

Hidden fingerprints: microbiome signatures in colorectal, breast, and prostate cancer

Tsen Chen Chien

ABSTRACT

Early detection of colorectal, breast, and prostate cancer remains difficult, and microbial signatures could support affordable screening tools. We trained four machine learning models (SVM, Random Forest, XGBoost, and L1-Logistic Regression) on public microbiome data from all three cancers. The models separated the cancers with 99.3% cross-validation accuracy, but this largely reflects sample type (stool, tissue, urine) rather than cancer biology. Within a single organ, breast tumors showed a modest signal (AUC 0.63) and colorectal showed none (AUC 0.49). SHAP analysis identified the genera driving the predictions. Overall, the results suggest that microbial cancer signals are organ-dependent rather than universal.

INTRODUCTION

- The human microbiome (bacteria, fungi, and viruses) supports digestion and the immune system.
- In 2020, researchers found bacteria living inside tumor cells across several cancer types, including breast, which showed that tumor microbes are more than contamination (Nejman et al., 2020).
- Colorectal cancer has the best-documented link: *Fusobacterium nucleatum* is abundant in tumors and helps them spread while dampening immune defenses (Kim & Lee, 2022).
- Breast cancer research and prostate cancer has less research done on. Both suggests hormone driven or processes which feeds tumors and increase risk of cancer (Zhang et al., 2020).
- Each cancer has been studied on its own, and to our knowledge, no study has compared all three in one machine learning model with explainability tools.
- Early detection is hard and screening is costly, so microbial signatures could support affordable tools for diverse communities.
- Question: can microbial profiles distinguish colorectal, breast, and prostate cancer, and which microbes drive the differences?
 - Hypothesis: the model can identify cancer type from microbes alone, and SHAP will show which microbes drive it.

METHODOLOGY

1. Collect

Three NCBI SRA datasets, 618 samples: colorectal (n = 268, PRJNA290926), breast (n = 319, PRJNA658160), prostate (n = 31, PRJNA1298576)

2. Clean

Filter low-quality samples and rare genera, Harmonize microbes to genus level, CLR transform (Aitchison, 1982)

3. Explore

Alpha diversity, PCoA / UMAP, PERMANOVA (Anderson, 2001), Mann-Whitney U with FDR correction

4. Classify

Four models (SVM / Random Forest (Breiman, 2001) / XGBoost (Chen & Guestrin, 2016) / L1- Logistic Regression), 5-fold cross validation, SMOTE inside training folds (Chawla et al., 2002)

5. Explain

SHAP (Lundberg & Lee, 2017), 15-genus biomarker panel

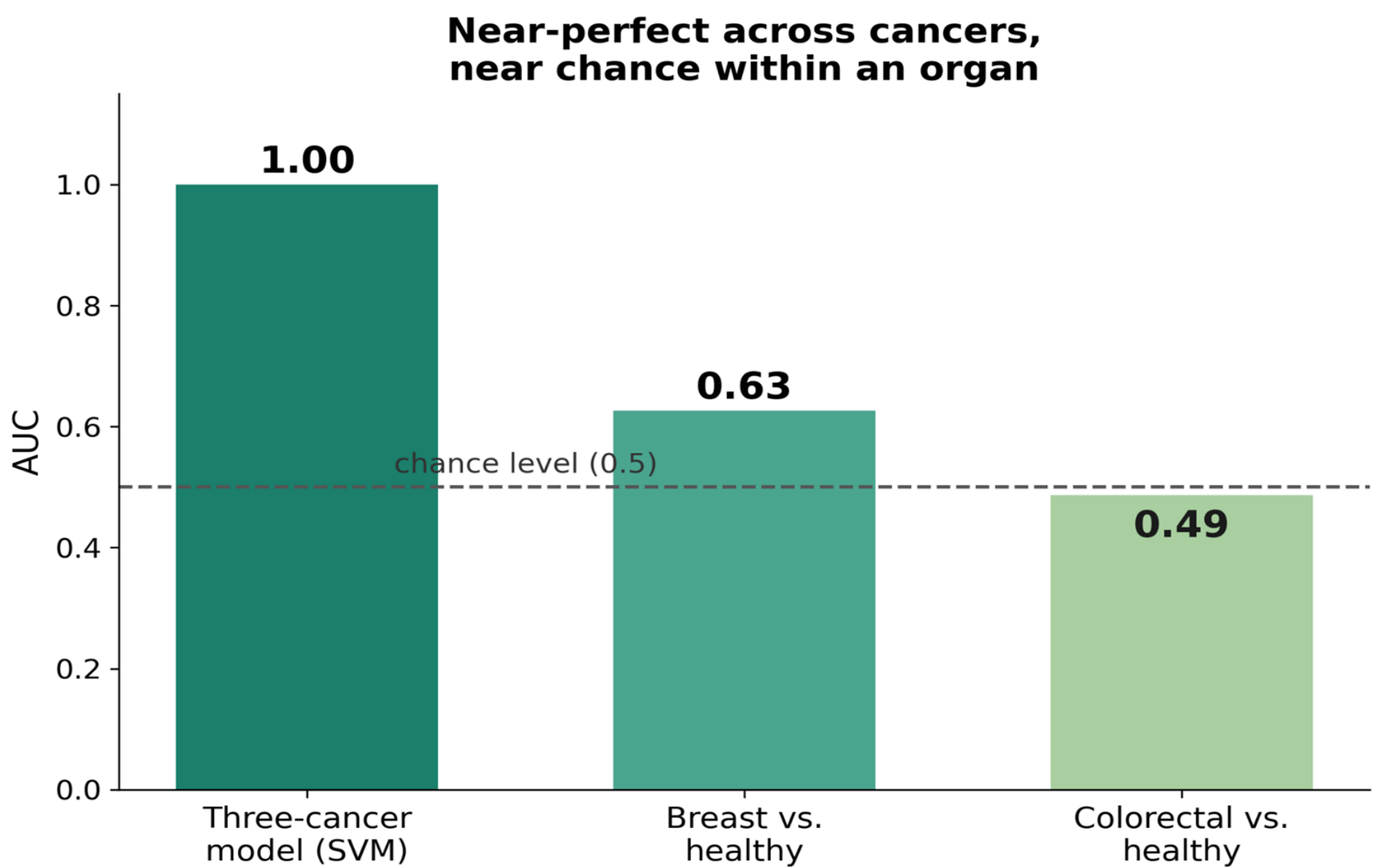
RESULTS

Figure 1: AUC Bar Chart

All four models reached 99.27% accuracy, but this mostly reflects sample type, not cancer biology.

Three-cancer model (SVM) scored near-perfect macro-AUC of 1.00 in cross-validation, but the score drops sharply within one organ (0.63 for breast and 0.49 for colorectal).

The 1.00 is a three-class cross-validation score; the breast and colorectal values come from held-out test sets.



RESULTS (CONTINUED)

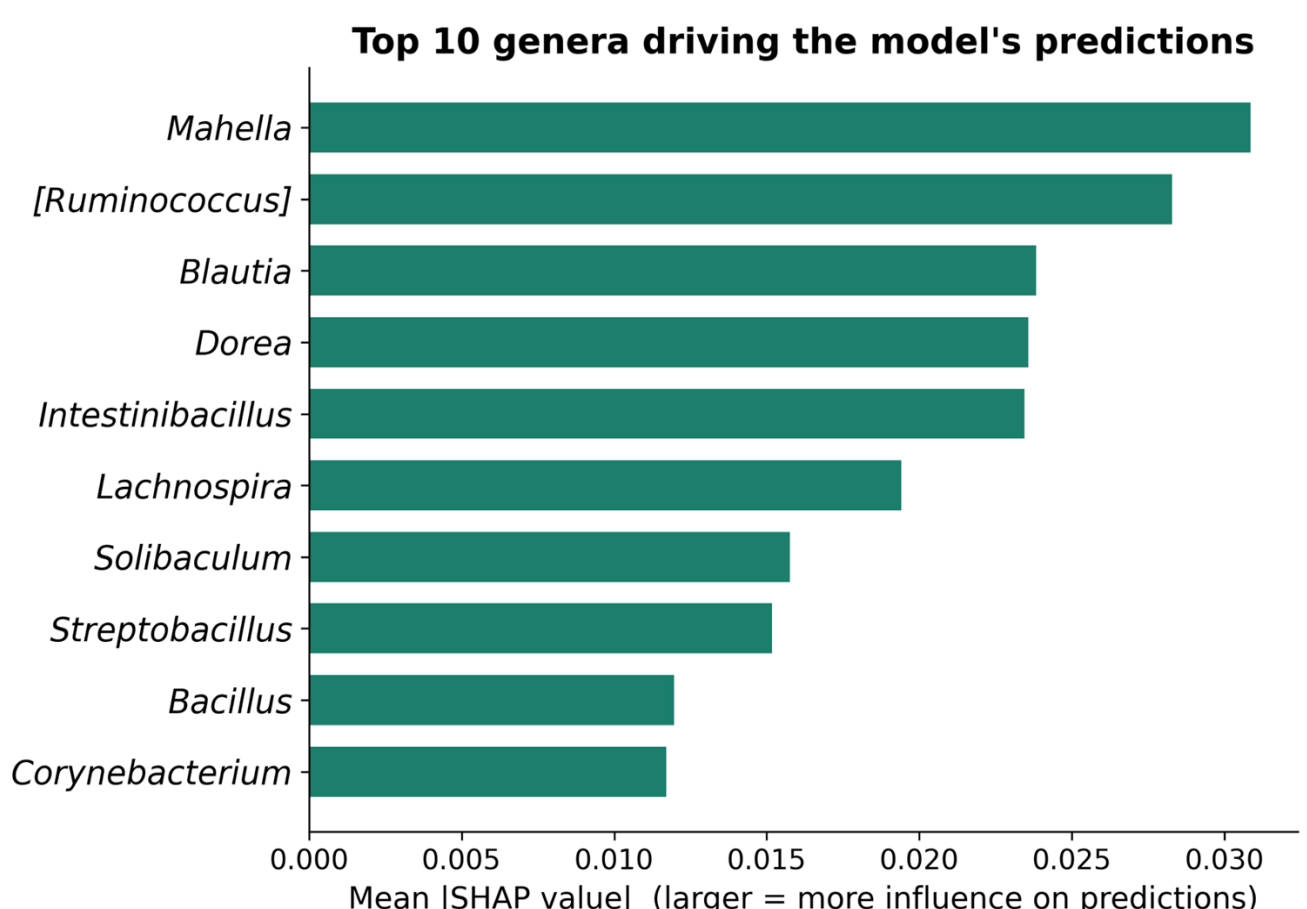


Figure 2: SHAP top 10

Top 10 genera driving the Random Forest's predictions (mean |SHAP value|, held-out test set). Longer bars mean greater influence. Many are common gut bacteria, consistent with the model mostly separating sample types.

Figure 3: ROC curves

Within one organ, only breast shows signal: AUC 0.63 for breast vs. 0.49 for colorectal.

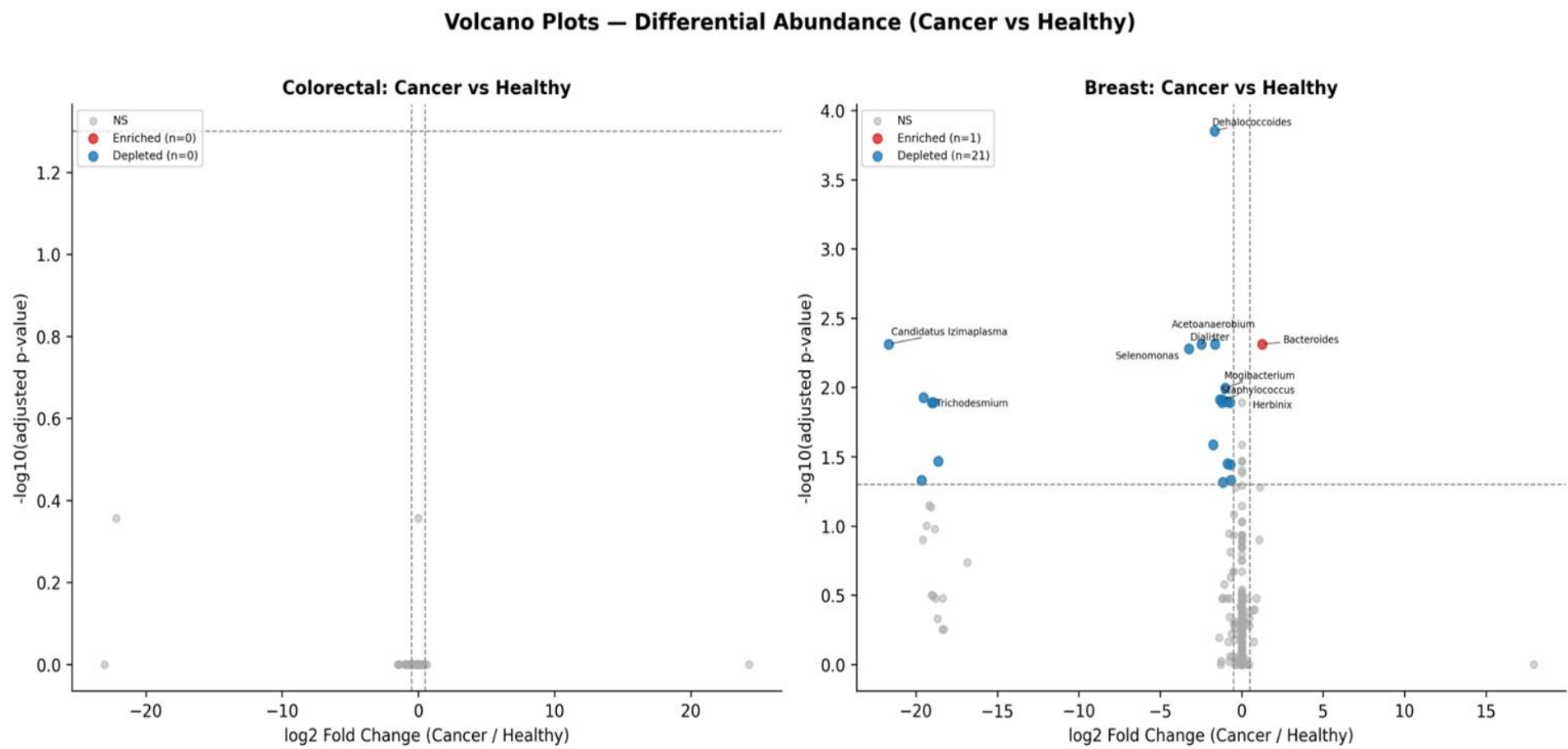
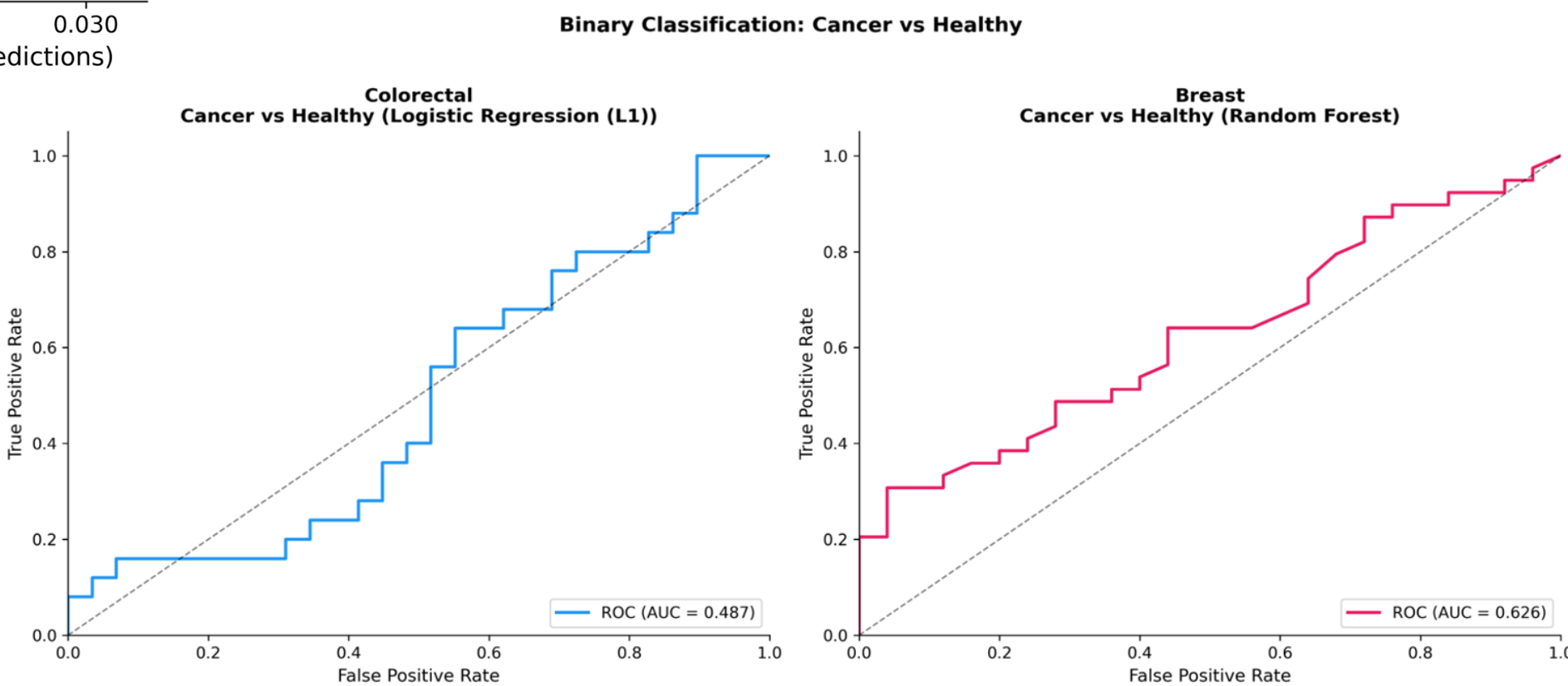


Figure 4: Volcano Plots

22 genera differ significantly in breast cancer (*Bacteroides* enriched); none do in colorectal.

DISCUSSION & CONCLUSION

The three-cancer model scored 99.27%, but this mostly reflects sample type (stool, tissue, urine), not cancer biology. SHAP highlights common gut genera (*Blautia*, *Dorea*, *Lachnospira*), consistent with the model separating body sites more than cancer.

Within one organ, signal was modest for breast (AUC 0.63) and absent for colorectal (AUC 0.49), so microbial cancer signals appear organ-dependent.

Limitations: A different sample type per cancer, a small prostate cohort (n = 31) with no healthy controls, and no independent validation.

Future research using larger, multi-center cohorts with consistent sample types is needed to determine whether biomarkers can be identified for specific cancer types.

REFERENCES

- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139–177. <http://www.jstor.org/stable/2345821>
- Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32–46. <https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Fujita, K., et al. (2023). The Gut-Prostate Axis: A New Perspective of Prostate Cancer Biology through the Gut Microbiome. *Cancers*, 15(5), Article 1375. <https://doi.org/10.3390/cancers15051375>
- Kim, J., & Lee, H. K. (2022). Potential role of the gut microbiome in colorectal cancer progression. *Frontiers in Immunology*, 12, Article 807648. <https://doi.org/10.3389/fimmu.2021.807648>
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv. <https://arxiv.org/abs/1705.07874>
- Nejman, D., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science*, 368(6494), 973–980. <https://doi.org/10.1126/science.aay9189>
- Zhang, J., et al. (2020). Breast and gut microbiome in health and cancer. *Genes & Diseases*. <https://doi.org/10.1016/j.gendis.2020.08.002>