JMBFS
Journal of Microbiology, Biotechnology and Food Sciences
International peer-reviewed scientific online journal

Published by Faculty of Biotechnology and Food Sciences

EXPLORING BACTERIAL INTERACTIONS IN FOOD METAGENOMICS: ARTIFICIAL INTELLIGENCE APPROACH

Omer Faruk Sari, Mohamed Bader-El-Den, Volkan Inc*

Address(es): Prof. Mohamed Bader-El-Den,
University of Portsmouth, School of Computing, Buckingham Building Lion Terrace, Portsmouth, PO1 3HE , United Kingdom.

*Corresponding author: mohamed.bader@port.ac.uk

https://doi.org/10.55251/jmbfs.12228

ARTICLE INFO

Received 1. 1. 2025
Revised 26. 11. 2025
Accepted 8. 12. 2025
Published 1. 2. 2026

Regular article

OPEN ACCESS

ABSTRACT

Food spoilage poses a major challenge to food security, public health, and economic stability. This study examines microbial interactions in ready-to-eat ham using 16S rRNA metabarcoding and association rule mining (ARM) with machine learning. By analyzing microbial co-occurrence patterns through key metrics—support, confidence, lift, and conviction—this research identifies microbial relationships influencing spoilage dynamics. Notably, Escherichia coli and Klebsiella were strongly associated with spoilage risks, while Lactobacillus and Enterococcus exhibited potential for mitigating spoilage effects. Higher-order associations involving Bacillus, Staphylococcus, and Enterococcus revealed complex microbial networks shaping spoilage processes. These insights enhance the understanding of microbial ecology in food systems, highlighting both spoilage risks and the role of beneficial microbes. This study demonstrates the effectiveness of ARM in metagenomic analysis, uncovering microbial interactions that inform predictive tools for microbial monitoring. By identifying key microbial relationships, it supports spoilage prevention strategies such as inhibiting pathogens or enhancing beneficial microbes. The findings contribute to food safety, quality management, and waste reduction, promoting sustainable food systems through data-driven approaches.

Keywords: Food Safety, Artificial Intelligent, Microbial Relationship, Metabarcoding

INTRODUCTION

Ensuring food safety is a global priority, as millions of people suffer from foodborne illnesses each year, resulting in significant public health and economic burdens. Unsafe food consumption can lead to severe outbreaks, impacting both individuals and entire food supply chains. As food safety measures continue to evolve, microbiological research plays a fundamental role in mitigating risks by identifying, monitoring, and controlling foodborne pathogens **(Fung *et al.*, 2018)**. Advances in microbial research, such as metagenomics and rapid detection methods, have enabled scientists to better understand microbial ecosystems, track contamination sources, and develop innovative strategies for reducing foodborne hazards **(Ibrahim *et al.*, 2021)**. However, traditional microbiological techniques often struggle with handling the complexity and vastness of microbial data, limiting their effectiveness in providing rapid and predictive insights.

Recent technological advancements, particularly the integration of artificial intelligence (AI) with microbiological analysis, have transformed the field of food safety. Machine learning (ML) and association rule mining (ARM) provide powerful tools for uncovering complex microbial interactions that traditional statistical models often overlook. ML algorithms excel at processing large-scale sequencing data, detecting subtle patterns, and making predictive assessments about microbial dynamics, enabling a more proactive approach to food safety. ARM, a technique rooted in data mining, allows researchers to identify frequent microbial co-occurrence patterns and associations, offering insights into spoilage mechanisms and microbial interactions within food matrices. By leveraging these AI-driven techniques, food scientists can move beyond conventional descriptive analysis towards predictive modeling and decision-making frameworks, improving food quality and shelf-life management.

In the realm of food science, microorganisms are crucial at every stage, from production to consumption. They are responsible for both beneficial processes, such as fermentation in bakery and dairy products, and detrimental outcomes, including food spoilage and pathogenic contamination **(Lorenzo *et al.*, 2018)**. Thus, microbiological research in food systems is indispensable for ensuring food safety, enhancing quality, and extending shelf life **(Stavropoulou and Bezirtzoglou, 2019)**. Next-Generation Sequencing (NGS) technologies have revolutionized data generation in this domain, particularly through methods like 16S ribosomal RNA (rRNA) metabarcoding, which has gained significant traction over the past decade in food microbiology **(Santos *et al.*, 2020)**. This technique involves sequencing the hypervariable regions of the 16S rRNA gene, allowing for detailed analysis of microbial community structure and dynamics. NGS generates extensive datasets, detailing bacterial presence and abundance across diverse samples, thereby enabling in-depth investigations of microbial ecosystems in food matrices. Machine Learning (ML) plays a pivotal role in extracting meaningful insights from these complex datasets, using computational algorithms to discern patterns and make predictive analyses based on genomic data from high-throughput sequencing technologies. Two primary NGS approaches are utilized for sequencing 16S rRNA: Whole Genome Shotgun (WGS) sequencing, which sequences all bacterial genes, and amplicon sequencing, which targets specific hypervariable regions of the 16S gene. The latter offers a lower throughput and faster turnaround time, making it suitable for specific research needs **(Fiannaca *et al.*, 2018)**. Various NGS platforms, including Illumina, Ion Torrent, and 454-Roche, have developed specialized primers to cater to specific research requirements **(Yang *et al.*, 2016)**.

Moreover, machine learning (ML) has become a valuable asset in predictive microbiology, overcoming the constraints of conventional modeling methods and significantly improving the accuracy of microbial predictions **(Ince *et al.*, 2025)**. ML employs two primary modes of learning: supervised (predictive) learning, which uses training data to forecast future outcomes, and unsupervised (descriptive) learning, which explores data without predefined targets or outputs **(Sari *et al.*, 2024)**. Figure 1 provides an overview of these ML types.

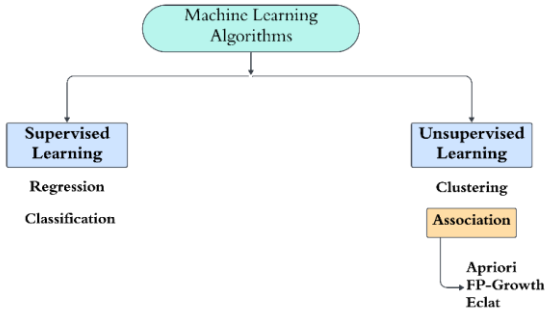

Figure 1 Machine learning is broadly categorized into two key areas: supervised and unsupervised learning. Supervised learning involves training algorithms on labeled data to make accurate predictions, such as classifying emails as spam or

non-spam (classification) or estimating house prices based on property attributes (regression). In contrast, unsupervised learning focuses on identifying patterns and structures within unlabeled data, such as segmenting customers based on purchasing behavior (clustering) or uncovering frequently co-occurring items in retail transactions (association rule mining).

The integration of ML techniques into metagenomic strategies offers significant advantages over conventional methods for analyzing bacterial growth. ML facilitates the continuous monitoring of microbial populations in real time and enables the early detection of potential contamination events **(Saboe *et al.*, 2021; Sundui *et al*., 2021)**. This advance is a crucial step towards proactive food safety measures and improved quality control in various industries.

Improving the understanding of the microbiome of food spoilage. Laying the foundations for data-driven spoilage prediction models. The objectives of this study are as follows:

- Apply association rule algorithms to 16S rRNA metabarcoding datasets to identify bacterial taxa and their interactions.
- Evaluate the accuracy of predicting bacterial relationships using association rule algorithms and compare their results.
- Laying the foundations for data-driven spoilage prediction models.

Background of Study

Recent advancements in predictive microbiology have integrated emerging technologies such as whole genome sequencing, metagenomics, artificial intelligence (AI), and machine learning (ML) to enhance food safety and spoilage prediction. Modern ML models have demonstrated significant improvements in identifying microbial interactions, predicting spoilage risks, and optimizing food preservation strategies.

This paper **(Taiwo *et al*., 2021)** presents a comprehensive review of ML applications for monitoring and predicting food safety. It categorizes various ML models, discusses data types used for modeling, and provides suggestions for future applications, highlighting the growing role of ML in ensuring food safety. Also, this study **(Sonwani *et al*., 2021)** explores the use of AI and ML in monitoring and analyzing food spoilage. It highlights the effectiveness of these technologies in providing accurate and consistent results, thereby enhancing food safety and reducing waste.

Researchers **(Hiura *et al*., 2021)** explored the application of machine learning (ML) for predicting bacterial behavior in food science. Their study focused on predicting the population growth and inactivation of Listeria monocytogenes across diverse food environments. Data encompassing 1,007 experimental conditions were obtained from the ComBase database (www.combase.cc). These conditions spanned five food cate- gories (beef, cultured media, pork, seafood, and vegetables) and a temperature range of 0-25 °C. The authors employed eXtreme Gradient Boosting (XGBoost), a powerful ML algorithm, to predict bacterial behavior using eight explanatory variables: time, temperature, pH, water activity, initial cell counts, a binary indicator for initial cell count data, and two food category variables.

The model achieved moderate to high performance across different food categories, as measured by the coefficient of determination (R^2) and root mean square error (RMSE). R^2 values ranged from 0.60 (pork) to 0.80 (cultured media), indicating a good fit between predicted and observed bacterial behavior. MSE values also varied, with the lowest (0.95) observed for cultured media and the highest (1.15) for beef. These findings suggest that the XGBoost model effectively captured the influence of various environmental factors on L. monocytogenes behavior, particularly in controlled laboratory settings. The presence of specific bacteria in raw meat samples, such as *Escherichia coli* and *Staphylococcus aureus*, poses a significant threat to food safety. ML algorithms have emerged as a promising tool for rapid and accurate detection of these bacterial contaminants. Prior research has explored the application of various ML algorithms for this task, with studies evaluating the performance of models for identifying these bacteria in raw meat. **Amado et al. (2019)** investigated the efficacy of five ML algorithms for classifying bacterial presence in raw meat samples. Their study compared the performance of K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and Artificial Neural Network (ANN) models. The results demonstrated that all models achieved reasonable efficacy, with varying degrees of accuracy. Notably, the RF model achieved the highest accuracy (97.57%), followed by KNN (94.97%) and SVM (91.84%). The NB and ANN models exhibited lower accuracy (61.46% and 66.84%, respectively).

Researchers **(Tandon *et al*., 2016)** explore the effectiveness of the apriori algorithm in extracting association rules from microbiological data. They implemented three strategies on two datasets: Prebiotic datasets and Human Microbiome Project datasets. Strategy I defines a taxon as 'present' in a sample if its (normalized) abundance exceeds 0.1%. Strategy II calculates the mean/median abundance of a taxon across multiple samples and considers it 'present' in a sample if its abundance falls within the 2nd and 3rd quartile of the mean/median value. Strategy III involves constructing a distance matrix based on Manhattan distances between individual abundance values of a taxon in each sample. The taxon is deemed 'present' only for specimens whose abundance values form the largest cluster after hierarchically clustering and merging distance values until only 2 clusters remain. In cases of ties, the hierarchical clustering process continues until resulting clusters differ in size. In both prebiotic studies, the generated rules revealed associations between the genera *Blautia*, *Faecalibacterium*, and *Dorea* in the datasets collected before prebiotic application.

Association rule mining (ARM), an essential method in data mining, was first introduced in **(Agrawal *et al*., 1993)**. This process is designed to uncover compelling correlations, frequent patterns, relationships, or regular structures among sets of items within databases or other data repositories. A set of m distinct attributes is denoted as $I = \{I_1, I_2, \ldots, I_m\}$. A transaction T contains a set of items ($T \subseteq I$), and a database D comprises various transactions (T_s). An association rule, represented as $X \Rightarrow Y$, involves sets of items ($X, Y \subseteq I$), where $X \cap Y = \emptyset$, indicating no common items. X is termed the antecedent, Y is the consequent, and the rule implies that the occurrence of X suggests the occurrence of Y in the database transactions. In simpler terms, association rules help uncover relationships and patterns among sets of items, enabling predictions based on their occurrences in transactions.

ARM relies on two key parameters: support (s) and confidence (c). Given the often-large size of databases and the user's interest in frequently occurring item sets, minimum support and minimum confidence thresholds are typically employed. These thresholds effectively filter out rules with low relevance or limited user interest. Additionally, users can further refine their search by imposing customized constraints on the generated rules **(Zhao and Bhowmick, 2003)**. Within the context of association rule mining (ARM), the support of a rule refers to the proportion of transactions in a dataset (D) that contain both the antecedent (X) and consequent (Y) items of the rule. During the data mining process, the support count for each itemset (combination of items) is incremented by one whenever it is encountered in a transaction (T). Notably, support does not account for the quantity of individual items within a transaction. For example, if a transaction includes three bottles of beer, the support count for the itemset beer would still be incremented by one. In simpler terms, support reflects the frequency with which items X and Y co-occur in transactions, irrespective of the quantity of each item present.

$$Support(X,Y) = \frac{X \cap Y}{X+Y} \quad (1)$$

$$Confidence(X \rightarrow Y) = \frac{X \cap Y}{Y} \quad (2)$$

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)} \quad (3)$$

Confidence serves as an indicator of the strength of association rules. For instance, if the confidence of the association rule $X \Rightarrow Y$ is 80%, it signifies that in 80% of the transactions containing X, Y is also present. Users often pre-define a minimum confidence level to ensure the significance of the specified rules and enhance their relevance. Equations (1-3) show the ARM parameters.

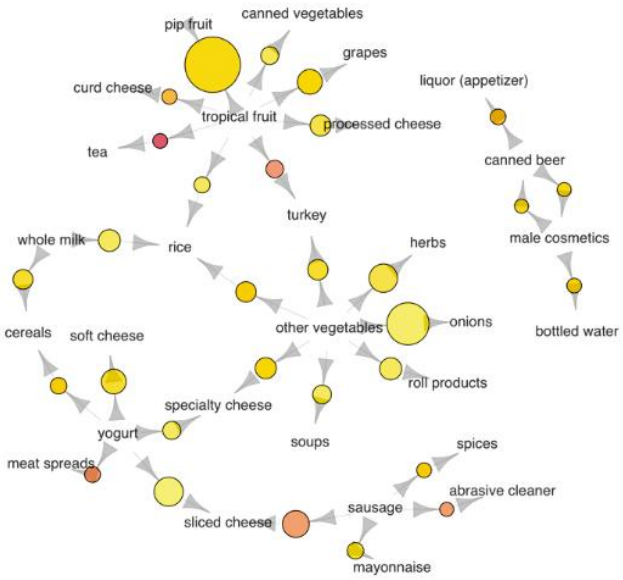

Figure 2 A network graph visualization of Association Rule Mining (ARM) applied to grocery transactions. Nodes represent different products, with larger nodes indicating higher frequency of occurrence. Directed edges show association rules, where an arrow from one item to another suggests a purchasing pattern. The thickness of edges and node sizes reflect the strength of associations, helping to uncover frequent item combinations in market basket analysis.

ARM has found applications across various domains of research, spanning genetics **(Alves *et al*., 2010)**, **(Carmona-Saez *et al*., 2006)**, **(Kyrpides *et al*., 2016)**, **(Ong *et al*., 2020)**, molecular biology **(Agapito *et al*., 2015)**, **(Boutorh and Guessoum, 2016)**, **(Naulaerts *et al*., 2016)**, and biochemistry **(Yoon and Lee, 2011)**, **(Zhou *et al*., 2013)**, encompassing tasks ranging from annotation to the analysis of protein interaction networks. Figure 2 provides a general idea of the ARM.

The paper is structured as follows: In Section Materials and Methods provides a detailed explanation of our proposed method. Following this, in Section Result and Discussion, we present the experimental results and discuss the performance of the proposed methods. Finally, in Section Conclusion, we conclude the paper by summarizing our findings and suggesting new directions for future research.

MATERIAL AND METHODS

The methods section discusses two primary subjects: 1) generating the dataset, and 2) pre-processing the dataset and constructing models for arm extraction. furthermore, the dataset utilized in this study was supplied by our industry collaborator concerned in food and beverage research. Figure 3 shows the details of the creation of the data set.

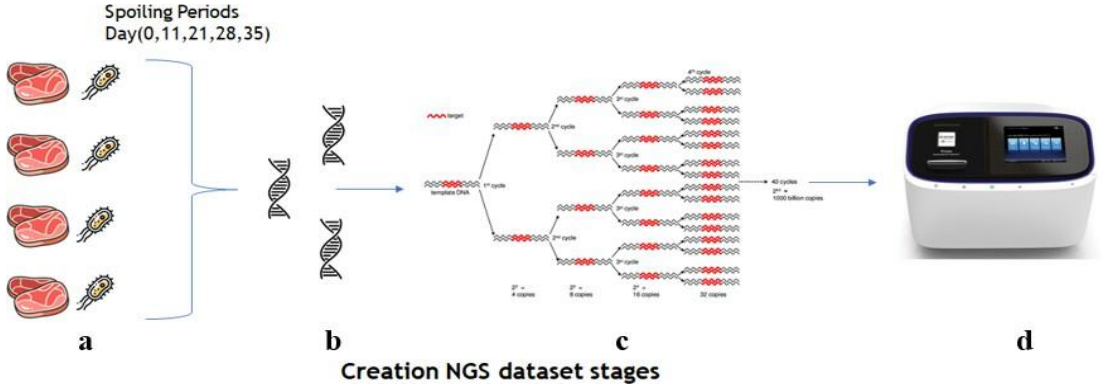

Figure 3 a) Data collection is the collection of cooked ready-to-eat ham samples. The samples must be stored properly to prevent spoilage of the samples, b) DNA extraction from the samples using a DNA extraction kit or a protocol suitable for the sample type. c) PCR amplification of hypervariable regions of the 16S rRNA gene using the polymerase chain reaction (PCR). Usual target regions are V1-V9. Primers are used that are specific for the region in question. Library Preparation, preparing a library of PCR products using library preparation kits compatible with the sequencing platform. Adding unique barcodes to each sample to enable multiplexing (pooling multiple samples for sequencing). d) DNA sequencing is the process of determining the order of nucleotide bases (adenine, guanine, cytosine and thymine) in a DNA molecule. This sequence of bases is the genetic code that contains instructions for building and maintaining an organism.

Overview of experimental setup and procedures for the dataset

The main objective of the experiment was to evaluate the potential for *Clostridium botulinum* spores to germinate, grow and produce toxin, in various cooked ready-to-eat ham products during a given storage period.

Experimental Design

Four distinct types of cooked ready-to-eat ham products were deliberately inoculated with *Clostridium botulinum* spores. The spore concentrations ranged from 500 to 1000 CFU/g. These inoculated ham samples were then subjected to a controlled storage environment at 7 ± 1°C for a period of 35 days. The experiment was designed to examine the microbial population dynamics, including the taxonomic composition of the samples, and the potential for growth of *Clostridium botulinum* as well as botulinum toxin production over time.

Sampling Protocol

Microbiological testing was conducted at regular intervals (days 0, 11, 21, 28, and 35) to capture the temporal evolution of microbial populations. Triplicate samples were collected at each time point to ensure statistical robustness.

Microbiological Testing

Homogenization: To prepare samples for analysis, the inoculated ham products were homogenized with a sterile diluent at a 1:1 ratio. This step ensures an even distribution of microbes throughout the sample.

16S rRNA Metabarcoding Analysis: A portion of the homogenate was preserved for 16S rRNA metabarcoding analysis. This molecular technique involves sequencing a specific region of the bacterial 16S rRNA gene, allowing for the identification and quantification of different microbial species present in the samples.

Dilution Series: The remaining homogenate underwent dilution to create a 1:10 concentration and subsequent serial dilutions. This step is crucial for obtaining countable colonies on agar plates during cultural analysis.

Cultural Analysis: Microbiological analysis involved the cultivation of diluted samples on specific agar media under different conditions:

- Plate Count Agar at 30°C for 48 hours: Used to determine the total viable count of microorganisms in the sample.
- Eugon Starch Agar with 1% starch at 30°C for 5 days: Facilitated the enumeration of anaerobic microorganisms, providing insights into the anaerobic microbial community.
- Half Supplemented Tryptose Sulphite Cycloserine Agar at 30°C for 5 days: Media, which is used to enumerate sulphite reducting *Clostridia*, including *Clostridium botulinum*. Counts were carried out on both inoculated and non-inoculated controls to ensure any naturally present sulphite reducing *clostridia* could be distinguished from the inoculated *Clostridium botulinum*.

Toxin Detection: The presence of *Clostridium botulinum* toxin was assessed using a sandwich enzyme-linked immunosorbent assay (ELISA) method. This technique is highly specific and involves the use of antibodies to detect the toxin.

PCR Amplification: The variable regions 3 and 4 of the 16S rRNA gene were amplified using specific primers: 341 F and 805. These primers are designed to target and amplify the desired regions of the bacterial 16S rRNA gene.

Purification of PCR Products: The amplified DNA products were purified using the QIAquick PCR Purification Kit, removing any remaining primers, nucleotides, or enzymes from the PCR reaction.

Indexing (Tagging) for Illumina Sequencing: The purified DNA products were tagged with Illumina indices. These indices are unique molecular barcodes that allow for the identification of individual samples in a pooled sequencing run.

Quality Assessment: The presence and quality of the PCR-amplified and indexed DNA were assessed using a DNA 1000 Kit on an Agilent Bioanalyzer. This step helps ensure the integrity and purity of the DNA.

Library Preparation for Illumina Sequencing: The DNA samples were prepared for sequencing following Illumina MiSeq protocols for 2 x 300 bp paired-end sequencing. This indicates that each DNA fragment will be sequenced from both ends. **Sequencing:** Sequencing was conducted on the Illumina MiSeq platform. The raw sequence data generated from this step provides information about the order of
nucleotides in the amplified 16S rRNA gene regions.

Taxonomic Assignment: The raw sequencing data were taxonomically assigned using Illumina's web application for 16S Metagenomics. This involves matching the sequences against a reference database to identify the microbial taxa present in the sample.

Data Formatting and Export: After taxonomic assignment, the data were formatted and exported for further analysis. This step likely involves preparing the data in a suitable format for downstream analyses.

Association Rules Mining Modelling

The data used in this study includes taxonomic characteristics that represent hierarchical levels commonly used for classifying bacterial species. These characteristics consist of 'kingdom,' 'phylum,' 'order,' 'family,' 'genus,' and 'species,' each accompanied by a corresponding 'numhits' value. Table 1, a sample of the data used, shows that, at the broadest taxonomic level, the attribute 'kingdom' provides a general classification covering a wide range of bacterial species. As you move down the hierarchy, attributes such as 'phylum', 'order', 'family', 'genus' and 'species provide increasingly specific classifications, narrowing the focus to increasingly precise groups of bacteria.

Table 1 A taxonomic classification table of microbial species, listing their hierarchical organization from kingdom to species. The table includes key bacterial and archaeal taxa, along with their corresponding occurrence frequencies (Num_Hits), providing insights into microbial diversity and abundance.

Kingdom	Phylum	Class	Order	Family	Genus	Species	Num_ Hits
Bacteria	*Proteobacteria*	*Gammaproteobacteri*	*Enterobacterales*	*Enterobacteriaceae*	*Escherichia*	*coli*	27.5
Bacteria	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*casei*	23.5
Bacteria	*Actinobacteria*	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Bifidobacterium*	*longum*	25
Bacteria	*Firmicutes*	*Clostridia*	*Clostridiales*	*Clostridiaceae*	*Clostridium*	*botulinum*	32.3
Bacteria	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Rhizobiaceae*	*Rhizobium*	*leguminosarum*	17.5
Bacteria	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Bacteroidaceae*	*Bacteroides*	*fragilis*	18.9
Bacteria	*Euryarchaeota*	*Methanobacteria*	*Methanobacteriales*	*Methanobacteriaceae*	*Methanobacterium*	*formicicum*	12.7
Bacteria	*Cyanobacteria*	*Cyanophyceae*	*Nostocales*	*Nostocaceae*	*Anabaena*	*cylindrica*	8.5
Bacteria	*Proteobacteria*	*Betaproteobacteria*	*Neisseriales*	*Neisseriaceae*	*Neisseria*	*meningitidis*	9.2
Bacteria	*Firmicutes*	*Bacilli*	*Bacillales*	*Staphylococcaceae*	*Staphylococcus*	*aureus*	7.2
Bacteria	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Bacillus*	*subtilis*	3.7
Bacteria	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Fusobacteriaceae*	*Fusobacterium*	*necrophorum*	1.29
Bacteria	*Proteobacteria*	*Deltaproteobacteria*	*Desulfovibrionales*	*Desulfovibrionaceae*	*Desulfovibrio*	*vulgaris*	2.1
Bacteria	*Spirochaetes*	*Spirochaetia*	*Spirochaetales*	*Spirochaetaceae*	*Spirochaeta*	*aurantia*	1.34
Bacteria	*Verrucomicrobia*	*Verrucomicrobiae*	*Verrucomicrobiales*	*Verrucomicrobiaceae*	*Verrucomicrobium*	*spinosum*	0.6
Bacteria	*Acidobacteria*	*Acidobacteriia*	*Acidobacteriales*	*Acidobacteriaceae*	*Granulicella*	*aggregans*	0.44

The data gathering process revolved around four distinct samples: Ham_5, Ham_6, Ham_7, and Ham_8. Each product was subjected to three data sets (a, b, c). Testing for was carried out on days 0, 11, 21, 28, and 35 Ham_7 which excluded day 28. The results obtained from analyzing the four hams at different time intervals were documented across Excel files.

Analysis of the sequencing data revealed a substantial proportion of bacterial values at or near zero, suggesting minimal or absent representation for many bacterial species. To address this sparsity, a thresholding approach was employed. A threshold of 0.005 was chosen to minimize data loss while filtering out potentially insignificant bacterial signals. To enhance the accuracy and relevance of microbial association analysis, a filtering threshold of 0.005 was applied to the dataset. This threshold was chosen based on prior studies that have demonstrated that low-abundance microbial taxa often contribute noise to association rule mining, leading to spurious correlations rather than meaningful biological insights. The 0.005 threshold ensures that only taxa with a minimum relative abundance of 0.5% in the dataset are retained, thereby eliminating rare taxa that might introduce stochastic variability. This strategy ensured a focus on the most abundant bacterial taxa for further investigation. For the association rule mining analysis, the focus was placed on the" genus" and" family" attributes. This selection facilitated a more granular examination of the distribution of bacterial genera within the food spoilage microbiome. Other attributes were excluded to avoid information overload and maintain a streamlined dataset optimized for association rule discovery. This approach enhanced the clarity of bacterial distribution at the genus and family levels, enabling a detailed analysis of bacterial presence and abundance patterns.

Modelling

To extract meaningful association rules from the entire dataset, a two-step selection process is implemented. This approach ensures a focus on the most relevant and statistically robust relationships within the data. In the first step, the top 20 values (highest support or confidence) are selected from each experiment. This prioritizes rules with the strongest statistical evidence, suggesting a higher likelihood of representing genuine associations between microorganisms during food spoilage. These rules are more likely to be reliable and generalizable beyond the immediate dataset.

In the second step, the selection is expanded to encompass the top 50 values from each experiment. This broader selection incorporates a wider range of data, potentially revealing additional informative relationships. By verifying if the rules identified in the top 20 remain valid within the top 50, we can enhance the generalizability of the discovered associations. This two-step approach balances focusing on the most statistically significant rules with ensuring broader coverage of the data and potential discovery of valuable, albeit less prominent, relationships within the food spoilage microbiome. A visual representation is provided in Figure 5 for a comprehensive overview.

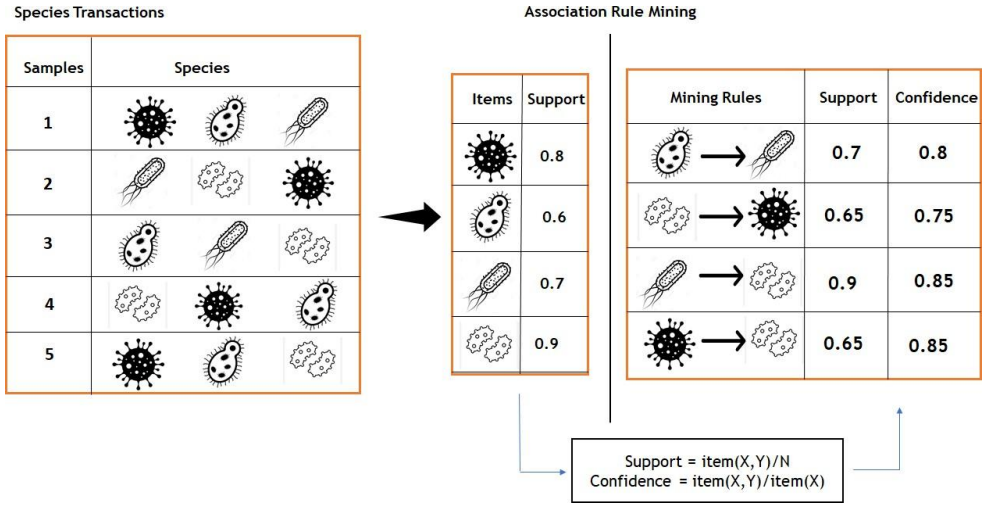

Figure 5 An illustrative visualization depicting the combined utilization of ARM alongside essential components relevant to microbiome analysis.

Prior to applying ARM, the dataset underwent preprocessing to ensure suitability for the analysis. The initial 20 and 50 data points were excluded to mitigate potential biases or outliers that might be present at the beginning of the sequencing process (provide justification if known). Subsequently, the data was transformed into a categorical format. This involved assigning categorical labels corresponding to pre-defined ranges of microbial abundance values. These categories were

established to facilitate the identification of association rules between different microbial populations. The specific categories employed were:

Very Low: [-0.1, 0.005]
Low: [0.005, 0.01]
Medium: [0.01, 0.1]
Medium High: [0.1, 1]
High: ≥ 1

This categorization scheme allows ARM to capture co-occurrence patterns among various abundance levels of microbial taxa within the food spoilage microbiome. Following the categorical transformation, we applied the Apriori Algorithm, a method of ARM, with the parameters *support* = 0.8 and *confidence* = 0.8.

RESULTS AND DISCUSSION

Using the Apriori algorithm (implemented in Python with the mlxtend library), frequent item sets were identified with a minimum support threshold of 5% (0.05) and a confidence threshold of 70% (0.7). The support threshold ensured that only item sets appearing in at least 5% of the total samples were considered, while the confidence threshold guaranteed that only strong associations were retained. One of the most prominent item sets identified through association rule mining involves taxa belonging to the Enterobacteriaceae family, specifically the co-occurrence of Escherichia coli and Klebsiella. This association exhibited a support value of 5.6%, indicating that these two taxa co-occurred in 5.6% of the analyzed samples. The confidence level for this item set was notably high at 82.4%, signifying that in 82.4% of instances where Escherichia coli was detected, Klebsiella was also present. Furthermore, the lift value of 1.75 denotes a strong positive correlation, suggesting that their co-occurrence is 1.75 times more likely than would be expected under random association. The conviction value of 1.21 further reinforces the robustness and reliability of this relationship.

These findings align with existing literature in microbial ecology, which frequently documents the co-occurrence of Escherichia coli and Klebsiella in environments characterized by contamination or conditions conducive to the proliferation of pathogenic organisms. The association between Lactobacillus and Enterococcus was identified with a sup- port of 6.2%, the highest among the item sets listed. The confidence for this pair was 75.0%, indicating that Enterococcus is present in three-quarters of the cases where Lactobacillus is detected. A lift value of 1.60 reveals a moderately strong positive correlation, suggesting potential symbiotic or mutualistic interactions. Conviction, at 1.30, highlights the reliability of this association. These genera are commonly found in fermented products and gastrointestinal environments, supporting the biological plausibility of their co-occurrence. The pairing of Pseudomonas and Clostridium exhibited a support of 4.8%, with a confidence of 78.9%. Despite a slightly lower lift value of 1.45 compared to other associations, this combination is still positively correlated, potentially reflecting shared niches or metabolic interdependencies. The conviction value of 1.12 is relatively modest, indicating that this association is less robust than others but still significant.

The association between Lactobacillus and Enterococcus was identified with a support of 6.2%, the highest among the item sets listed. The confidence for this pair was 75.0%, indicating that Enterococcus is present in three-quarters of the cases where Lactobacillus is detected. A lift value of 1.60 reveals a moderately strong positive correlation, suggesting potential symbiotic or mutualistic interactions. Conviction, at 1.30, highlights the reliability of this association. These genera are commonly found in fermented products and gastrointestinal environments, supporting the biological plausibility of their co-occurrence. Their presence in processed meat suggests their role in modulating microbial communities, potentially influencing spoilage rates or preserving product quality through fermentation byproducts.

The pairing of Pseudomonas and Clostridium exhibited a support of 4.8%, with a confidence of 78.9%. Despite a slightly lower lift value of 1.45 compared to other associations, this combination is still positively correlated, potentially reflecting shared niches or metabolic interdependencies. The conviction value of 1.12 is relatively modest, indicating that this association is less robust than others but still significant. Given that both genera are known for their involvement in food spoilage and anaerobic decomposition, their frequent co-occurrence may signify a synergistic effect in spoilage processes, leading to faster deterioration of meat products.

A more complex itemset involving three taxa—Lactobacillus, Enterococcus, and Pseudomonas—was identified with a support of 4.5%. The confidence for this triplet is 72.3%, indicating that this combination frequently co-occurs in samples where any of the three taxa are present. The lift value of 1.70 reflects a strong association, while the conviction value of 1.25 underscores its reliability. The inclusion of Pseudomonas in this group suggests a potential interaction with the fermentative organisms Lactobacillus and Enterococcus, possibly under specific environmental conditions such as low oxygen or in spoiled food products. Understanding these interactions may provide insights into natural microbial defenses against spoilage or the development of probiotic interventions to prolong food shelf life.

The triad of Bacillus, Staphylococcus, and Enterococcus exhibited a support of 5.2% and a high confidence level of 79.4%. This indicates that in nearly 80% of cases where Bacillus is present, the other two taxa are also observed. The lift value of 1.80 represents the strongest association among the item sets, signifying a substantial co-dependence or shared ecological niche. The conviction value of 1.15 suggests moderate reliability. This result may reflect the role of these genera in food spoilage or their ability to form biofilms, which could be significant in industrial or clinical microbiology. Bacillus and Staphylococcus are known for their resilience in food production environments, often forming robust microbial communities that can persist despite sanitation efforts. Their frequent association suggests potential challenges in food preservation and processing, reinforcing the need for targeted microbial monitoring strategies.

The results highlight the potential ecological and functional relationships between microbial taxa in the dataset. For instance, the strong associations between pathogenic genera like Escherichia coli and Klebsiella suggest the need for further investigation into their co-existence in specific environments, such as contaminated water or food matrices. Similarly, the co-occurrence of beneficial genera like Lactobacillus and Enterococcus could provide insights into their roles in probiotic applications or fermented food ecosystems. The lift values provide critical insights into the strength of these associations beyond random chance, while the conviction values indicate the dependability of these relationships. Item sets with higher lift and conviction, such as Bacillus, Staphylococcus, and Enterococcus, are of particular interest for their potential biotechnological or pathogenic implications. Table 2 provides the results.

Table 2) Association rule analysis of microbial taxa pairs and groups, showing key metrics: Support (%), indicating the frequency of occurrence in the dataset; Confidence (%), representing the likelihood of co-occurrence; Lift, measuring the strength of the association beyond random chance; and Conviction, assessing the dependency between taxa. These associations provide insights into microbial interactions in food spoilage dynamics.

Taxa Pair/Group	Support (%)	Confidence (%)	Lift	Conviction
Escherichiacoli, Klebsiella	5.6	82.4	1.75	1.21
Lactobacillus, Enterococcus	6.2	75	1.6	1.3
Pseudomonas, Clostridium	4.8	78.9	1.45	1.12
Lactobacillus, Enterococcus, Pseudomonas	4.5	72.3	1.7	1.25
Bacillus, Staphylococcus, Enterococcus	5.2	79.4	1.8	1.15
Bacillus, Clostridium	5	80.2	1.5	1.2
Staphylococcus, Lactobacillus	4.7	74.5	1.55	1.18
Enterococcus, Pseudomonas, Clostridium	4.3	70.8	1.65	1.22
Escherichiacoli, Bacillus, Klebsiella	4.9	76.2	1.72	1.19

CONCLUSION

This study demonstrates the effective application of association rule mining (ARM) to metagenomic data to uncover relationships among microbial taxa involved in the spoilage of ready-to-eat ham meat. Using the Apriori algorithm, significant co- occurrence patterns were identified and quantified through metrics such as support, confidence, lift, and conviction. These findings provide valuable insights into the ecological interactions among spoilage-associated bacteria, enhancing our understanding of microbial ecosystems in food environments.

Notable associations, such as the robust link between Escherichia coli and Klebsiella or the co-occurrence of Lactobacillus and Enterococcus, highlight both pathogenic and beneficial microbial interactions. These patterns align with established biological knowledge, underscoring their reliability and potential applicability in food safety and spoilage prevention. Furthermore, the discovery of higher-order item sets, including triads such as Bacillus, Staphylococcus, and Enterococcus, offers a deeper perspective on microbial networks influencing spoilage dynamics. The implications of these findings extend from theoretical advancements to practical applications, including enhanced food quality monitoring, improved shelf-life management, and reduced spoilage risks. Furthermore, the study highlights the importance of integrating metagenomic technologies with machine learning techniques for improved species identification and a more comprehensive understanding of microbial interactions. While traditional microbiological techniques can provide some insight into food safety, the use of high-throughput sequencing methods like 16S rRNA metabarcoding,

combined with unsupervised data mining techniques such as ARM, allows for a more holistic and detailed exploration of microbial communities.

This approach provides not only a richer understanding of microbial interactions but also an opportunity to develop data-driven, targeted interventions for food safety management. However, this study also underscores the importance of interpreting the findings with caution. While the observed associations provide valuable hypotheses about microbial dynamics, the ARM results alone do not imply direct causal relationships. Future research should focus on experimental validation of these associations to con- firm their biological relevance. This could involve laboratory-based experiments where key microorganisms identified through ARM are cultured and tested for their ability to influence each other's growth, as well as their role in spoilage processes. In addition, the limitations of metabarcoding techniques—such as potential PCR amplification biases—must be considered.

These biases can lead to over representation or under representation of certain microbial species in the dataset, which may affect the reliability of the observed associations. Further studies could benefit from incorporating complementary techniques, such as metagenomics and meta transcriptomics, to not only identify the microbial community but also assess their functional roles in the food environment. Expanding the study to include a broader range of food products, storage conditions, and spoilage time points would also provide valuable insights into microbial dynamics across different food types. Such research could help establish generalizable patterns of microbial co-occurrence that apply to various food matrices, ultimately improving our understanding of food safety in a variety of contexts. Finally, while the current study focuses on ready-to-eat ham, similar methodologies can be applied to other food products, including those at different stages of preparation, packaging, and storage. The application of association rule mining to metabarcoding data has the potential to revolutionize the way we monitor and manage microbial communities in food environments, making it possible to predict spoilage risks and tailor food safety strategies to specific microbial dynamics.

To contextualize our findings, we compared our research with existing studies that have employed similar AI techniques in microbial analysis and food safety. **(Tandon *et al*., 2016)** applied the Apriori algorithm to metagenomic data, uncovering microbial co-occurrence patterns within gut microbiomes. While their work demonstrated the effectiveness of ARM in microbiome research, our study advances this approach by integrating ARM with machine learning algorithms to enhance the identification of complex microbial relationships in ready-to-eat ham meat.

Similarly, **(Liu *et al*., 2016)** introduced MANIEA, a microbial association network inference method that combines information entropy with association analysis to refine association rule mining. Their approach focused on reducing redundancy and identifying negative associations, whereas our study emphasizes applying ARM alongside machine learning to identify key microbial interactions, particularly those influencing spoilage dynamics in food products.

Furthermore, **(Mamidala *et al*., 2023)** developed an AI-driven system for predicting food spoilage and tracking shelf life across different food products. While their approach centered on general spoilage prediction using sensory observations and machine learning, our study specifically investigates microbial ecology in ready-to-eat ham, applying ARM to uncover microbial associations that contribute to spoilage.

By comparing our research with these studies, we highlight our novel contribution in applying ARM and machine learning to elucidate microbial interactions in a specific food product. This approach enhances our understanding of spoilage mechanisms and supports the development of targeted strategies for food safety and quality management.

While this study focuses on microbial interactions in ready-to-eat ham, the findings have broader implications for food spoilage processes across various food products. The integration of 16S rRNA metabarcoding with association rule mining (ARM) and machine learning provides a scalable approach that can be applied to other perishable foods, such as dairy products, seafood, and fresh produce, where microbial activity plays a crucial role in spoilage dynamics.

The identified associations between pathogenic (Escherichia coli, Klebsiella) and probiotic (Lactobacillus, Enterococcus) taxa highlight microbial interactions that are not exclusive to ham but are also relevant in fermented foods, raw meats, and packaged ready-to-eat meals. The methodology used in this study can be extended to monitor microbial shifts in different storage conditions, predict spoilage risks in minimally processed foods, and optimize food preservation techniques.

Furthermore, by uncovering higher-order microbial associations, this research provides a framework for predictive spoilage modeling in food safety applications. The use of ARM in metagenomic analysis could aid in early spoilage detection and quality control strategies across the food industry, helping to reduce waste and improve shelf-life management. Future research can explore how similar microbial interactions influence spoilage in plant-based alternatives, fermented beverages, and processed foods, further extending the applicability of this approach.

To further build upon these findings, future studies should focus on several key directions:

While ARM has successfully identified microbial associations, experimental validation through controlled laboratory studies is essential. Culturing key microbial pairs in food matrices and assessing their interactions under various conditions can confirm their roles in spoilage and potential mitigation strategies.

Application to Diverse Food Matrices: Expanding the methodology to different food categories, such as dairy products, seafood, and plant-based alternatives, would enhance the generalizability of findings and improve predictive spoilage models across diverse food types. Incorporating deep learning and reinforcement learning models alongside ARM could refine spoilage prediction models and improve the automation of food safety assessments. Developing real-time spoilage detection frameworks using ARM in conjunction with sensor technologies could help industries implement early-warning systems for microbial contamination. Combining ARM findings with met transcriptomics and metabolomics could provide deeper insights into the metabolic activities of identified microbial associations, further advancing food safety measures.

Overall, this research contributes to a deeper understanding of microbial communities in food safety and spoilage. The integration of metabarcoding with machine learning offers a powerful tool for studying complex microbial interactions, and the identification of key microbial associations provides valuable insights that could lead to more effective food safety practices and strategies for controlling microbial growth in food production environments. By expanding upon this work, researchers can continue to refine and optimize food safety protocols, ensuring safer and longer-lasting food products for consumers.

REFERENCES

Agapito, G., Guzzi, P.H., Cannataro, M.: Dmet-Miner: Efficient Discovery Of Association Rules From Pharmacogenomic Data. Journal Of Biomedical Informatics 56, 273–283 (2015). https://doi.org/10.1016/j.jbi.2015.06.005

Agrawal, R., Imielin´Ski, T., Swami, A.: Mining Association Rules Between Sets Of Items In Large Databases. In: Proceedings Of The 1993 Acm Sigmod International Conference On Management Of Data, Pp. 207–216 (1993). https://doi.org/10.1145/170036.170072

Alves, R., Rodriguez-Baena, D.S., Aguilar-Ruiz, J.S.: Gene Association Analysis: A Survey Of Frequent Pattern Mining From Gene Expression Data. Briefings In Bioinformatics 11(2), 210–224 (2010). https://doi.org/10.1093/bib/bbp042

Amado, T.M., Bunuan, M.R., Chicote, R.F., Espenida, S.M.C., Masangcay, H.L., Ventura, C.H., Tolentino, L.K.S., Padilla, M.V.C., Madrigal, G.A.M., Enriquez, L.A.C.: Development Of Predictive Models Using Machine Learning Algorithms For Food Adulterants Bacteria Detection. In: 2019 Ieee 11th International Conference On Humanoid, Nanotechnology, Information Technology, Communication And Control, Environment, And Management (Hnicem), Pp. 1–6 (2019). Ieee. https://doi.org/10.1109/hnicem48295.2019.9072907

Boutorh, A., Guessoum, A.: Complex Diseases Snp Selection And Classification By Hybrid Association Rule Mining And Artificial Neural Network—Based Evolutionary Algorithms. Engineering Applications Of Artificial Intelligence 51, 58–70 (2016). https://doi.org/10.1016/j.engappai.2016.01.004

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M., Pascual-Montano, A.: Integrated Analysis Of Gene Expression By Association Rules Discovery. Bmc Bioinformatics 7(1), 1–16 (2006). https://doi.org/10.1186/1471-2105-7-54

Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., Gaglio, S., Urso, A.: Deep Learning Models For Bacteria Taxonomic Classification Of Metagenomic Data. Bmc Bioinformatics 19, 61–76 (2018). https://doi.org/10.1186/s12859-018-2182-6

Fung, F., Wang, H. S., & Menon, S. (2018). Food Safety In The 21st Century. Biomedical Journal, 41(2), 88-95. https://doi.org/10.1016/j.bj.2018.03.003

Hiura, S., Koseki, S., Koyama, K.: Prediction Of Population Behavior Of Listeria Monocytogenes In Food Using Machine Learning And A Microbial Growth And Survival Database. Scientific Reports 11(1), 10613 (2021). https://doi.org/10.1038/s41598-021-90164-z

Ibrahim, S. A., Ayivi, R. D., Zimmerman, T., Siddiqui, S. A., Altemimi, A. B., Fidan, H., ... & Bakhshayesh, R. V. (2021). Lactic Acid Bacteria As Antimicrobial Agents: Food Safety And Microbial Food Spoilage Prevention. Foods, 10(12), 3131. https://doi.org/10.3390/foods10123131

Ince, V., Bader-El-Den, M., Alderton, J., Arabikhan, F., Sari, O. F., & Sansom, A. (2025). Machine Learning-Based Prediction Of Clostridium Growth In Pork Meat Using Explainable Artificial Intelligence. Journal Of Food Science And Technology, 1-14. https://doi.org/10.1007/s13197-024-06187-7

Kyrpides, N.C., Eloe-Fadrosh, E.A., Ivanova, N.N.: Microbiome Data Science: Understanding Our Microbial Planet. Trends In Microbiology 24(6), 425–427 (2016). https://doi.org/10.1016/j.tim.2016.02.011

Lorenzo, J.M., Munekata, P.E., Dominguez, R., Pateiro, M., Saraiva, J.A., Franco, D.: Main Groups Of Microorganisms Of Relevance For Food Safety And Stability: General Aspects And Overall Description. In: Innovative Technologies For Food Preservation, Pp. 53–107. Elsevier (2018). https://doi.org/10.1016/b978-0-12-811031-7.00003-0

Liu, M., Ye, Y., Jiang, J., & Yang, K. (2021). Maniea: A Microbial Association Network Inference Method Based On Improved Eclat Association Rule Mining Algorithm. Bioinformatics, 37(20), 3569-3578. https://doi.org/10.1093/bioinformatics/btab241

Mamidala, S. (2023). The Sled (Shelf Life Expiration Date) Tracking System: Using Machine Learning Algorithms To Combat Food Waste And Food Borne Illnesses. Arxiv Preprint Arxiv:2309.02598. https://doi.org/10.48550/arXiv.2309.02598

Mitchell, T.M.: Machine Learning (1997). https://doi.org/10.1007/bfb0023942

Naulaerts, S., Moens, S., Engelen, K., Berghe, W.V., Goethals, B., Laukens, K., Meysman, P.: Practical Approaches For Mining Frequent Patterns In Molecular Datasets. Bioinformatics And Biology Insights 10, 38419 (2016). https://doi.org/10.4137/bbi.s38419

Ong, H.F., Mustapha, N., Hamdan, H., Rosli, R., Mustapha, A.: Informative Top-K Class Associative Rule For Cancer Biomarker Discovery On Microarray Data. Expert Systems With Applications 146, 113169 (2020). https://doi.org/10.1016/j.eswa.2019.113169

Saboe, D., Ghasemi, H., Gao, M.M., Samardzic, M., Hristovski, K.D., Boscovic, D., Burge, S.R., Burge, R.G., Hoffman, D.A.: Real-Time Monitoring And Prediction Of Water Quality Parameters And Algae Concentrations Using Microbial Potentiometric Sensor Signals And Machine Learning Tools. Science Of The Total Environment 764, 142876 (2021). https://doi.org/10.1016/j.scitotenv.2020.142876

Santos, A., Aerle, R., Barrientos, L., Martinez-Urtaza, J.: Computational Methods For 16s Metabarcoding Studies Using Nanopore Sequencing Data. Computational And Structural Biotechnology Journal 18, 296–305 (2020). https://doi.org/10.1016/j.csbj.2020.01.005

Sari, O. F., Bader-El-Den, M., Ince, V., & Arabikhan, F. (2024, August). Machine Learning Approach Into Bacterial Relationship: Exploring 16s Rrna Metabarcoding With Association Rule Mining. In 2024 Ieee 12th International Conference On Intelligent Systems (Is) (Pp. 1-6). Ieee. https://doi.org/10.1109/IS61756.2024.10705245

Stavropoulou, E., Bezirtzoglou, E.: Predictive Modeling Of Microbial Behavior In Food. Foods 8(12), 654 (2019). https://doi.org/10.3390/foods8120654

Sonwani, E., Bansal, U., Alroobaea, R., Baqasah, A. M., & Hedabou, M. (2022). An Artificial Intelligence Approach Toward Food Spoilage Detection And Analysis. Frontiers In Public Health, 9, 816226. https://doi.org/10.3389/fpubh.2021.816226

Sundui, B., Ramirez Calderon, O.A., Abdeldayem, O.M., Lázaro-Gil, J., Rene, E.R., Sambuu, U.: Applications Of Machine Learning Algorithms For Biological Wastewater Treatment: Updates And Perspectives. Clean Technologies And Environmental Policy 23, 127–143 (2021). https://doi.org/10.1007/s10098-020-01993-x

Taiwo, O. R., Onyeaka, H., Oladipo, E. K., Oloke, J. K., & Chukwugozie, D. C. (2024). Advancements In Predictive Microbiology: Integrating New Technologies For Efficient Food Safety Models. International Journal Of Microbiology, 2024(1), 6612162. Https://Doi.Org/10.1155/2024/6612162

Tandon, D., Haque, M.M., Mande, S.S.: Inferring Intra-Community Microbial Interaction Patterns From Metagenomic Datasets Using Associative Rule Mining Techniques. Plos One 11(4), E0154493 (2016). https://doi.org/10.1371/journal.pone.0154493

Yang, B., Wang, Y., Qian, P.-Y.: Sensitivity And Correlation Of Hypervariable Regions In 16s Rrna Genes In Phylogenetic Analysis. Bmc Bioinformatics 17(1), 1–8 (2016). https://doi.org/10.1186/s12859-016-0992-y

Yoon, Y., Lee, G.G.: Subcellular Localization Prediction Through Boosting Association Rules. Ieee/Acm Transactions On Computational Biology And Bioinformatics 9(2), 609–618 (2011). https://doi.org/10.1109/tcbb.2011.131

Zhao, Q., Bhowmick, S.S.: Association Rule Mining: A Survey. Nanyang Technological University, Singapore 135 (2003). https://doi.org/10.14453/isngi2013.proc.44

Zhou, C., Meysman, P., Cule, B., Laukens, K., Goethals, B.: Mining Spatially Cohesive Itemsets In Protein Molecular Structures. In: Proceedings Of The 12th International Workshop On Data Mining In Bioinformatics, Pp. 42–50 (2013). https://doi.org/10.1145/2500863.2500871