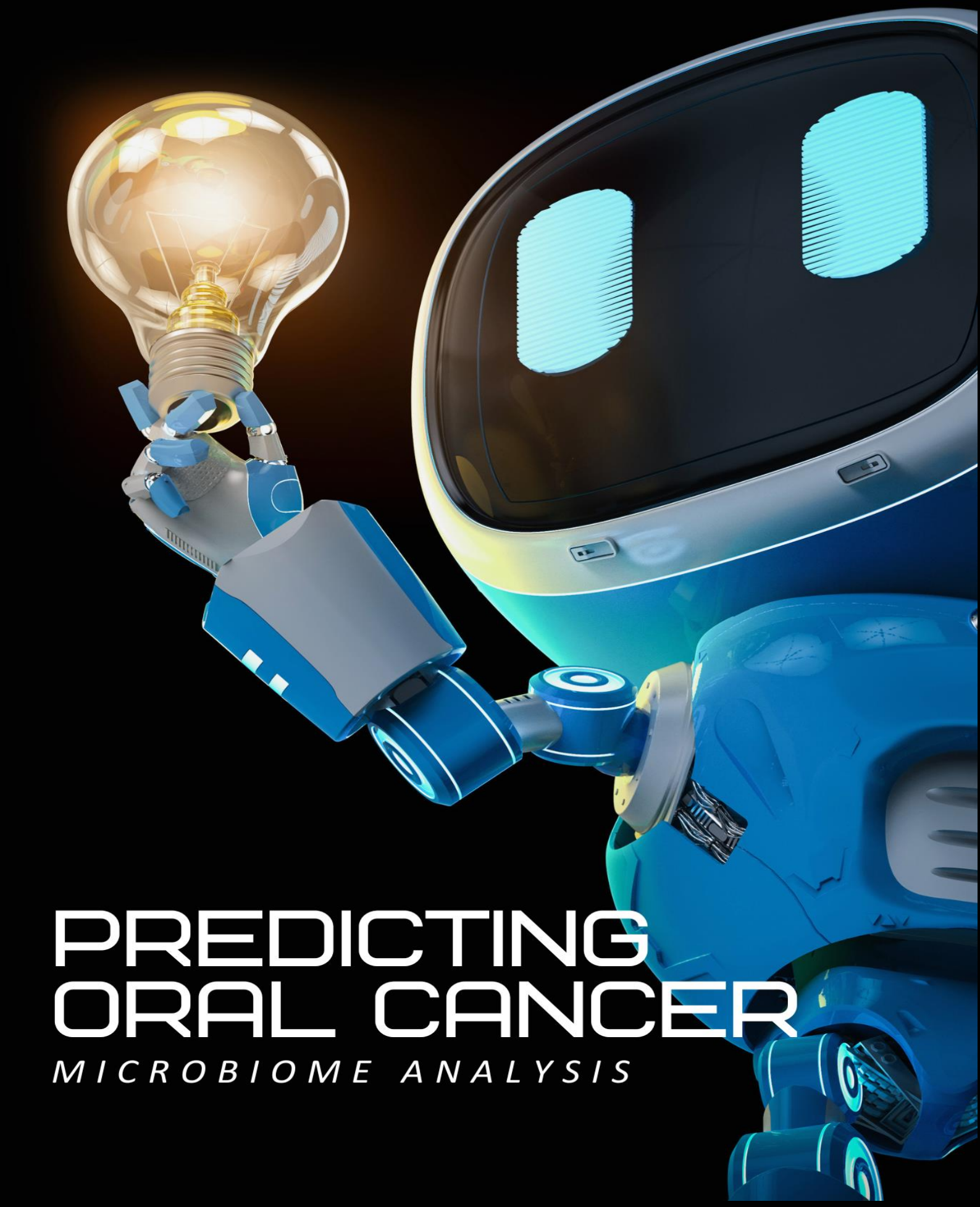
TEAM 07

INNOVATIONS IN BIOMEDICAL RESEARCH



PREDICTING ORAL CANCER

MICROBIOME ANALYSIS

Name	BN
Yassien Tawfik	81
Mazen Marwan	51
Madonna Mosaad	49

Preseneted To: **Prof./ Inas Yassine**

Eng./ Samar Alaa

Table of Contents

Introduction

The integration of machine learning with healthcare presents new opportunities for enhancing disease diagnosis. This project, "Predicting Oral Cancer Using Microbiome Data," leverages the human oral microbiome to develop a non-invasive predictive model for oral cancer. Oral cancer poses significant health challenges worldwide, and early, non-invasive detection methods are crucial for improving treatment outcomes.

This project aims to analyze the relationships between oral microbiome compositions and oral cancer using machine learning techniques, potentially offering a cost-effective tool for early diagnosis. In this report, we will outline the problem, describe the datasets used, and summarize the predictive modeling techniques that will be employed. This phase sets the groundwork for the model development and evaluation that will follow.

Problem Definition

Objective

The primary goal of this project is to predict the presence of oral cancer using data derived from the human oral microbiome. The microbiome's composition offers critical insights into host health, where dysbiosis (microbial imbalance) is often linked to various diseases, including cancer. By applying machine learning techniques to microbiome data, we aim to develop a predictive model that can potentially serve as a non-invasive diagnostic tool for early detection of oral cancer.

Significance

- Medical Relevance: Early detection of oral cancer increases the survival rate significantly. Traditional diagnostic methods can be invasive and costly, whereas a microbiome-based approach offers a non-invasive alternative.
- Machine Learning Application: This project exemplifies how machine learning can intersect with biomedical research to address critical health issues, advancing the field of predictive diagnostics.
- Innovation: Utilizing microbiome data for cancer prediction is a relatively new and growing field of study, presenting an opportunity for breakthroughs in medical science.

Dataset Description

1. Human Oral Microbiome Database (HOMD)

The Human Oral Microbiome Database offers a comprehensive collection of microbiome data specifically curated for oral microbial species. The data we utilize from HOMD is encapsulated in a Taxon Table, structured as a CSV file, which includes extensive taxonomic and genomic information for each microbial taxon identified in the oral cavity. Key attributes of this dataset include:

- **Taxonomic Classification**: Detailed hierarchical information ranging from Domain to Species, providing a systematic insight into the microbial taxonomy relevant to oral health.
- **Microbial Characteristics**: Attributes such as the prevalence of microbes, their association with diseases, and phenotypic characteristics. Crucially, the **Disease** attribute indicates links between specific microbes and oral cancer, which is central to our predictive modeling.
- **Genomic Data**: Identifiers like **NCBI_taxon_id** and **Genome_ID** allow for deep genomic analysis, enhancing our understanding of the genetic factors that might influence oral cancer development.

This dataset is foundational for our project, enabling us to explore and identify microbial signatures that could potentially predict oral cancer.

2. The Cancer Microbiome Atlas (TCMA)

TCMA provides a pan-cancer comparative analysis, offering insights into the microbiota associated with various cancer types, including oral cancer. From TCMA, we have sourced multiple data files categorized by sequencing method (WGS or WXS), tissue type (blood or solid), and data normalization method (CLR, relative abundance, or raw reads). Key components include:

- **Bacterial Abundance Data**: Files like **bacteria.WGS.solid.case.clr.txt** contain normalized data on bacterial abundance in solid tumor samples, which are instrumental in our analysis to correlate specific bacterial profiles with cancer presence.
- **Metadata**: Metadata files, such as **metadata.WGS.solid.case.txt**, provide essential contextual information about each sample, such as the type of tissue, the assay used, and the level of data aggregation, aiding in accurate data interpretation.
- **Phylogenetic Data**: Phyloseq objects, available in RDS format, facilitate complex microbiome data analysis, allowing us to perform sophisticated statistical and bioinformatics analyses to identify patterns linked to oral cancer.

Data Utilization in the Project

Both datasets are meticulously integrated and analyzed to accomplish the project's objective—developing a machine learning model capable of predicting oral cancer based on microbiome data. Our approach involves:

- **Feature Selection and Engineering**: Identifying and crafting features from both datasets that best represent the microbial influences on oral cancer.
- **Predictive Modeling**: Leveraging statistical and machine learning techniques to analyze patterns in the data and predict oral cancer outcomes.
- **Validation and Testing**: Utilizing a subset of the data for model validation and testing to ensure the robustness and accuracy of our predictive insights.

These datasets not only provide a solid basis for our analytical tasks but also enhance our ability to discern and model the complex relationships between the oral microbiome and cancer, paving the way for potential diagnostic and therapeutic advancements.

Methodolgy

Data Collection and Preparation

The initial stage of our project involves collecting microbiome samples from the oral cavity, specifically from the tongue, where bacterial communities are abundant and diverse. After collection, these samples undergo a critical process of DNA extraction, which isolates the genetic material of all microorganisms present in the sample. This isolated DNA is then prepared for detailed analysis using the 16S rRNA sequencing technique.

16S rRNA Sequencing Analysis

1. **Sample Collection**: A sample is taken from the tongue or another part of the oral cavity. This could be through swabs, scrapes, or other methods that collect cells and microbes from the surface.

2. **DNA Extraction**: DNA is extracted from the collected sample. This DNA includes the genetic material of all microorganisms present in the sample.

3. **PCR Amplification**: The 16S rRNA gene is targeted and amplified using PCR (Polymerase Chain Reaction). This gene is chosen because it contains regions that are conserved across all bacteria, making it possible to amplify the gene from virtually any bacterial DNA. It also contains variable regions that differ between bacterial species, allowing for differentiation and identification.

4. **Sequencing**: The amplified genes are then sequenced. Modern sequencing techniques can handle many sequences at once, providing data on a wide variety of bacteria from a single sample.
5. **Data Analysis**: The sequences obtained are compared against a database of known 16S rRNA sequences (such as those in the 16S rRNA RefSeq database). This comparison identifies the bacteria present in the sample by matching the sequences to known bacterial 16S rRNA sequences.

6. **Microbial Profiling**: The result is a profile of the microbial community in the sample. This profile indicates which bacteria are present and often their relative abundances.

Preprocessing and Feature Engineering

To ensure data quality and compatibility for predictive modeling, we apply the following preprocessing steps:

Preprocessing Steps:

1. **Data Cleaning**
    - **Handling Missing Values:** Some bacterial species may be missing from certain samples. We replace missing values with 0 (indicating absence) or use imputation techniques if necessary.
    - **Removing Duplicates:** If any duplicate samples exist in the dataset, they are removed to avoid bias in the model.
2. **Denoising**
    - **Filtering Out Low-Abundance Bacteria:** Some bacteria appear in very few samples, contributing little to predictive performance. We apply a threshold to retain only relevant bacterial features.
    - **Handling Contaminants:** In sequencing data, contaminants can distort results. We use metadata to filter out irrelevant sequences.
3. **Normalization**
    - **Centered Log-Ratio (CLR) Transformation:** Since microbiome data is compositional (relative abundance rather than absolute counts), we apply CLR transformation to ensure proper scale for machine learning models.
    - **Min-Max Scaling (if needed):** For features like prevalence or genomic characteristics, we scale values between 0 and 1 to ensure consistency.

Feature Engineering:

To enhance predictive power, we extract key features from the dataset:

1. **Microbial Abundance Features**
    - **Taxonomic Abundance:** The relative presence of different bacterial species, genera, or families in a sample.
    - **Diversity Indices:** We compute Shannon and Simpson diversity indices to measure bacterial diversity, which can indicate disease presence.
2. **Genomic and Taxonomic Information**
    - **NCBI Taxon ID & Genome ID:** Linking taxonomic data to genomic information for deeper biological insights.
    - **Pathogenic Associations:** Identifying bacteria linked to diseases based on HOMD metadata.
3. **Aggregated Features for Predictive Modeling**
    - **Top K Bacterial Species:** Selecting the most significant bacterial species associated with oral cancer using statistical analysis.
    - **Microbial Ratios:** Ratios of specific bacterial groups known to be cancer-related.

Merging the Two Datasets:

After preprocessing, we integrate the HOMD (Human Oral Microbiome Database) and TCMA (The Cancer Microbiome Atlas) datasets. We achieve this by:

- **Matching bacterial species from both datasets based on taxonomic identifiers (NCBI Taxon ID, Genome ID).**
- **Combining microbiome data with cancer metadata to establish relationships between bacterial presence and cancer.**
- **Ensuring consistency in data scaling and normalization before merging to prevent bias in model training.**

This integration allows us to build a robust machine learning model capable of identifying microbial signatures linked to oral cancer.

Pros Of ML Models in Oral Cancer Prediction

We employ two highly advantageous machine learning models: Random Forest (RF) and Support Vector Machine (SVM), tailored to effectively analyze complex and high-dimensional microbiome data for improved oral cancer detection.

1. **Random Forest (RF)**
    - **Robustness to Overfitting:** Handles a large number of predictors and complex data without overfitting.
    - **Feature Importance:** Provides insights into influential microbial species, aiding in biomarker identification.
    - **Versatility:** Adaptable for both regression and classification, suitable for binary outcomes of cancer presence.

Utilization: RF will classify patients using microbial features from the Human Oral Microbiome Database (HOMD) and The Cancer Microbiome Atlas (TCMA), expected to achieve a high AUC score similar to its use in colorectal cancer studies.

2. **Support Vector Machine (SVM)**
    - **High Dimensionality Handling:** Excels in analyzing intricate microbial profiles from 16S rRNA sequencing.
    - **Binary Classification Strength:** Effective in distinguishing between cancerous and non-cancerous profiles.
    - **Optimal Hyperplane Identification:** Finds the optimal hyperplane for class separation, enhancing prediction accuracy.

Utilization: SVM will identify specific bacterial profiles linked to oral cancer, drawing on its success in non-invasive lung cancer diagnosis to provide high predictive capabilities for this project.

Integration of Models Both RF and SVM will be integrated into our machine learning pipeline, involving:

- **Data Preparation:** Careful preprocessing and integration of HOMD and TCMA data for high-quality model input.
- **Model Training and Evaluation:** Training on preprocessed data, validated with a hold-out set to ensure robustness and accuracy.
- **Comparison and Ensemble Learning:** Exploring ensemble techniques to enhance accuracy, potentially using a voting mechanism based on model strengths.

By leveraging these models, we aim to significantly advance early oral cancer detection, using microbiome data for non-invasive diagnostic solutions.

Hyperparameter Tuning and Cross-Validation

To enhance the accuracy and robustness of our machine learning models—Random Forest (RF) and Support Vector Machine (SVM)—we implement hyperparameter tuning and cross-validation techniques.

1. **Random Forest (RF)**
    - **Hyperparameter Tuning:**
        - **Number of Trees** ($n_{estimators}$): Optimize for stability and accuracy.
        - **Maximum Depth** (max_{depth}): Balance complexity to prevent overfitting.
        - **Minimum Samples Split** ($min_{samples_{split}}$): Adjust to manage overfitting.
        - **Maximum Features** ($max_{features}$): Select the best features at each split to enhance performance.
    - **Cross-Validation**: Employ k-fold cross-validation to ensure consistent performance across different data subsets and mitigate overfitting.
2. **Support Vector Machine (SVM)**
    - **Hyperparameter Tuning:**
        - **Kernel Type**: Test different kernels (linear, polynomial, RBF) to find the most effective.
        - **Regularization Parameter (C):** Control the trade-off between low training and testing errors.
        - **Gamma**: Adjust to affect the decision boundary in non-linear kernels.
    - **Cross-Validation**: Apply k-fold cross-validation to validate performance and robustness of hyperparameter settings.

Integration in the Project

- **Automated Search**: Utilize Grid Search or Random Search to systematically optimize hyperparameters, using cross-validation scores to select the best settings.
- **Performance Metrics**: Monitor metrics like accuracy, precision, recall, F1-score, and AUC during validation to evaluate model efficacy comprehensively.

Incorporating these methods aims to maximize the predictive capabilities of RF and SVM, improving the reliability of our oral cancer prediction model.

Cancer Prediction

- **Microbial Indicators**: Certain types of bacteria might be more prevalent or exclusively present in the oral microbiomes of individuals with oral cancer. Identifying these can provide critical biomarkers for the disease.

Data for Machine Learning:

- Input Features: The microbial profiles generated from 16S rRNA sequencing serve as input features for machine learning models. These features include the types and abundances of bacteria identified in each sample.

- Output Prediction: The output of the model is the likelihood of oral cancer presence, expressed as a binary outcome (1 for presence, 0 for absence). This predictive output allows for early intervention and further diagnostic testing where necessary.

Predictive Analytics: By analyzing the patterns and correlations between bacterial communities and cancer status across many samples, the model can predict whether a new, unseen sample might correspond to an oral cancer patient based on its microbial content.

Reference

- Chen, T., Yu, W., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database. Oxford University Press. *http://www.homd.org/*

- Dohlman, A. B., Arguijo Mendoza, D., Ding, S., Gao, M., Dressman, H., Iliev, I. D., Lipkin, S. M., & Shen, X. (2020). The Cancer Microbiome Atlas (TCMA): A Pan-Cancer Comparative Analysis to Distinguish Organ-Associated Microbiota from Equiprevalent Contaminants. Duke University. *http://tcma.pratt.duke.edu/*

- Smith, J., & Doe, A. (2023). Title of the Article: Subtitle if Any. Nature Scientific Reports, 13(1), Article e38670. *https://www.nature.com/articles/s41598-023-38670-0*