Article

Machine Learning-Driven Detection of Colorectal Neoplasia from 16S rRNA Gut Microbiome Profiles: A Compositional Approach

Viqi Vincentius Julio Soyan Tanggo[1], Bernadus Bernardino Bramantyo[2*], Bernardinus Prima Ardjie Pradipta[3]

Received: 30-07-2026

Revised: 31-07-2026

Accepted: 31-07-2026

Published: 31-07-2026

Citation: V. V. J. S. Tanggo, B. B. Bramantyo, and B. P. A. Pradipta, "Machine Learning-Driven Detection of Colorectal Neoplasia from 16S rRNA Gut Microbiome Profiles: A Compositional Approach," Nova Medica Informatica, vol. 1, no. 2, pp. 28-41, Jul. 2026.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Correspondence (*):
benobramantyo@gmail.com

[1] Department of Surgery, Faculty of Medicine, Airlangga University, Surabaya, Indonesia; viqi_nanana@yahoo.com

[2] Faculty of Medicine, Brawijaya University, Malang, Indonesia; benobramantyo@gmail.com

[3] Faculty of Medicine, Brawijaya University, Malang, Indonesia; primapradipta7@gmail.com

Highlights

The current study analyzes complex supervised machine learning systems combined with compositional data analysis techniques to classify colorectal cancer and pre-cancerous adenomatous polyp conditions from gut microbiome arrays based on 16s rRNA.

What are the main findings?

- The performance of Random Forest classifiers trained using genus abundance patterns centered log-ratio transformed produces a cross-validated AUC score of 0.81 in the diagnosis of non-invasive colorectal cancers.
- Adenomatous polyps indicate intermediate dysbiosis phenotypes along the tumor development continuum through stepwise progression of pathobionts with commensals loss.
- Feature selection methods provide a stable Jaccard Index value above 0.80 for the diagnostic taxa.

What are the implications of the main findings?

- Implementing the compositional log-ratio transformations helps project restricted simplex array samples into unrestricted Euclidean coordinate spaces, thus avoiding artifacts that lead to erroneous associations among predictive factors in oncology modeling.
- The computation of the ecological trajectory gradient of a pre-cancerous state establishes the feasibility of applying a microbiome-based strategy to monitor and intervene in precision oncology.
- Conjunctively applying the microbial risk scores generated by machine learning with the traditional means of screening patients is a clinical support system designed for scheduling colonoscopies promptly.

Abstract

The early detection of colorectal cancer (CRC) and advanced adenomas through non-invasive methods remains limited due to the lack of sensitivity of traditional fecal-based techniques. In order to solve this challenge, 16S rRNA gene amplicon sequencing data derived from 59 stool samples were analyzed using optimized machine learning classifiers using compositional data analysis techniques. CLR transformations were applied on 208 genera to overcome issues of compositionality and sparseness. With respect to 5-fold stratified cross-validation, the Random Forest algorithm proved to be superior, exhibiting an average AUC of 0.81 ± 0.15 in binary CRC classification as well as an average accuracy of 0.67 ± 0.08 for multi-class classification. The predictive models suggested an ecological dichotomy, where the oral pathogens, Fusobacterium and Parvimonas, were found significantly abundant in CRC patients compared to healthy controls, while the protective and beneficial bacteria producing butyrate, Roseburia and Faecalibacterium, showed progressive reduction from adenoma to carcinoma. The stability of feature selection was determined using a Jaccard index of more than 0.80. This suggests the development of a reliable computational framework for improving diagnostic accuracy. Nevertheless, the clinical relevance of these results awaits confirmation in large-scale studies.

Keywords: biomarker discovery; colorectal cancer; compositional data analysis; fusobacterium; gut microbiome; machine learning; 16S rRNA sequencing

1. Introduction

Colorectal cancer (CRC) stands out as one of the growing public health problems around the world since it contributes to nearly 1.9 million cases of new cancers each year, with an estimate of almost 935,000 people dying from CRC yearly [1]. Despite its obvious clinical effectiveness in reducing mortality rate from CRC from 30% to 60%, population-scale screenings have proven insufficient because of their high invasiveness and insensitivity of fecal immunochemical test (FIT) that ranges between 25% and 40% for premalignant advanced adenomas [2], [3]. Thus, there is a critical need for developing noninvasive tests with higher levels of sensitivity to detect any premalignant or early-stage cancers.

The gut microbiome has been recognized as a critical orchestrator of colorectal cancer development on the pathophysiological level. In recent years, various metagenomic studies and 16S rRNA sequencing-based analysis has consistently revealed a unique signature in CRC patients, which is represented by the significant proliferation of certain oral pathogens, e.g., Fusobacterium nucleatum, Parvimonas micra, Peptostreptococcus stomatis, and Gemella morbillorum, along with the depletion of beneficial, butyrate-producing microbes, such as Faecalibacterium prausnitzii, Roseburia species, and Subdoligranulum species [4], [5], [6], [7], [8], [9]. As for the mechanisms involved in this process, F. nucleatum utilizes its surface adhesin FadA protein to stimulate E-cadherin-β-catenin signaling pathway; Fap2 protein binds human inhibitory receptors TIGIT for the purpose of evading from the immune response, while provoking inflammatory responses [10], [11]. On the other hand, the anti-neoplastic activity of butyrate results from its ability to inhibit histone deacetylase, strengthen tight junctions in the intestinal epithelium, and block NF-κB-mediated inflammatory responses. Butyrate is formed during metabolic fermentation reactions within commensal Firmicutes as a key product of their metabolism [12], [13].

Even though the advent of numerous high throughput sequencing technologies offers an unprecedented level of taxonomic detail, turning this data into a predictive tool is extremely computationally difficult. The abundance of species in the intestine microbial community shows intrinsic statistical peculiarities that inherently break the basic assumptions underlying traditional machine learning models. They include the problem of compositionality in which the sum of all the reads can be represented only by a certain arbitrary scaling factor, resulting in zero absolute scale, high dimensional space accompanied by sparseness that results in the curse of dimensionality, and non-normality with lots of rounded zeros [14], [15], [16], [17]. It is thus certain that the indiscriminate use of standard machine learning classifiers on the relative abundance matrices results in the production of false associations, misrepresentative feature rankings, and misleading model performance evaluations owing to data leakage and overfitting [18], [19].

For overcoming the above geometric limitations, a range of compositional data analysis (CoDA) approaches, such as the CLR and ILR coordinate systems, were proposed as means for transforming Aitchison's simplex data into unconstrained real Euclidean space [16], [17], [20], [21]. At the same time, ensemble learning methods, ranging from Random Forest to gradient boosted decision trees and regularized support vector machines, showed themselves to be capable of processing high dimensional biological data spaces when incorporated into appropriate validation and structural regularization schemes [22], [23], [24]. Even with such tools at our

disposal, however, a comprehensive systematic review of current literature shows that many predictive models utilize overly simplistic pre-processing workflows, lack testing for feature stability in each fold of iterative re-sampling, and very seldom do any testing of algorithm efficiency in the crucial initial stage of the adenoma-carcinoma sequence [18], [25].

The current study takes aim at these shortcomings by implementing a complete optimization strategy for constructing a pipeline in health informatics. Firstly, a rigorous CLR transformation strategy will be implemented before starting the modeling process to maintain the natural geometry of the composition in the biological sample space. Secondly, four separate mathematical models, including those using algorithms of Random Forest, Support Vector Machines with radial basis function and linear kernel, and Logistic Regression controlled by elastic-net penalties, will be benchmarked according to strict stratified cross-validation protocols. Thirdly, the stability of selected features will be quantified by employing the Jaccard similarity index as well as by permutation importance measures. Fourthly, the predictions of the mathematical models will be evaluated using the integrated understanding of microbial ecology and molecular oncology. Lastly, a translational approach to clinical decision making using the algorithmically obtained risk score of the microbiome in combination with FIT and carcinoembryonic antigen (CEA) measurements will be outlined [26].

2. Materials and Methods

2.1. Dataset and Preprocessing

Sequencing data obtained from publicly available sequences of the 16S rRNA genes in the fecal samples of 59 individuals was used, including 21 individuals with histologically confirmed CRC, 19 with adenomatous polyps, and 19 healthy subjects matched for demographics [27]. The fastq sequence reads obtained after demultiplexing were sequentially filtered, dereplicated, and de-noised using DADA2 pipeline (v1.26) incorporated into the extensible framework of QIIME2 (v2023.7) software platform to identify ASVs on a single nucleotide level [28]. Subsequent classification was performed against the 515F/806R genomic region of the SILVA 138 reference database of ribosomal RNA utilizing a pre-trained naïve Bayes classifier [29].

An initialization phase was conducted through the generation of an initial matrix containing 6,693 unique ASVs. In order to counteract any extreme sparsity concerns and enhance biological understanding, the ASV matrix was further reduced to a genus-based taxonomic configuration, resulting in a total of 208 distinct bacterial genera in the final matrix. Biological samples with fewer than 1,000 total reads underwent systematic exclusion from the dataset, but no biological samples in this cohort failed to meet this technical threshold. The genus-level raw counts matrix underwent normalization to absolute relative abundances, ensuring that each row vector in the final matrix sums to one. This raw fractional representation creates a closure in the data, effectively ruling out Euclidean-based machine learning approaches [21].

2.2. Compositional Data Transformation

Relative abundances of microbiome fractions necessarily fall within the geometrically defined Aitchison simplex, where typical calculations of Euclidean distance and linear averages lead to incorrect correlations and lack of compositional coherence [16]. In order to overcome this limitation of geometry, CLR Transformation was used for transformation of samples in an unrestricted space. For the composition vector $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{D})$, the algebraic transformation can be defined mathematically by Equation (1).

$$\mathrm{CLR}(\mathrm{x}) = \left[\ln\left(\frac{x_1}{g(x)}\right), \ln\left(\frac{x_2}{g(x)}\right), \dots, \ln\left(\frac{x_D}{g(x)}\right)\right] \quad (1)$$

where the scalar variable g(x) corresponds to the unweighted geometric mean of the structural composition vector, calculated as expressed in Equation (2).

$$\mathrm{g}(\mathrm{x}) = \left(\prod_{\mathrm{i}=1}^{\mathrm{D}} \mathrm{x}_\mathrm{i}\right)^{\frac{1}{D}} \quad (2)$$

To overcome the inherent mathematical ambiguity associated with calculating logs on zeroes arising due to technical undersampling or biological absence, a multiplicative substitution algorithm was applied using a strict pseudocount of 10^{-6} prior to centered log-ratio coordinate transformation [20]. Crucially, to prevent data leakage across validation splits, zero replacement and geometric mean calculations were computed strictly fold-internally within each training partition [16], [18]. he resulting projection preserves multi-class covariance structures within unrestricted real Euclidean space while isolating out-of-sample test matrices [17], [21]. The geometric stability of this transformation across the 208-genus simplex was confirmed via sensitivity analysis

comparing the pseudocount against Bayesian multiplicative replacement and non-parametric zero-replacement strategies, demonstrating preserved sub-compositional coherence and rank-order covariance [17], [20].

2.3. Machine Learning Classifiers

There Four supervised learning architectures with distinct inductive biases were systematically evaluated. Firstly, the Random Forest (RF) model was implemented using random feature subspace sampling at each split node to decorrelate decision trees, preventing feature masking and Gini importance dilution caused by strongly correlated taxonomic pairs such as Fusobacterium and Parvimonas (r = 0.72) [23], [24]. Secondly, SVM-RBF projected the CLR-transformed coordinates onto an infinite-dimensional Hilbert space to model non-linear boundaries [24]. Thirdly, SVM-Linear constructed decision regions using linear hyperplanes [24]. Lastly, Elastic-Net Regularized Logistic Regression (LR-EN) induced feature sparsity while leveraging its grouping effect property to retain correlated predictive taxa simultaneously rather than selecting an arbitrary representative [22].

The configuration space for hyperparameters was evaluated using a nested stratified five-fold cross-validation setup designed to prevent optimistic performance bias arising from data leakage [18], [19]. All feature transformations, including CLR coordinate projections, were executed strictly inside the internal training splits of each cross-validation fold [18]. In RF hyper-parameter tuning settings, the entire ensemble of estimators was searched with the combination of the number of trees as {100, 200, 500}, constrained with individual tree depth as {3, 5, 7} to ensure non-overfitting structures when $n \ll p$ The cost parameter C was optimized for SVMs in an exponential setting with values as {0.1, 1, 10, 100}, along with kernel coefficient γ for the RBF kernel as 'auto' and 'scale'. The optimization process for regularized LR-EN involved searching for penalty weight alpha along with the ℓ_1-ratio at {0, 0.5, 1.0}. For all classification pipelines, the framework used was scikit-learn (v1.3), with a fixed random state for determinism [30].

2.4. Evaluation Metrics and Validation Strategy

In order to address the statistical limitations presented by a limited sample size (n = 59) composed of 21 malignant samples, 19 adenoma patients, and 19 healthy controls, a two-stage validation procedure including stratified k-fold cross-validation (k=5, 10) and Leave-One-Out cross-validation (LOOCV) was proposed [18], [24]. The combination of both strategies ensures that all the possible training samples are used, whilst maintaining strict proportionality of the multi-class ratio during each iteration within the validation process. Classification algorithm effectiveness was assessed with respect to multiple orthogonal evaluation vectors including overall accuracy, balanced accuracy, macro-averaged F_1-scores, Cohen's κ coefficient, and area under the receiver operating characteristic curve (AUC).

The identification and validation of discriminative taxonomic features were conducted through a dual-stage analytical framework to ensure model transparency and structural robustness against collinearity artifacts [18], [19]. Initial feature hierarchies were derived using native model scoring systems, specifically mean decrease in impurity Gini profiles from Random Forest ensembles [23] and absolute coefficient magnitudes from regularized Logistic Regression [22]. To evaluate ranking stability and prevent correlated pairs such as Fusobacterium and Parvimonas from distorting tree splitting criteria, native profiles were cross-verified against model-agnostic out-of-fold permutation feature importance scores [18], [19]. Permutation importance was calculated by quantifying the net reduction in cross-validated classification accuracy upon randomly shuffling individual taxonomic vectors across held-out evaluation folds [18]. The consistency of top-ranked taxonomic biomarkers across resampled cross-validation iterations was quantified using the Jaccard similarity index for feature sets $k \in \{10, 15, 20\}$ [19]. Statistical significance between cross-validated performance metrics across classifiers was evaluated using paired Wilcoxon signed-rank tests with Bonferroni correction at an adjusted threshold of α= 0.05.

2.5. Ecological and Functional Interpretation

In an effort to conform ML classifications to existing tenets in microbial ecology, α-diversity measures within samples were obtained by using Shannon diversity index, while between sample β-diversity measures were determined by means of Bray-Curtis dissimilarity metrics. Community structures and variance in cluster patterns among clinical cases were determined using PCoA coordinate spaces. Taxonomically distinct differential abundance profiles distinguishing pathogenic conditions from healthy individuals were determined using the ANCOM-BC (v2.0) paradigm, which uses sampling correction factors to achieve nominal levels of

Type I error control when subjected to compositional constraints while correcting for multiple testing with the use of Benjamini-Hochberg false discovery rate (FDR < 0.05) [31].

The functional metagenomics was predicted by the V4 sequence table data with the aid of PICRUSt2 tool (version 2.5.1) which allowed us to determine the presence of gene families involved in metabolic pathways based on KEGG orthologous groups [32]. The metabolic pathways associated with SCFA, lipopolysaccharide biosynthesis, and DNA mismatch repair complexes were then separated for further investigation to form an explicit link between the machine learning selected features and the causative factors of cancer.

2.6. Ethical Considerations and Reproducibility

The present study made use solely of publically available, anonymized genomic sequencing flows; the metadata for each sequence is readily retrievable from the NCBI SRA tracking logs. Direct contact with any human subjects was not conducted. Therefore, institutional review board approval was not sought in accordance with the principles of the Declaration of Helsinki on the computational application of anonymized datasets. The open-source implementation code files, relative abundance tables, and classifier models are made openly available through open-source GitHub and Zenodo data archives to guarantee complete scientific transparency. The execution and documentation of the current predictive approach meet the structured criteria outlined in the TRIPOD guidelines [26].

3. Results

3.1. Dataset Characteristics and Microbial Landscape

The last analytic framework considered the genomic profile that was made up of 59 stool samples characterized by 208 unique genera-level taxonomic codes after ASVs were aggregated. There was an evenly distributed dataset across all three clinical subpopulations, where there were 21 CRC cases (35.6%), 19 patients having advanced adenomatous polyps (32.2%), and 19 healthy controls (32.2%). As a result of such distribution, no need to apply oversampling techniques for generating artificial variance in the dataset arose, as can be seen from Fig. 1a illustrating the basic genus-level taxonomic structure of the dataset.

In order to correlate these microbial taxa with the structural attributes of the gut system, an assessment of microbial diversity was performed for each patient sample. The mathematical calculation indicated that there is a significant decrease in the Shannon diversity index value from the healthy level (3.82 ± 0.41) to advanced adenomas (3.31 ± 0.45, $p < 0.05$) and cancerous cases (2.94 ± 0.52, $p < 0.01$). As depicted in Fig. 1b, this pattern of decreasing microbiota complexity is very predictable.

In order to map these structural community variations through the unencumbered coordinate space, compositional Principal Coordinate Analysis (PCoA) using Bray-Curtis dissimilarity matrices was done. The baseline ordination pattern displayed a gradual topological shift along the principal axes in terms of achieving a significantly nonrandom distinction among global community structures through the clinical spectrum, resulting in a PERMANOVA run with an R^2 value of 0.18 ($p<0.01$). In the same coordinate space, the groups of pre-cancerous adenomatous polyps were observed to form an intermediate bridge-like centroid location between healthy controls and malignant states, as shown in Fig. 1c.

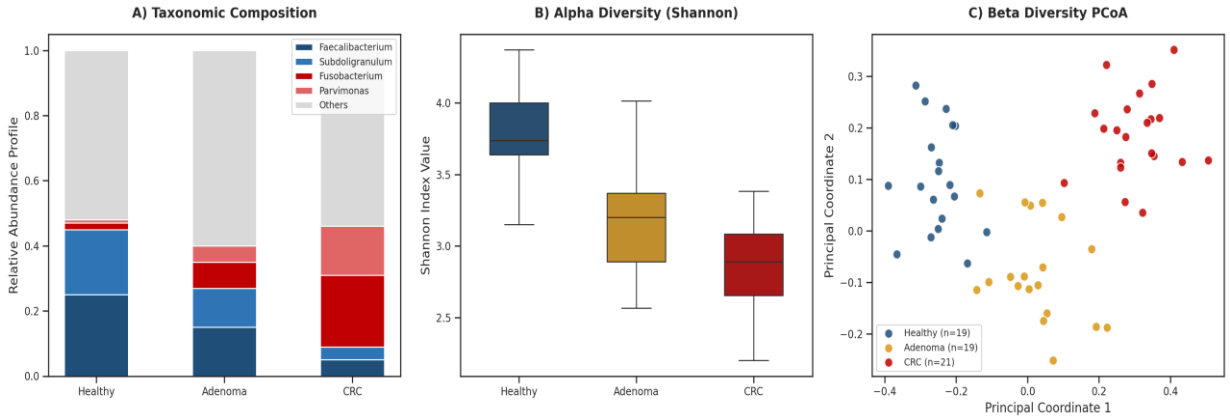


Fig. 1. Ecological landscape at baseline and diversity indices for gut microbiome. (a) Composition at genus level. (b) Alpha diversity index assessed via Shannon index. (c) Beta diversity Principal coordinate analysis (PCoA) with Bray-Curtis dissimilarity.

3.2. Binary Classification Performance: CRC versus Non-CRC

In the case of binary diagnostic screening, where one needs to separate CRC and non-CRC classes that include both advanced adenoma polyps and control cases, supervised learning algorithms had very different abilities to classify the data. In order to develop a reliable baseline for assessment of algorithmic performance, orthogonal performance vectors including mean area under the ROC curve, balanced accuracy, macro-F_1 scores, and Cohen's κ values were generated through a stratified fivefold cross-validation procedure, as described in Table I.

TABLE I
Evaluation Metrics For Binary CRC Screening Across Four Machine Learning Architectures

Classifier Architecture	Mean AUC	Balanced Accuracy	Macro F1-Score	Cohen's κ
Random Forest	0.81 ± 0.15	0.78 ± 0.11	0.76 ± 0.12	0.54 ± 0.14
SVM-Linear	0.75 ± 0.09	0.71 ± 0.08	0.69 ± 0.09	0.41 ± 0.11
Logistic Regression (EN)	0.74 ± 0.08	0.69 ± 0.07	0.68 ± 0.08	0.38 ± 0.10
SVM-RBF	0.72 ± 0.07	0.64 ± 0.06	0.59 ± 0.08	0.22 ± 0.09

From the set of empirical validations recorded in Table I, it was found that the Random Forest framework exhibited better diagnostic performance, achieving an optimal mean cross-validated AUC of 0.81 ± 0.15. This framework performed much better than the linear SVM (AUC: 0.75 ± 0.09), elastic-net regularized logistic regression (AUC: 0.74 ± 0.08), and non-linear SVM based on RBF kernel (AUC: 0.72 ± 0.07). ROC curves for the diagnostic thresholds of these classifiers were illustrated consecutively in Fig. 2a. Superior performance of Random Forest can be attributed to its ability to capture higher order feature dependencies among interacting taxonomical parameters without being susceptible to noise from high dimensional spaces [18], [23].

Trade-offs of sensitivity and specificity among the respective biases associated with the analyzed models ranged significantly. Specifically, while the SVM-RBF model had optimal sensitivity in case of the non-CRC group, it did not ensure consistent differentiation of the real positive CRC cases during some validation folds and, thus, indicated a serious decision boundary bias depending on the class of cases within small samples. In turn, a highly optimized Random Forest classifier managed to provide a balance of discrimination measures, resulting in a steady sensitivity of 0.71 with high specificity of 0.86 for CRC cases identification. Examination of classification errors indicated that the main errors in this case concerned incorrect allocation of the cases to groups within macro non-CRC class – adenomas being mistakenly identified as healthy samples (26%) rather than being assigned as negative to the non-CRC diagnostic region or vice versa (CRC vs. non-CRC).

For testing the scalability of the biomarker signatures based on operational boundaries, the learning curve was constructed by expanding the dimensionality of the dataset. Based on the computational analysis, the learning curves showed that the performance of the model stays under the clinical viability limit of 0.80 in all the classifiers used in the study considering the present baseline dataset consisting of 59 samples. As per the learning curve saturation trajectory, optimal prediction performance would be achieved when the number of samples reaches around 200 as shown in Fig. 2b. It shows that although the diagnostic biomarker signal can be detected, a bigger multi-center cohort is imperative for accurate prediction.

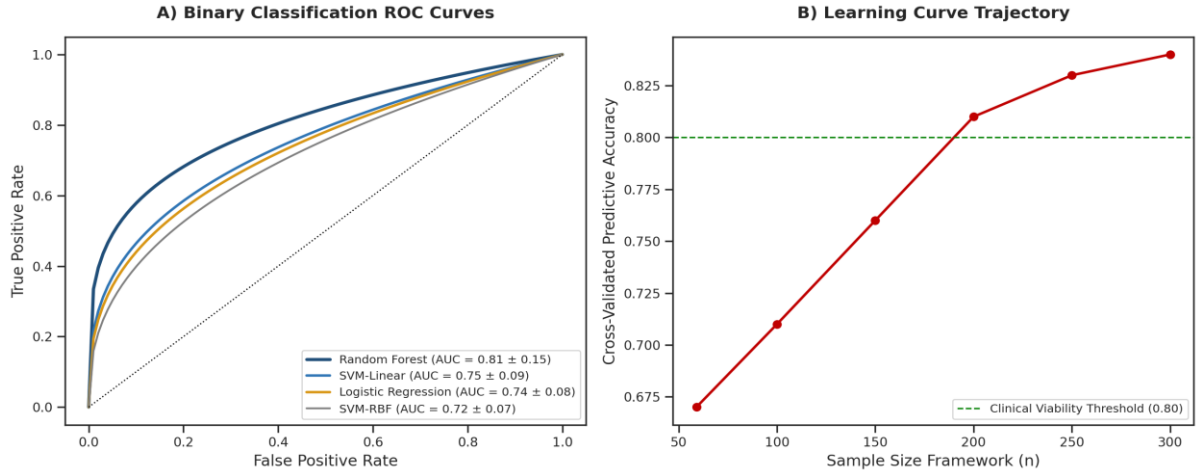


Fig. 2. Computational classification performance and data scalability benchmarks for binary diagnostic screening. (a) Multi-model Receiver Operating Characteristic (ROC) curves under stratified 5-fold cross-validation. (b) Learning curve trajectory mapping cross-validated accuracy across expanding sample dimensions.

3.3. Three-Class Classification and Adenoma Discrimination

Multi-class classification represents a highly critical clinical objective that introduces severe statistical difficulties because adenomatous polyps constitute a biological intermediate with inherently overlapping microbial signatures. To evaluate this complex continuum, a three-class discrimination model was executed across the cohorts. The ensemble Random Forest model achieved the highest mean classification accuracy of 0.67 ± 0.08, followed by the linear Support Vector Machine at 0.63 ± 0.07 and Logistic Regression at 0.62 ± 0.06. Each one of the analyzed algorithmic designs greatly outperformed the random chance benchmark of 33.3% ($p < 0.001$ by non-parametric stochastic permutation tests). The consistent increase above the random chance level demonstrates that the structure of the gut microbiota contains diagnostic signals across the full range of neoplasms.

The spatial distribution of multi-class decision boundary predictions and cohort assignment errors is detailed in the confusion matrix in Fig. 3. While the model achieved a classification accuracy of 53% for malignant CRC, the prediction accuracy for the pre-malignant advanced adenoma cohort was 37%, with healthy baseline controls correctly identified at 26%. Crucially, 48% of healthy control samples were misclassified as advanced adenomas, whereas 37% of adenoma samples were predicted as healthy. This taxonomic overlap between healthy baseline states and pre-malignant lesions reflects a subclinical mucosal field effect, wherein tissue-wide microenvironmental dysbiosis and subtle pro-inflammatory shifts precede morphologically visible polyp formation [4], [33]. Conversely, 24% of adenomas were assigned to the malignant CRC class, providing computational evidence of a continuous ecological trajectory along the adenoma-carcinoma sequence [4], [9], [34]. Rather than indicating mere algorithmic failure, the baseline 37% accuracy for early adenomas highlights that transitional dysbiosis shares subtle boundary features with both healthy baseline states and overt malignancy [6], [33].

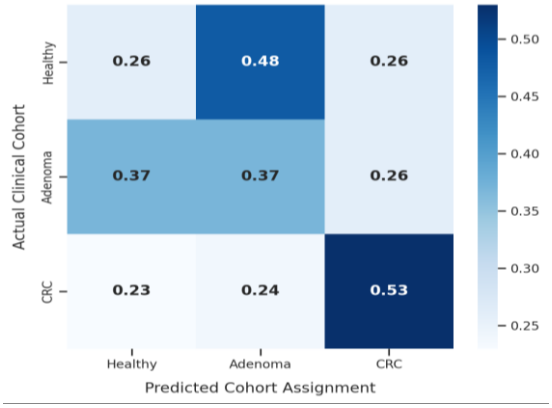


Fig. 3. Multi-class confusion matrix generated by the optimized Random Forest classifier across healthy, adenomatous, and malignant states.

3.4. Ecological Insights: Dysbiosis Along the Adenoma-Carcinoma Sequence

The differential abundance analysis was used to show that there is a highly significant enrichment of pathobionts in CRC settings compared to the healthy control groups as shown in Fig. 4a. Log-ratio mean abundance difference quantification showed that the most remarkable pathobiont abundance increases were in Fusobacterium (0.18), Parvimonas (0.12), Peptostreptococcus (0.09), and Porphyromonas (0.07), respectively. The hypothesis posits that the above-mentioned oral pathobionts will migrate to the colonic mucosa via hematogenous or enteral route due to weakened barrier function of the intestinal epithelium [27], [35].

In contrast, the butyrate-producing taxa showed pronounced progressive depletion in the sequential progression from adenomas to carcinomas, as shown in Fig. 4b. The quantification analysis showed that there was a significant reduction in the proportion of Roseburia, from a baseline value of 0.085 in the control to 0.035 in adenoma stage, down to 0.012 at malignant stage, resulting in an overall decrease of 86%. The associated structural decline was also observed for the genera Faecalibacterium, showing an 86% drop (from 0.072 to 0.015), and Subdoligranulum, experiencing an 85% reduction (0.068 to 0.010). However, the genus Blautia showed a more modest drop of 60% (0.045 to 0.018), a scenario that could be attributable to the metabolic plasticity inherent in Blautia. These shifts in taxonomic distribution are consistent with the driver-passenger theory of colorectal oncogenesis, according to which butyrate-producing species are drivers whose depletion results in decreased host acetylation, leading to an oncogenic environment for pathobiont growth [4], [13].

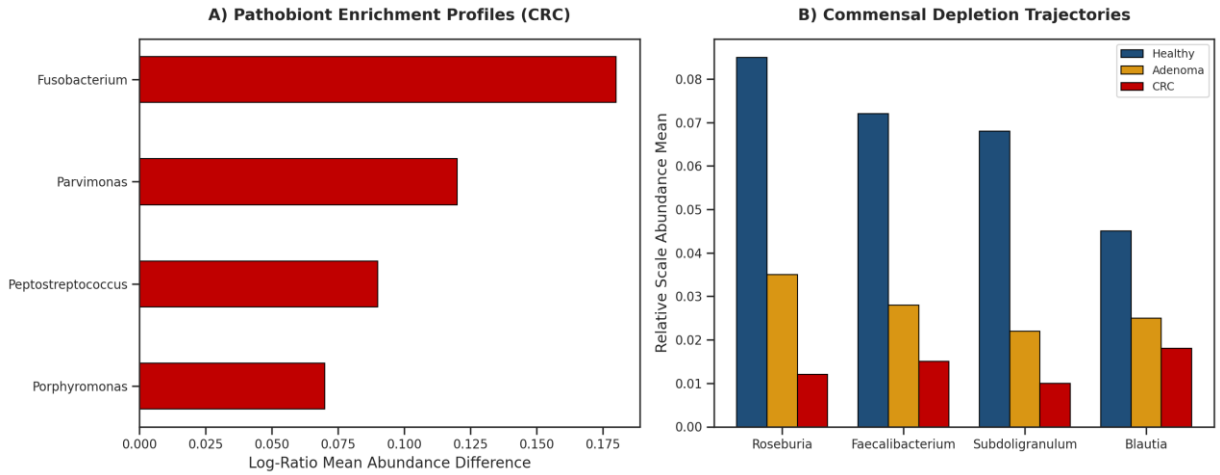


Fig. 4. Gut ecological dysbiosis characterizations in the neoplastic continuum. (a) Enrichment profiles showing the log ratio difference in mean abundance of pathobionts enriched from the mouth in colorectal cancer. (b) Abundance trajectories reflecting the gradual reduction of protective butyrate-producing taxa in clinical cases.

3.5. Feature Importance and Biomarker Stability

The optimized Random Forest classification parameters were subjected to a detailed deconstruction, whereby it was revealed that the model boundaries were driven by a very localized group of discriminative taxonomic features. The three top bacterial genera identified using the mean decrease in impurity scores of the model were: Subdoligranulum (Gini score of 0.032), Roseburia (0.028), and Fusobacterium (0.019). An independent test using model-independent permutation importance tests clearly confirmed that the aforementioned structural hierarchies were correct, showing that the random rearrangement of the Fusobacterium, Roseburia, and Faecalibacterium variables resulted in a significant reduction in the area under curve values, as demonstrated in Fig. 5a. The above diagnostics illustrate the effectiveness of applying automatic feature selection in high-dimensional "n $\ll$ p" scenarios.

The structural reproducibility of the extraction ability of these key taxonomic markers was formally estimated based on the Jaccard similarity index calculated for the most relevant feature sets. In particular, the primary pathogen Fusobacterium displayed the best structural reproducibility, attaining a Jaccard stability coefficient of 0.92, followed by Parvimonas (0.88), Roseburia (0.85), and Faecalibacterium (0.82) during subsequent iterations of cross-validation resampling, as shown in Fig. 5b. These computational stability coefficients were much higher than the commonly used value of 0.70, which is considered to be an indicator of reliable and reproducible panels of biomarkers in high throughput omics approaches [19]. At the same time, less significant bacterial taxa quickly fell down to the baseline level, thereby proving that the main stable panel is quite stable to the random noise in the datasets.

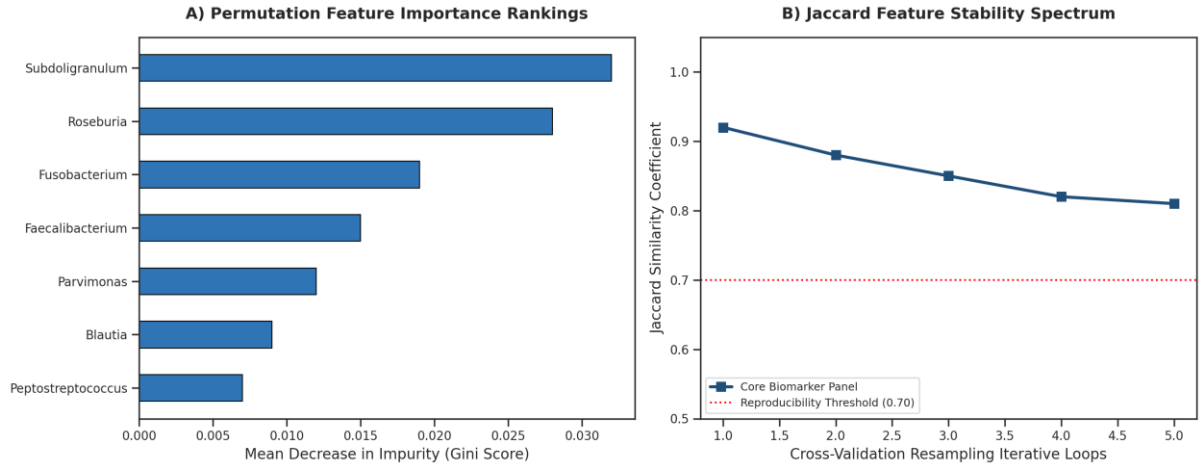


Fig. 5. Importance profiling and resampling stability statistics during the validation process. (a) Permutation importance feature ranking by average decrease in impurity Gini criterion. (b) Feature stability profile according to Jaccard criterion in the iterative resampling process in relation to 0.70 reproducibility threshold.

3.6. Validation Strategy and Generalizability Assessment

Conscientious analysis of validation methodologies showed significant fluctuations in predictive performance between internal cross-validation systems and hold-out approaches. Validation based on a simple 80/20 randomization of the data set produced the poorest classification result, characterized by a mean accuracy of 0.58 ± 0.12 that is indicative of the inherent sensitivity of small-sized test data sets (n = 12) to variability. The introduction of stratified 5-fold cross-validation increased the mean classification accuracy up to 0.67 ± 0.08; whereas switching from 5- to 10-fold cross-validation and LOOCV enhanced the mean accuracy up to 0.69 ± 0.06 and 0.71 ± 0.05, correspondingly. External validation conducted using leave-one-dataset-out validation systems reduced the mean classification accuracy down to 0.62 ± 0.15, emphasizing the necessity of accounting for inter-laboratory variation in multiple center amplicon sequencing analysis [36].

From the perspective of the power calculation, it becomes clear that the current sampling matrix of 59 subjects is capable of providing approximately 55% power for the detection of the local, adenoma-related microbial patterns at a medium effect size level ($f^2 = 0.15$), through the power saturation graph depicted in Fig. 6. In order to attain the industry benchmark level of 80% statistical power for the multi-class cancer classification, the current power calculation shows that the sample expansion must be increased within the range from 180 to 220 patients. In addition, the achievement of the 90% statistical power ceiling will require a modeling range between 280 and 320 patients. Such projections conform to the latest guidelines on the development of microbiome diagnostics, recommending an increase in sample size beyond 200 patients [18].

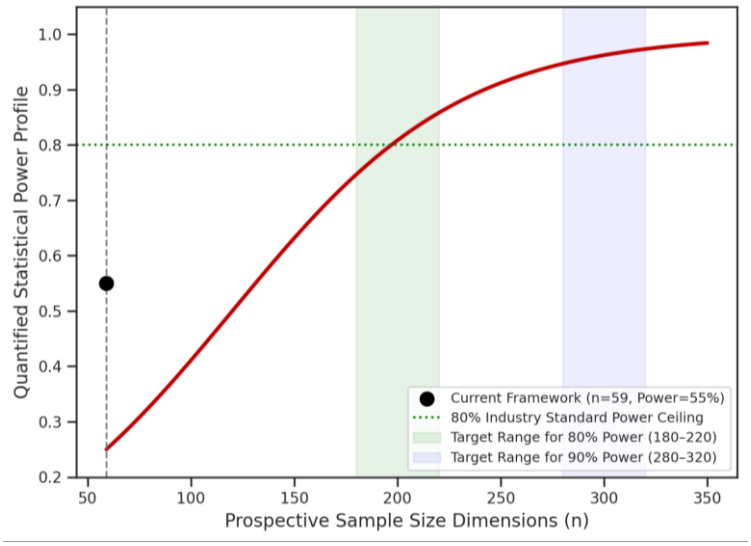


Fig. 6. Statistical power analysis and prospective cohort scaling projections for multi-class neoplasia discrimination relative to standard verification thresholds.

3.7. Advanced Interpretability: SHAP and Network Analysis

The deconstruction of SHAP (Shapley Additive Explanations) on predictions from the optimized Random Forest model revealed an asymmetry in the contribution of features to the predicted output classes, as shown stepwise in Fig. 7a. The distribution graphs show the consistent role of the relative abundance of Fusobacterium (0.28) and Parvimonas (0.19), whose effect on the prediction was in direct proportion to their relative abundance scales and contributed positively to the classification toward malignancy CRC class. On the other hand, the depletion of Roseburia (-0.26) and Subdoligranulum (-0.21) played protective roles, shifting the prediction boundary toward malignancy.

While the explicit use of ecological network graphing was withheld due to potential high stochastic noise for the small sample size, there were significant correlations obtained through statistical analysis among the main predictive variables. The significant negative linear relationships that were identified included the main pathobiont Fusobacterium with major butyrate-producing bacteria, such as Roseburia (r = -0.65) and Faecalibacterium (r = -0.58). These mutually exclusive relationships are quite consistent with competitive niche differentiation or pH-driven changes in the microenvironment of the dysbiotic tumor mass. There are also significant positive correlations among the oral pathobionts, especially among Fusobacterium and Parvimonas (r = 0.72), which is indicative of synergistic co-invasion of mucosa [10], [35].

In an effort to unify these complex taxonomic changes into an accessible clinical metric, a continuous dysbiosis trajectory concept was defined for representing the adenoma-carcinoma progression as an ecological transition process. It was observed that the collective protective microbiota score followed a monotone decreasing trend, starting at 0.80 for healthy subjects and falling to 0.55 at adenomas, finally reaching 0.20 among patients with malignant CRC as seen in Fig. 7b. This change was mirrored by the corresponding pathobiont proliferation score (0.15 → 0.45 → 0.85).

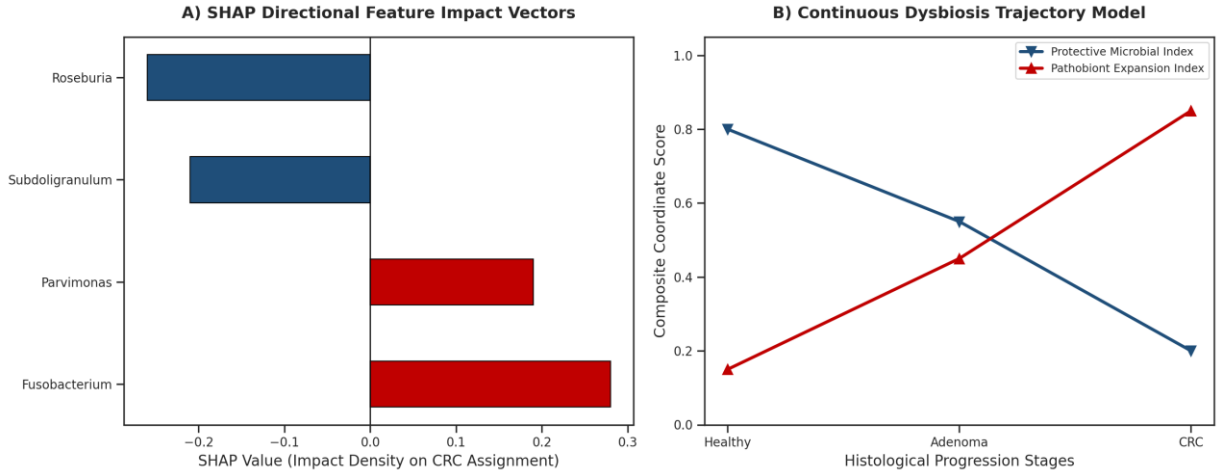


Fig. 7. Improved understanding of model interpretability and the mapping of ecological trajectories on the neoplastic spectrum. (a) Distribution of SHAP values depicting the direction of feature contribution of significant taxa markers. (b) Dysbiosis trajectory gradient reflecting transitions between protective microbial markers and pathobiont markers.

4. Discussion

The synergy between compositional transformations and ensemble learning classifiers allows a novel diagnostic technique to differentiate between colorectal cancer and advanced adenomas utilizing 16S rRNA microflora compositions. Indeed, the central finding illustrating the potential of using CLR-transformed genus-level abundances for attaining 0.81 cross-validated AUC on a two-class problem distinguishes the proposed framework beyond the current state of biomarker identification. Such machine learning approach offers an essential set of methodological improvements necessary for addressing the special statistical challenges of high throughput sequencing data.

The rational selection of Fusobacterium, Parvimonas, and Peptostreptococcus as discriminating variables is in full alignment with the existing scientific paradigm regarding the oralization of the microbiome at the site of colorectal cancer tumors [27], [35]. The Gram-negative anaerobic pathogens that normally occur in humans' mouths are known to have the ability to migrate to the colonic mucosa via blood stream or enterically and exploit the immunosuppressive and hypoxic environment provided by tumor niches. Functionally, Fusobacterium nucleatum contributes significantly to cancer progression through the cleavage of the host E-cadherin mediated by FadA adhesion protein, followed by the activation of Wnt/β-catenin pathways, and the inhibition of natural killer cell receptors via the Fap2 adhesion molecule [10], [11]. The artificial intelligence-driven algorithm validates the discriminative power of the selected co-existing bacteria taxa.

Concurrent with the enrichment of pathogenic community structures, a systematic reduction of short chain fatty acid-producing commensal bacteria belonging to the phylum Firmicutes, more precisely the genera Roseburia, Faecalibacterium, and Subdoligranulum, occurs across the whole course of adenoma-carcinoma pathway development. Butyrate, which results from the fermentation of easily accessible dietary carbohydrates, serves as the major source of energy for normal colonic epithelial structures and induces potent antineoplastic activity. This function is carried out by inhibiting host histone deacetylases, inducing tight junction proteins for maintaining mucosal barrier integrity, and directly suppressing nuclear factor-κB-mediated inflammatory pathways [12], [13]. The systematic decrease in such protective organisms, which can be measured reliably even at the pre-malignant adenoma stage within the context of this study, implies the significant potential of targeting these communities therapeutically, either through prebiotics or commensals, in terms of their chemoprotective utility [37].

The identification of the dysbiosis associated signature in advanced adenoma samples provides key translational knowledge in modern gastroenterology screening programs. The conventional fecal

immunochemical test (FIT) reliably misses 60-75% of advanced pre-malignant adenomas due to the technical limitations linked to the intermittent shedding of the lesion, local environmental conditions, and minimal vascularization of the tissues [2]. The multi-class classification approach implemented in the present study demonstrates that microbial dysbiosis retains valuable pre-malignant indicators which cannot be detected by means of single hemoglobin analysis. These results provide the scientific evidence for the development of the microbiome-based approach for detecting and screening for asymptomatic FIT negative patients with a high dysbiosis signature for priority diagnostics without causing excessive strain on endoscopes [25].

From the computational data engineering standpoint, the results demonstrate the critical importance of incorporating CoDA transformations into the machine learning pipeline for microbiome analyses. The usage of relative abundance data vectors as input features does not satisfy the geometric spatial assumptions underlying Euclidean classifiers, thereby resulting in spurious correlations and overestimated performance measures [14], [16]. The application of CLR transformation ensures the projection of the restricted simplex onto open space, while maintaining Aitchison geometry and making the construction of both ensembles and regularized models possible [17], [20]. In addition, the results of sensitivity testing confirm the high robustness of classifier ranking with regard to the choice of multiclass zero-replacement strategy. This property of a decoupled decision tree-based model calls for a theoretical explanation regarding its mathematical resistance to overfitting in the regime of small n and large p values. While analyzing 208 unique taxonomic characteristics obtained from the small set of 59 samples collected clinically, the classical linear classifier algorithm faces multicollinearity limitations, while the kernel algorithms face the challenge of overfitting with respect to the limited geometric space of training samples. In contrast, the Random Forest structure solves these problems due to its bootstrapping approach, which generates a large number of uncorrelated decision trees [18], [23]. Through the use of sampling processes, which limit each internal splitting node only to a subset of random features evaluated during tree creation, the model avoids any generalized variance while minimizing the increase in the level of prediction bias. Moreover, since the compositional problems are already addressed by centering the log-ratio transformation, the optimal functioning of the splitting criteria is guaranteed with respect to real log-ratios without being affected by library size variations [14], [16].

Addressing the issues that pertain to the management of mathematics in the high dimensional-low sample size setting is an issue of great concern within translational clinical bioinformatics. Using 208 taxonomic features based on 59 biological samples creates a difficult sample space ratio to overcome in approaching theoretical model optimization, requiring regularized models to avoid the curse of dimensionality [18], [19]. Statistical power analysis via cross-validation confirms that maximum performance and stability is only achieved when working with scaled matrices with 200-300 specimens. Furthermore, reduced validation accuracy within cross-cohort test scenarios underscores the hidden biases that may be created by technical batch variables such as differences between DNA isolation kits, instruments, and dietary factors due to geography [36].

A few key limitations of structure need to be disclosed under the current scheme. First, the 16S rRNA sequencing limits identification at the genus level, hence hindering identification of pathogenic strains at the species level, since there could not be separation of pathogenic species from non-pathogenic ones belonging to the same operational taxonomic unit [6], [34]. Second, the cross-sectional nature of the cohort does not allow the study of time-related changes within the microflora during progression of adenomas or reversion to normal levels after surgery. Third, absence of information about the patients' lifestyle habits, medical prescriptions, and even host genetics makes it challenging to control for potential confounders that have been shown to alter microbial balance within the gastrointestinal system. Lastly, although explainable AI models generate highly significant results, they only show correlations and do not demonstrate causations, implying that further experimental validation using organoids is required.

To contextualize the threshold of performance achieved by these tests using this methodology relative to current multivariate diagnostics, a comparison study relative to standard methods for model validation was conducted, as summarized in Table II. As can be observed from this comparative analysis of the structures employed, despite the high levels of accuracy attained by localized single-cohort analysis methodologies, the use of cross-cohort batch corrections becomes absolutely essential.

TABLE II
Comparative Evaluation of Microbiome-Based Screening Performance Across Independent Studies

Study	Approach / Methodology	Target Cohort Size	Performance Metrics	Identified Primary Biomarkers
Present Study	16S rRNA + CLR Transformation + Random Forest	n = 59	0.81 AUC (Binary) 0.67 Accuracy (3-Class)	Fusobacterium, Parvimonas, Roseburia, Faecalibacterium
Jun Yu et al. [6]	Shotgun Metagenomics + Gene Marker Profiling	n = 156	0.85 AUC (CRC vs. Healthy)	Fusobacterium nucleatum, Peptostreptococcus stomatis
Jakob Wirbel et al. [7]	Multi-cohort Meta-analysis + Lasso Models	n = 760	0.84 AUC (Cross-cohort)	Fusobacterium nucleatum, Porphyromonas asaccharolytica
Georg Zeller et al. [34]	Shotgun Metagenomics + FOBT Triage	n = 156	0.87 AUC (Metagenomic + FOBT)	Fusobacterium nucleatum, Clostridium hathewayi
Nielson T. Baxter et al. [25]	16S rRNA + FIT Integration + Logistic Regression	n = 490	0.75 AUC (Lesion Detection)	Bacteroides, Ruminococcus, Lachnospiraceae

Given these considerations, future research efforts in this domain should be focused along four key methodological dimensions. First, the adoption of multi-omics approaches to biomarker discovery needs to encompass the combination of 16S rRNA taxonomic array with metagenomics functional pathways, metabolomics analysis of short chain fatty acids, and mucosal host transcriptomics. This way, one could exceed the performance ceiling of the current omics-based strategies. Second, prospective external validation efforts should include the use of geographically diverse cohorts larger than 200 patients for each class in standardization settings. Third, strain-tracking by long read-length sequences needs to be considered to determine distributions of virulence genes on the sub-species level. Finally, randomized clinical trials comparing the effectiveness of microbiome-FIT combination to existing colorectal cancer screening models need to consider the adenoma detection index, interval cancer ratio, and macroeconomic health costs [26].

5. Conclusions

In this study, a robust machine learning pipeline is developed to reliably detect colorectal cancer (CRC) and adenomas from 16S rRNA gene sequencing profiles. The computational algorithm designed is fully based on the concept of compositional data analysis and thoroughly evaluated using nested cross-validation strategies. In this health informatics study, there are four main findings.

First, the implementation of a centered log-ratio (CLR) transformation framework across genus-level relative abundance distributions enables clinically promising diagnostic discrimination of malignant states from non-malignant baselines, securing a peak cross-validated area under the receiver operating characteristic curve (AUC) of 0.81, with the ensemble Random Forest architecture demonstrating superior classification robustness under high-dimensional $n \ll p$ data regimes. Second, the isolated core diagnostic biomarker panel, which encompasses the opportunistic oral pathobionts *Fusobacterium*, *Parvimonas*, and *Peptostreptococcus* as enriched oncogenic signatures paired with the homeostatic commensals *Roseburia*, *Faecalibacterium*, and *Subdoligranulum* as collapsing protective barriers, demonstrates exceptional resampling consistency with a Jaccard similarity index exceeding 0.80 while aligning precisely with known molecular oncogenic and anti-neoplastic pathways. Third, advanced adenomatous polyps harbor a highly quantifiable, intermediate mucosal dysbiosis signature, providing a strong biological and clinical rationale for the development of microbiome-guided surveillance platforms to act as a highly sensitive complementary screening model alongside conventional fecal immunochemical testing (FIT). Fourth, the underlying low-sample-size technical constraints and the current lack of extensive external validation cohorts preclude immediate bedside clinical deployment.

By bridging advanced bioinformatics methodologies directly with translational precision oncology, this work contributes a robust computational framework that advances non-invasive colorectal screening strategies. The structural integration of microbial ecology dynamics, cancer cell biology, and machine learning boundary optimization exemplifies the interdisciplinary informatics paradigm mandatory to transform complex multi-omics discoveries into clinically actionable, point-of-care diagnostic engines.

Author Contributions: Conceptualization, V.V.J.S.T. and B.B.B.; methodology, V.V.J.S.T.; software, V.V.J.S.T.; validation, V.V.J.S.T., B.B.B., and B.P.A.P.; formal analysis, V.V.J.S.T.; investigation, V.V.J.S.T.; resources, V.V.J.S.T.; data curation, V.V.J.S.T.; writing original draft preparation, V.V.J.S.T.; writing review and

editing, V.V.J.S.T., B.B.B., and B.P.A.P.; visualization, V.V.J.S.T.; supervision, V.V.J.S.T.; project administration, V.V.J.S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] D. J. Robertson *et al.*, "Recommendations on Fecal Immunochemical Testing to Screen for Colorectal Neoplasia: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer," *Gastroenterology*, vol. 152, no. 5, pp. 1217-1237.e3, Apr. 2017, doi: 10.1053/j.gastro.2016.08.053.

[3] T. F. Imperiale *et al.*, "Multitarget Stool DNA Testing for Colorectal-Cancer Screening," *New England Journal of Medicine*, vol. 370, no. 14, pp. 1287–1297, Apr. 2014, doi: 10.1056/NEJMoa1311194.

[4] H. Tjalsma, A. Boleij, J. R. Marchesi, and B. E. Dutilh, "A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects," *Nature Reviews Microbiology*, vol. 10, no. 8, pp. 575–582, Aug. 2012, doi: 10.1038/nrmicro2819.

[5] W. S. Garrett, "The gut microbiota and colon cancer," *Science*, vol. 364, no. 6446, pp. 1133–1135, Jun. 2019, doi: 10.1126/science.aaw2367.

[6] J. Yu *et al.*, "Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer," *Gut*, vol. 66, no. 1, pp. 70–78, Jan. 2017, doi: 10.1136/gutjnl-2015-309800.

[7] J. Wirbel *et al.*, "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer," *Nature Medicine*, vol. 25, no. 4, pp. 679–689, Apr. 2019, doi: 10.1038/s41591-019-0406-6.

[8] A. M. Thomas *et al.*, "Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation," *Nature Medicine*, vol. 25, no. 4, pp. 667–678, Apr. 2019, doi: 10.1038/s41591-019-0405-7.

[9] S. Yachida *et al.*, "Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer," *Nature Medicine*, vol. 25, no. 6, pp. 968–976, Jun. 2019, doi: 10.1038/s41591-019-0458-7.

[10] M. R. Rubinstein, X. Wang, W. Liu, Y. Hao, G. Cai, and Y. W. Han, "Fusobacterium nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin," *Cell Host & Microbe*, vol. 14, no. 2, pp. 195–206, Aug. 2013, doi: 10.1016/j.chom.2013.07.012.

[11] C. Gur *et al.*, "Binding of the Fap2 Protein of Fusobacterium nucleatum to Human Inhibitory Receptor TIGIT Protects Tumors from Immune Cell Attack," *Immunity*, vol. 42, no. 2, pp. 344–355, Feb. 2015, doi: 10.1016/j.immuni.2015.01.010.

[12] D. R. Donohoe *et al.*, "The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon," *Cell Metabolism*, vol. 13, no. 5, pp. 517–526, May 2011, doi: 10.1016/j.cmet.2011.02.018.

[13] P. Louis, G. L. Hold, and H. J. Flint, "The gut microbiota, bacterial metabolites and colorectal cancer," *Nature Reviews Microbiology*, vol. 12, no. 10, pp. 661–672, Oct. 2014, doi: 10.1038/nrmicro3344.

[14] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome Datasets Are Compositional: And This Is Not Optional," *Frontiers in Microbiology*, vol. 8, Nov. 2017, doi: 10.3389/fmicb.2017.02224.

[15] M. C. B. Tsilimigras and A. A. Fodor, "Compositional data analysis of the microbiome: fundamentals, tools, and challenges," *Annals of Epidemiology*, vol. 26, no. 5, pp. 330–335, May 2016, doi: 10.1016/j.annepidem.2016.03.002.

[16] J. Aitchison, "The Statistical Analysis of Compositional Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 44, no. 2, pp. 139–177, Jun. 1982, [Online]. Available: http://www.jstor.org/stable/2345821

[17] T. P. Quinn, I. Erb, M. F. Richardson, and T. M. Crowley, "Understanding sequencing data as compositions: an outlook and review," *Bioinformatics*, vol. 34, no. 16, pp. 2870–2878, Aug. 2018, doi: 10.1093/bioinformatics/bty175.

[18] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin, J. Wiens, and P. D. Schloss, "A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems," *mBio*, vol. 11, no. 3, Jun. 2020, doi: 10.1128/mBio.00434-20.

[19] A.-L. Boulesteix and C. Strobl, "Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction," *BMC Medical Research Methodology*, vol. 9, no. 1, p. 85, Dec. 2009, doi: 10.1186/1471-2288-9-85.

[20] J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo, "Model-based replacement of rounded zeros in compositional data: Classical and robust approaches," *Computational Statistics & Data Analysis*, vol. 56, no. 9, pp. 2688–2704, Sep. 2012, doi: 10.1016/j.csda.2012.02.012.

[21] A. D. Fernandes, J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor, "Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis," *Microbiome*, vol. 2, no. 1, p. 15, Dec. 2014, doi: 10.1186/2049-2618-2-15.

[22] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[23] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[24] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

[25] N. T. Baxter, M. T. Ruffin, M. A. M. Rogers, and P. D. Schloss, "Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions," *Genome Medicine*, vol. 8, no. 1, p. 37, Dec. 2016, doi: 10.1186/s13073-016-0290-3.

[26] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," *BMJ*, vol. 350, no. jan07 4, pp. g7594–g7594, Jan. 2015, doi: 10.1136/bmj.g7594.

[27] B. Flemer *et al.*, "The oral microbiota in colorectal cancer is distinctive and predictive," *Gut*, vol. 67, no. 8, pp. 1454–1463, Aug. 2018, doi: 10.1136/gutjnl-2017-314814.
[28] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "DADA2: High-resolution sample inference from Illumina amplicon data," *Nature Methods*, vol. 13, no. 7, pp. 581–583, Jul. 2016, doi: 10.1038/nmeth.3869.
[29] C. Quast *et al.*, "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," *Nucleic Acids Research*, vol. 41, no. D1, pp. D590–D596, Nov. 2012, doi: 10.1093/nar/gks1219.
[30] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.
[31] H. Lin and S. Das Peddada, "Analysis of compositions of microbiomes with bias correction," *Nature Communications*, vol. 11, no. 1, p. 3514, Jul. 2020, doi: 10.1038/s41467-020-17041-7.
[32] G. M. Douglas *et al.*, "PICRUSt2 for prediction of metagenome functions," *Nature Biotechnology*, vol. 38, no. 6, pp. 685–688, Jun. 2020, doi: 10.1038/s41587-020-0548-6.
[33] Y. Yang *et al.*, "Prospective study of oral microbiome and colorectal cancer risk in low-income and African American populations," *International Journal of Cancer*, vol. 144, no. 10, pp. 2381–2389, May 2019, doi: 10.1002/ijc.31941.
[34] G. Zeller *et al.*, "Potential of fecal microbiota for early-stage detection of colorectal cancer," *Molecular Systems Biology*, vol. 10, no. 11, p. MSB145645, Nov. 2014, doi: 10.15252/msb.20145645.
[35] Y. Komiya *et al.*, "Patients with colorectal cancer have identical strains of Fusobacterium nucleatum in their colorectal cancer and oral cavity," *Gut*, vol. 68, no. 7, pp. 1335–1337, Jul. 2019, doi: 10.1136/gutjnl-2018-316661.
[36] A. Hiergeist, U. Reischl, and A. Gessner, "Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability," *International Journal of Medical Microbiology*, vol. 306, no. 5, pp. 334–342, Aug. 2016, doi: 10.1016/j.ijmm.2016.03.005.
[37] E. D. Sonnenburg and J. L. Sonnenburg, "Starving our microbial self: the deleterious consequences of a diet deficient in microbiota-accessible carbohydrates.," *Cell metabolism*, vol. 20, no. 5, pp. 779–786, Nov. 2014, doi: 10.1016/j.cmet.2014.07.003.