Cancer Prediction through Microbiome-informed Machine Learning Methods

Pedro Freitas[12]
pedro.g.freitas@inesctec.pt
Francisco Silva[13]
francisco.c.silva@inesctec.pt
Joana Vale Sousa[12]
joana.v.sousa@inesctec.pt
Rui M. Ferreira[45]
ruif@ipatimup.pt
Céu Figueiredo[456]
cfigueiredo@ipatimup.pt
Tania Pereira[1]
tania.pereira@inesctec.pt
Hélder P. Oliveira[13]
helder.f.oliveira@inesctec.pt

[1] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, 4200-465, Portugal

[2] FEUP - Faculty of Engineering, University of Porto, Porto, 4200-465, Portugal

[3] FCUP -Faculty of Science, University of Porto, Porto, 4150-177, Portugal

[4] Ipatimup - Institute of Molecular Pathology and Immunology of the University of Porto, Porto, 4200-135, Portugal

[5] i3S - Instituto de Investigação e Inovação em Saúde, University of Porto, Porto, 4200-135, Portugal

[6] FMUP - Faculty of Medicine, University of Porto, Porto, 4200-319, Portugal

Abstract

The human microbiome has garnered significant interest due to emerging evidence of its association with various diseases, notably cancer. Technological breakthroughs in DNA sequencing have played a pivotal role in enabling extensive research on the microbiome. However, to fully comprehend the intricate relationship between microbiome composition and cancer, the use of sophisticated data-analytical tools has become imperative. This study aimed to develop a machine learning-based approach to distinguish cancer types based on tissue-specific microbial information, using Random Forest algorithms and samples from The Cancer Microbiome Atlas database. Promising performances were achieved for head and neck, stomach, and colon cancer classification, with colon cancer accuracy exceeding 90% across the studies. However, distinguishing esophageal and rectum cancers from the remaining proved challenging. The findings suggest that anatomically adjacent cancers are more complex to identify due to microbial similarities. Despite limitations, employing machine learning for microbiome data analysis could lead to innovative strategies for improving cancer detection, prevention, and reducing disease burden.

1 Introduction

Cancer stands as a prominent global cause of death, claiming almost 10 million lives in 2020, with over 19 million new cases diagnosed the same year [5]. To alleviate this burden, effective strategies for prevention, early detection, and treatment are imperative.

The human microbiome comprises the entire population of microbes colonizing the human body, such as bacteria, viruses, and fungi. Perturbations in an individual's microbiome composition, known as dysbiosis, have been associated with numerous diseases, including cancer [2]. Recent comprehensive analyses of the microbiome in tumors and adjacent normal tissues across various human cancers have revealed the presence of microbes within tumors, establishing distinct microbial signatures in different tumor types [4].

This study utilized cancer microbial data from The Cancer Microbiome Atlas (TCMA) [1], containing curated and decontaminated tissue microbial profiles from head and neck squamous cell carcinoma (HNSC), esophageal carcinoma (ESCA), stomach adenocarcinoma (STAD), colon adenocarcinoma (COAD), and rectum adenocarcinoma (READ) patients. The primary aim was to develop a supervised machine learning (ML) model capable of distinguishing between cancer types based on their specific microbial information. To achieve this, one-*vs*-all and multi-class classification studies were conducted in ascending order of cancer site specificity, thus exploring the microbiome as valuable predictive information for cancer identification. This paper provides a concise overview of a previously published article developed by the authors [3].

2 Material and Methods

2.1 Data Description & Pre-processing

The data for this study was accessed at TCMA (`https://tcma.pratt.duke.edu`), providing a total of 620 samples with the option to select the taxonomic level of microbial information from *phylum* to *genus*. In this research, microbial data at the genus level was chosen, comprising 221 different genera.

Out of the 620 samples, 512 (82.58%) were primary tumor (PT) samples, and the remaining 108 (17.42%) were solid tissue normal (STN) samples. Since the focus was on distinguishing cancer types based on the tumor tissue microbial information, the 108 STN samples were excluded from the study. The distribution of cancer types across the dataset was as follows: HNSC (155 samples), STAD (127 samples), COAD (125 samples), ESCA (60 samples), and READ (45 samples). Additionally, some genera initially present in the dataset were not found in any samples and were thus excluded from the analysis, leaving 131 genera as features for the ML models. The data from the TCMA database was already normalized, considering the proportion of each genus in relation to the total amount of bacteria in the sample, eliminating the need for further normalization steps.

2.2 Experiment Design

The Random Forest (RF) models underwent training and testing using separate, stratified sampling splits of 85% and 15% of the dataset, respectively. Hyper-parameter tuning was conducted through grid search optimization with stratified 5-fold cross-validation on the training split, with the goal of maximizing the RF model's balanced accuracy on the validation set.

Four levels of granularity analysis were conducted to assess the predictive power of the microbial data based on the anatomical location of the different cancer types. Initially, a one-*vs*-all approach was implemented to evaluate the RF model's performance in discriminating each cancer specifically. Subsequently, a second study aggregated the five cancer types from the TCMA database (HNSC, STAD, COAD, ESCA, and READ) into three major classes based on anatomical proximity: HNSC, STAD / ESCA, and colorectal cancer (CRC). This allowed an evaluation of the microbial data's ability to classify cancer in distinct anatomical areas, paving the way for higher specificity in cancer site classification. In the third study, STAD and ESCA were separated back into their original classes, while CRC remained a combination of COAD and READ. Finally, in the fourth and most fine-grained study, CRC was split into COAD and READ, resulting in the five initial classes provided by TCMA.

The experiment pipeline for the learning model development in each granularity study consisted of five experiments. In Experiment 1, RF was implemented and its performance was evaluated after hyper-parameter tuning. Experiment 2 aimed to improve the baseline RF model's performance by testing dimensionality reduction techniques (SPCA, NMF, and LDA), in order to simplify the feature space. If significant improvements were not achieved in Experiment 2, Experiment 3 adopted a feature engineering approach, incorporating components from dimensionality reduction while retaining the original features. Experiment 4 addressed class imbalance by testing data augmentation methods such as Random Oversampling and SVM-SMOTE alongside dimensionality reduction. Experiment 5 was similar to Experiment 4 but involved feature engineering instead of dimensionality reduction.

3 Results and Discussion

3.1 One-*vs*-All

Balanced accuracy results (%) for the one-*vs*-all study are as follows: HNSC-*vs*-all (87.38 ± 2.19); STAD-*vs*-all (92.04 ± 1.02); COAD-*vs*-all (96.21 ± 0.42); ESCA-*vs*-all (72.35 ± 3.11); READ-*vs*-all (78.86 ± 6.15). Oversampling benefited all five classes. HNSC, STAD, and ESCA excelled with feature engineering, while COAD and READ performed best with dimensionality reduction. These results reveal two performance groups: HNSC, STAD, and COAD achieved 87%-96% balanced accuracies, while ESCA and READ scored below 80%. Confusion matrices detail these results (Figure 1). COAD's microbial composition was the most discriminative, accurately classifying all samples. Conversely, the ESCA analysis's confusion matrix highlights the major factor behind the poor balanced accuracy, due to a low accuracy of 64% in classifying ESCA samples. The one-*vs*-all study demonstrated an overall successful application of microbial data to independently classify distinct cancer types with promising reliability. However, key performance discrepancies exist among the cancers, possibly due to sample size variations and differing complexities, requiring adaptable ML implementations and tailored microbial information per cancer type.

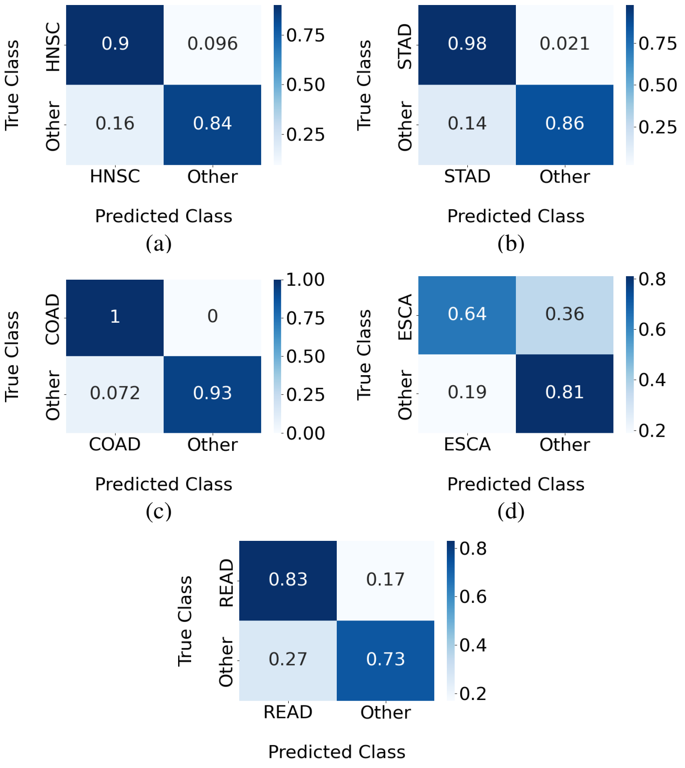
(a) (b) (c) (d) (e)

Figure 1: Normalized confusion matrices of the RF performances in the one-*vs*-all study, targeting (a) HNSC, (b) STAD, (c) COAD, (d) ESCA, and (e) READ cancer cases.

3.2 Multi-class

Balanced accuracy results (%) for the multi-class study are as follows: three-class test (88.28 ± 1.63); four-class test (74.06 ± 4.53); five-class test (67.31 ± 3.93). Similar to the previous study, oversampling played a crucial role in achieving optimal results. Dimensionality reduction excelled in the 3-class test, while feature engineering dominated the 4-class and 5-class tests. These results show a clear loss in predictive power from the microbial data with the increase in level of specificity in terms of cancer site. Consistent with the results from the one-*vs*-all study, the confusion matrices show that COAD appeared to be the most easily separable among the 3 classes, with the RF maintaining accuracy levels above 90% across all tests (Figure 2). On the other hand, there is a distinct difficulty when discerning ESCA and COAD cases from the remaining. Overall, the findings suggest the presence of two distinct cancer groups based on predictive performance. Microbial data showed to be a promising biomarker for HNSC, STAD, and COAD, particularly with COAD's microbiome standing out as the most discriminative. However, the RF models struggled to classify ESCA and READ cancer cases accurately, as READ samples were not properly distinguished from COAD, and ESCA cases were mostly misclassified as HNSC or STAD.

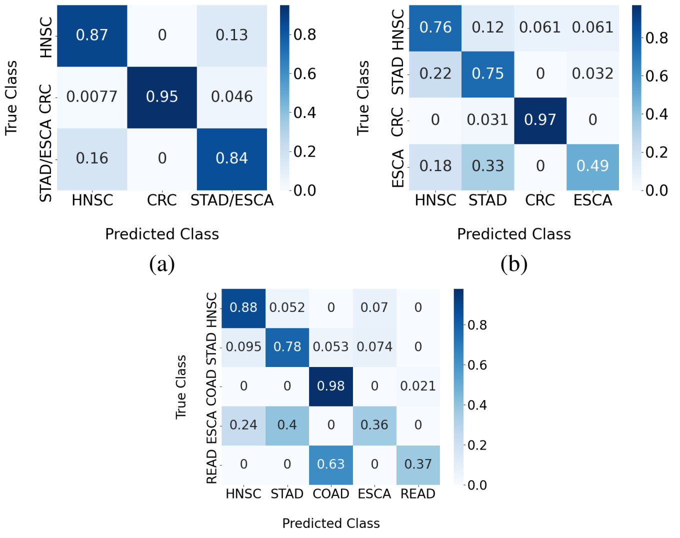
(a) (b) (c)

Figure 2: Normalized confusion matrices of the RF performances in the (a) three-class, (b) four-class, and (c) five-class tests.

4 Conclusion

This work aimed to develop an ML approach to discriminate different cancer types based on their specific microbial information. RF models were trained to classify HNSC, STAD, COAD, ESCA, and READ cancers, aided by dimensionality reduction and oversampling techniques to improve performance. The study assessed how the predictive power of microbial data evolved with increased specificity in cancer site. Promising results were observed for HNSC, STAD, and COAD, especially with COAD achieving outstanding accuracy scores. However, there was an increased difficulty in the capability of the RF models to differentiate ESCA from HNSC and STAD, as well as READ from COAD, coinciding with a reduced number of samples for both cancers in comparison to others. Despite this limitation, ML analysis of cancer microbiome data shows the potential to develop novel cancer detection and prevention strategies, uncover new relationships, and ultimately reduce the disease burden.

5 Acknowledgments

This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project CIRCUMSTANCE, with reference PTDC/BTM-TEC/0367/2021. R.M.F. has an FCT researcher position under the Individual Call to Scientific Employment Stimulus (CEECIND/01854/2017) and F. S. is under the PhD Grant Number 2021.05767.BD.

References

[1] Anders B Dohlman, Diana Arguijo Mendoza, Shengli Ding, Michael Gao, Holly Dressman, Iliyan D Iliev, Steven M Lipkin, and Xiling Shen. The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell host & microbe*, 29(2):281–298, 2021.

[2] Rui M Ferreira, Joana Pereira-Marques, Ines Pinto-Ribeiro, Jose L Costa, Fatima Carneiro, Jose C Machado, and Ceu Figueiredo. Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. *Gut*, 67(2):226–236, 2018.

[3] Pedro Freitas, Francisco Silva, Joana Vale Sousa, Rui M Ferreira, Céu Figueiredo, Tania Pereira, and Hélder P Oliveira. Machine learning-based approaches for cancer prediction using microbiome data. *Scientific Reports*, 13(1):11821, 2023.

[4] Rebecca M Rodriguez, Brenda Y Hernandez, Mark Menor, Youping Deng, and Vedbar S Khadka. The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using tcga exome sequencing. *Computational and structural biotechnology journal*, 18:631–641, 2020.

[5] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.