines and Its Role in Health and Disease: The Black Box Opens". en. In: *Microorganisms* 10.12 (Dec. 2022).

[5] Maimaiti Tuniyazi, Wenqing Wang, and Naisheng Zhang. "A Systematic Review of Current Applications of Fecal Microbiota Transplantation in Horses". In: *Veterinary Sciences* 10.4 (2023). ISSN: 2306-7381. DOI: 10.3390/vetsci10040290. URL: https://www.mdpi.com/2306-7381/10/4/290.

[6] Divya Sharma, Andrew D Paterson, and Wei Xu. "TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction". In: *Bioinformatics* 36.17 (May 2020), pp. 4544–4550. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa542. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/17/4544/50677219/btaa542.pdf. URL: https://doi.org/10.1093/bioinformatics/btaa542.

[7] Christopher Irwin et al. *Graph Neural Networks for Gut Microbiome Metaomic data: A preliminary work*. 2024. arXiv: 2407.00142 [cs.LG].

[8] OpenAI. *ChatGPT (version GPT-4)*. https://chat.openai.com/chat. 2025.

[9] National Cancer Institute. *Microorganism - NCI Dictionary of Cancer Terms*. Accessed: 2025-04-25. 2025. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/microorganism.

[10] Encyclopedia Britannica. *Types of Microorganisms*. Accessed: 2025-04-25. 2025. URL: https://www.britannica.com/science/microbiology/Types-of-microorganisms.

[11] Srinivas Sulugodu Ramachandra et al. "Evaluating models and assessment techniques for understanding oral biofilm complexity". In: *MicrobiologyOpen* 12.6 (2023).

[12] Kaijian Hou et al. "Microbiota in health and diseases". In: 7.1 (Apr. 2022), p. 135.

[13] Vanessa K. Ridaura et al. "Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice". In: *Science* 341.6150 (2013), p. 1241214. DOI: 10.1126/science.1241214.

[14] Elaine Y Hsiao et al. "Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders". en. In: *Cell* 155.7 (Dec. 2013), pp. 1451–1463.

[15] Yasmine Belkaid and Timothy W Hand. "Role of the microbiota in immunity and inflammation". en. In: *Cell* 157.1 (Mar. 2014), pp. 121–141.

[16] Patrice D. Cani et al. "Metabolic Endotoxemia Initiates Obesity and Insulin Resistance". In: *Diabetes* 56.7 (July 2007), pp. 1761–1772. DOI: 10.2337/db06-1491. URL: https://doi.org/10.2337/db06-1491.

[17] Timothy R Sampson et al. "Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease". en. In: *Cell* 167.6 (Dec. 2016), 1469–1480.e12.

[18] Anne Kauter et al. "The gut microbiome of horses: current research on equine enteral microbiota and future perspectives". In: *Animal Microbiome* 1.1 (Nov. 2019), p. 14.

[19] Laurie Boucher et al. "Current Understanding of Equine Gut Dysbiosis and Microbiota Manipulation Techniques: Comparison with Current Knowledge in Other Species". en. In: *Animals (Basel)* 14.5 (Feb. 2024).

[20] Sanju Tamang. *Illumina Sequencing: Principle, Steps, Uses, Diagram*. Consulté le 3 mai 2025. 2024. URL: https://microbenotes.com/illumina-sequencing/#stepsprocess-of-illumina-sequencing.

[21] J Michael Janda and Sharon L Abbott. "16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls". en. In: *J Clin Microbiol* 45.9 (July 2007), pp. 2761–2764.

[22] Royal Veterinary College. *Laminitis*. Accessed: 2025-04-22. 2024. URL: https://www.rvc.ac.uk/equine-vet/information-and-advice/fact-files/laminitis#panel-key-points-about-laminitis-in-horses.

[23] M.M. Sloet van Oldruitenborgh-Oosterbaan and. "Laminitis in the horse: A review". In: *Veterinary Quarterly* 21.4 (1999). PMID: 10568001, pp. 121–127. DOI: 10.1080/01652176.1999.9695006. eprint: https://doi.org/10.1080/01652176.1999.9695006. URL: https://doi.org/10.1080/01652176.1999.9695006.

[24] Southwest Equine Veterinary Group. *Understanding Laminitis*. Accessed: 2025-04-23. n.d. URL: http://www.southwestequine.com.au/understanding-laminitis-fouunder/.

[25] University of Liverpool Equine Hospital. *What Causes Laminitis?* Accessed: 2025-04-23. 2023. URL: https://www.liverpool.ac.uk/equine/common-conditions/laminitis/whatcauseslaminitis/.

[26] Claire E Wylie et al. "Risk factors for equine laminitis: a systematic review with quality appraisal of published evidence". en. In: *Vet J* 193.1 (Nov. 2011), pp. 58–66.

[27] Brian Beasley. *Laminitis in Horses*. 2024/06 2024. URL: https://www.msdvetmanual.com/musculoskeletal-system/disorders-of-the-foot-in-horses/laminitis-in-horses.

[28] SUCCEED Equine. *Colitis in Horses: Causes, Symptoms, and Treatment*. Accessed: 2025-04-23. 2023. URL: https://www.succeed-vet.com/educational-resources/disease-library/colitis/.

[29] MadBarn Inc. *Colitis in Horses: Causes, Symptoms & Treatment*. Accessed: 2025-04-23. 2023. URL: https://madbarn.com/colitis-in-horses/.

[30] Marcio C. Costa et al. "Comparison of the Fecal Microbiota of Healthy Horses and Horses with Colitis by High Throughput Sequencing of the V3-V5 Region of the 16S rRNA Gene". In: *PLOS ONE* 7 (July 2012). URL: https://doi.org/10.1371/journal.pone.0041484.

[31] MSD Veterinary Manual. *Overview of Colic in Horses*. Accessed April 22, 2025. 2023. URL: https://www.msdvetmanual.com/digestive-system/colic-in-horses/overview-of-colic-in-horses.

[32] Royal Veterinary College. *Colic - Information and Advice*. Accessed April 22, 2025. 2023. URL: https://www.rvc.ac.uk/equine-vet/information-and-advice/fact-files/colic#panel-key-points.

[33] University of Minnesota Extension. *Colic in your horse*. Accessed: 2025-04-22. 2023. URL: https://extension.umn.edu/horse-health/colic-your-horse#types-of-colic--71561.

[34] Blue Cross. *Horse colic: prevention and management*. Accessed April 22, 2025. n.d. URL: https://www.bluecross.org.uk/advice/horse/health-and-injuries/horse-colic-prevention-and-management#:~:text=Treatment%20for%20colic,and%20treatments%20to%20the%20stomach.

[35] Daniel Berrar. "Cross-Validation". In: Jan. 2018. ISBN: 9780128096338. DOI: 10.1016/B978-0-12-809633-8.20349-X.

[36] Gopi Sambasivam. *K-Fold Cross Validation in Keras*. https://medium.com/the-owl/k-fold-cross-validation-in-keras-3ec4a3a00538. Accessed: 2025-05-06. 2018.

[37] Peshawa J. Muhammad Ali and Rezhna H. Faraj. *Data Normalization and Standardization: A Technical Report*. Tech. rep. 1. Machine Learning Technical Reports, 2014, pp. 1–6.

[38] Spiceworks. *What Is Principal Component Analysis?* Accessed: 2025-05-10. 2023. URL: https://www.spiceworks.com/tech/big-data/articles/what-is-principal-component-analysis/.

[39] John A. Lee and Michel Verleysen. *Nonlinear Dimensionality Reduction*. New York: Springer, 2007. ISBN: 9780387393513. DOI: 10.1007/978-0-387-39352-3.

[40] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

[41] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. 1996, pp. 226–231.

[42] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[43] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Draft of January 12, 2025, Chapter 5: Logistic Regression. Copyright 00A9 2024. All rights reserved., 2025.

[44] IBM Corporation. *What are support vector machines (SVMs)?* Consulté le 28 avril 2025. 2023. URL: https://www.ibm.com/think/topics/support-vector-machine.

[45] William S. Noble. "What is a support vector machine?" In: *Nature Biotechnology* 24 (2006), pp. 1565–1567. DOI: 10.1038/nbt1206-1565.

[46] Constantin F. Aliferis and Ioannis Tsamardinos. "Support Vector Machines". In: *MEDINFO 2004, T02: Machine Learning Methods for Decision Support and Discovery*. Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University. 2004.

[47] IBM. *What Is Random Forest?* Accessed: 2025-04-28. 2024. URL: https://www.ibm.com/think/topics/random-forest.

[48] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[49] M.W Gardner and S.R Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". In: *Atmospheric Environment* 32.14 (1998), pp. 2627–2636. ISSN: 1352-2310. DOI: https://doi.org/10.1016/S1352-2310(97)00447-0. URL: https://www.sciencedirect.com/science/article/pii/S1352231097004470.

[50] Michel Verleysen and Cyril de Bodt. *Nonlinear regression with Multi-Layer Perceptrons*. Lecture slides, ELEC2870 - Machine Learning: Regression and Dimensionality Reduction. Version 4.5. 2023. URL: https://www.uclouvain.be.

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[52] Sumit Saha. *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. https://medium.com/data-science/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. Accessed: 2025-05-06. 2018.

[53] Ole Dreessen. "Training Convolutional Neural Networks: What Is Machine Learning?—Part 2". In: *Analog Dialogue* 57.1 (Mar. 2023).

[54] Serkan Kiranyaz et al. "1D convolutional neural networks and applications: A survey". In: *Mechanical Systems and Signal Processing* 151 (2021), p. 107398. ISSN: 0888-3270. DOI: https://doi.org/10.1016/j.ymssp.2020.107398. URL: https://www.sciencedirect.com/science/article/pii/S0888327020307846.

[55] Bharti Khemani et al. "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions". In: *Journal of Big Data* 11.1 (Jan. 2024), p. 18.

[56] Jie Zhou et al. "Graph Neural Networks: A Review of Methods and Applications". In: *AI Open* 1 (2020), pp. 57–81. DOI: 10.1016/j.aiopen.2021.01.001.

[57] Sequence Read Archive Submissions Staff. *Downloading and Accessing Data*. SRA Knowledge Base [Internet]. [Updated 2014 Mar 18]. Bethesda (MD): National Center for Biotechnology Information (US), 2011.

[58] NCBI. *SRA Toolkit. National Center for Biotechnology Information*. https://github.com/ncbi/sra-tools. 2025.

[59] Robert C. Edgar. *FASTQ files*. https://www.drive5.com/usearch/manual7/fastq_files.html. Accessed: 2025-05-28. 2013.

[60] R Thanissery et al. "Clostridioides difficile carriage in animals and the associated changes in the host fecal microbiota". In: *Anaerobe* 66 (2020), p. 102279. DOI: 10.1016/j.anaerobe.2020.102279.

[61] Zachary L. McAdams et al. "A novel dataset of 2,362 equine fecal microbiomes from eight veterinary teaching hospital on three continents reveals dominant effects of geography, breed, and disease". In: *bioRxiv* (2024). DOI: 10.1101/2024.10.21.619412. URL: https://www.biorxiv.org/content/early/2024/10/22/2024.10.21.619412.

[62] F Lara, R Castro, and P Thomson. "Changes in the gut microbiome and colic in horses: Are they causes or consequences?" In: *Open Veterinary Journal* 12.2 (2022), pp. 242–249. DOI: 10.5455/OVJ.2022.v12.i2.12.

[63] CA McKinney et al. "The fecal microbiota of healthy donor horses and geriatric recipients undergoing fecal microbial transplantation for the treatment of diarrhea". In: *PLoS One* 15.3 (2020), e0230148. DOI: 10.1371/journal.pone.0230148.

[64] C Loublier et al. "Longitudinal Changes in Fecal Microbiota During Hospitalization in Horses With Different Types of Colic". In: *Journal of Veterinary Internal Medicine* 39.2 (2025), e70039. DOI: 10.1111/jvim.70039.

[65] C Arnold et al. "Alterations in the Fecal Microbiome and Metabolome of Horses with Antimicrobial-Associated Diarrhea Compared to Antibiotic-Treated and Non-Treated Healthy Case Controls". In: *Animals* 11.6 (2021), p. 1807. DOI: 10.3390/ani11061807.

[66] Y Kinoshita, H Niwa, and T Ueno. "Minimal disruption of equine gut microbiota by intravenous cephalothin treatment". In: *Journal of Veterinary Medical Science* (2025). Epub ahead of print. DOI: 10.1292/jvms.25-0105.

[67] Susan M Steelman et al. "Pyrosequencing of 16S rRNA genes in fecal samples reveals high diversity of hindgut microflora in horses and potential links to chronic laminitis". In: *BMC Veterinary Research* 8 (2012), p. 231. DOI: 10.1186/1746-6148-8-231.

[68] Núria Mach et al. "Gut microbiota resilience in horse athletes following holidays out to pasture". In: *Scientific Reports* 11.1 (2021), p. 5007. DOI: 10.1038/s41598-021-84497-y. URL: https://doi.org/10.1038/s41598-021-84497-y.

[69] Evan Bolyen et al. "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2". In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 852–857.

[70] Benjamin J Callahan et al. “DADA2: High-resolution sample inference from Illumina amplicon data”. en. In: *Nat. Methods* 13.7 (July 2016), pp. 581–583. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3869. URL: http://dx.doi.org/10.1038/nmeth.3869.

[71] Isabel Abellan-Schneyder et al. “Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing”. en. In: *mSphere* 6.1 (Feb. 2021).

[72] Andrei Prodan et al. “Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing”. en. In: *PLoS One* 15.1 (Jan. 2020), e0227434.

[73] Pelin Yilmaz et al. “The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks”. In: *Nucleic Acids Research* 42.D1 (Nov. 2013), pp. D643–D648. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1209. URL: https://doi.org/10.1093/nar/gkt1209.

[74] Michael S. Robeson et al. “RESCRIPt: Reproducible sequence taxonomy reference database management for the masses”. In: *bioRxiv* (2020). DOI: 10.1101/2020.10.05.326504. URL: https://www.biorxiv.org/content/early/2020/10/05/2020.10.05.326504.

[75] Derrick E Wood, Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2”. In: *Genome Biology* 20.1 (Nov. 2019), p. 257.

[76] Daniel McDonald et al. “An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea”. In: *The ISME Journal* 6.3 (Mar. 2012), pp. 610–618.

[77] James R Cole et al. “Ribosomal Database Project: data and tools for high throughput rRNA analysis”. en. In: *Nucleic Acids Res* 42.Database issue (Nov. 2013), pp. D633–42.

[78] Gorkem Gunay. *Understanding Parameters of SVM*. Accessed: 2025-05-11. 2019. URL: https://www.kaggle.com/code/gorkemgunay/understanding-parameters-of-svm.

[79] Lars Buitinck et al. “API design for machine learning software: experiences from the scikit-learn project”. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[80] Data Science. *Hyperparameter Tuning the Random Forest in Python using Scikit-learn*. Accessed: 2025-05-11. 2018. URL: https://medium.com/data-science/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74.

[81] Maxim Gusarov. *Do I Need to Tune Logistic Regression Hyperparameters?* Accessed: 2025-05-11. 2022. URL: https://medium.com/codex/do-i-need-to-tune-logistic-regression-hyperparameters-1cb2b81fca69.

[82] Sejal Jaiswal. *Multilayer Perceptrons in Machine Learning: A Comprehensive Guide*. Accessed: 2025-05-11. 2025. URL: https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning.

[83] William L. Hamilton, Rex Ying, and Jure Leskovec. "Inductive Representation Learning on Large Graphs". In: *CoRR* abs/1706.02216 (2017). arXiv: 1706.02216. URL: http://arxiv.org/abs/1706.02216.

[84] George Armstrong et al. "Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data". In: *Frontiers in Bioinformatics* Volume 2 - 2022 (2022). ISSN: 2673-7647. DOI: 10.3389/fbinf.2022.821861. URL: https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.821861.

[85] Jingtai Ma et al. "Interpretable machine learning algorithms reveal gut microbiome features associated with atopic dermatitis". In: *Frontiers in Immunology* Volume 16 - 2025 (2025). ISSN: 1664-3224. DOI: 10.3389/fimmu.2025.1528046. URL: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2025.1528046.

[86] Hengameh Chloé Mirsepasi-Lauridsen et al. "Escherichia coli Pathobionts Associated with Inflammatory Bowel Disease". In: *Clinical Microbiology Reviews* 32.2 (2019), e00060–18. DOI: 10.1128/CMR.00060-18. URL: https://doi.org/10.1128/CMR.00060-18.

7 Appendix

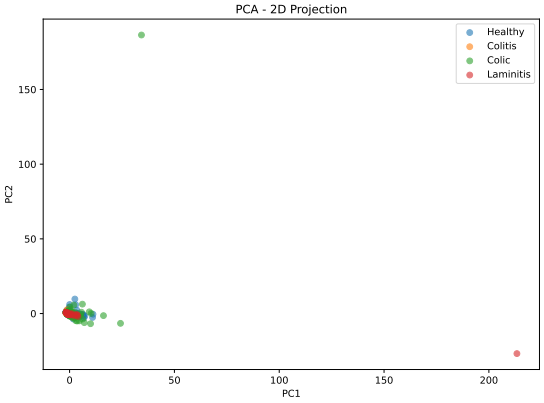


Figure 7.1: 2D PCA projection of QIIME2-based features. Each point represents a sample colored by its clinical class (Healthy, Colic, Colitis, Laminitis).

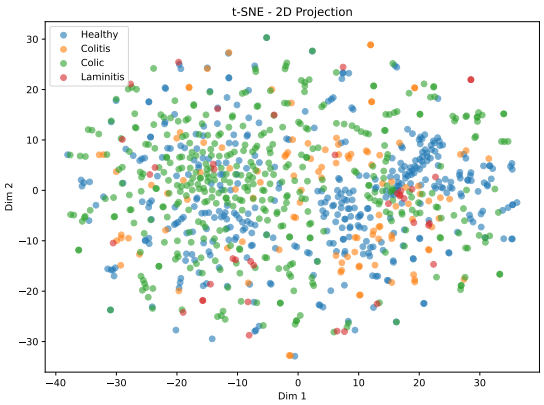


Figure 7.2: 2D t-SNE projection of QIIME2-based features, colored by clinical class.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**