

An Optimised Computational Framework for Cross-Cohort Metagenomic Classification Using P-Value Statistical Filtering

Hien Thanh Thi Nguyen¹, Hai Thanh Nguyen^{2*}

^{1,2}*College of Information and Communication Technology, Can Tho University, Can Tho, Vietnam*

Abstract – Metagenomic data analysis faces significant challenges due to high dimensionality, sparsity, and geographical heterogeneity, which limit the generalizability of diagnostic models. This study proposes a robust computational framework that integrates p-value-based statistical filtering with ensemble machine learning to identify stable microbial signatures for colorectal cancer (CRC). By ranking features via T-tests, we identified an optimal subset of 170 features, significantly reducing dimensionality from the original 2000 features while enhancing predictive performance. Results demonstrate high system efficiency: the Gradient Boosting Classifier (GBC) achieved an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.88 on the Zeller cohort, compared to 0.80 using the full feature set. Extensive cross-cohort evaluations across four global datasets (Germany/France, Austria, China, and USA) identified the Zeller dataset as a “Gold Standard” for feature extraction, yielding biomarkers that are highly transferable across diverse populations. Furthermore, ensemble models (GBC and Random Forest) exhibited superior resilience to distribution shifts compared to SVM, which suffered from performance collapse in cross-dataset scenarios. This framework provides a computationally lean and scalable solution for large-scale metagenomic studies, emphasising the importance of accounting for population-specific variability in robust system design.

Keywords – Disease prediction, feature selection, machine learning, metagenomic data, T-test.

I. INTRODUCTION

Microbial communities inhabiting the human body play a fundamental role in maintaining physiological homeostasis, and alterations in their composition have been strongly associated with a wide range of diseases. Recent advances in metagenomic sequencing have enabled large-scale profiling of the human microbiome, leading to the widespread use of species-level relative abundance profiles to characterise disease-associated microbial shifts. Growing evidence indicates that changes in microbial species abundance can serve as robust, non-invasive biomarkers for disease diagnosis and prognosis, including colorectal cancer, where consistent microbial signatures have been observed across multiple cohorts [1]–[3]. Importantly, accumulating evidence suggests that these microbial signatures are not entirely universal, but are partly shaped by geographical

and demographic factors, particularly dietary patterns that differ across populations. Large-scale metagenomic studies have shown that while a core set of colorectal cancer (CRC)-associated microbes is shared globally, substantial variability exists between Western and Asian cohorts due to differences in lifestyle, diet, and environmental exposures [2], [3].

In parallel, machine learning (ML) has become a core analytical paradigm for biomedical data analysis due to its ability to learn complex patterns from high-dimensional data. In microbiome research, ML models trained on species abundance profiles have demonstrated strong predictive performance for disease classification and biomarker discovery. Notably, Su et al. showed that ML models using species-level abundance features could accurately distinguish multiple disease phenotypes, including colorectal cancer, and generalise across independent cohorts [4]. More recent benchmarking studies further confirm that abundance-based ML pipelines can achieve competitive diagnostic performance without relying on gene- or pathway-level features [5].

Despite these advances, applying machine learning to microbial abundance data remains challenging. Species abundance matrices are typically characterised by high dimensionality, sparsity, compositionality, and noise, while the number of samples is often limited. These properties can negatively affect model generalisation and increase the risk of overfitting when raw abundance profiles are directly used for training [6], [7]. In addition, inter-cohort heterogeneity and technical variation further complicate the development of robust and transferable microbiome-based diagnostic models [5].

Feature selection has therefore emerged as a critical step in abundance-based machine learning pipelines. By reducing dimensionality and retaining only the most informative microbial species, feature selection can improve classification performance, enhance model interpretability, and reduce computational cost. Recent studies and reviews on microbiome-based cancer prediction highlight that appropriate feature selection and statistical filtering play a key role in stabilising ML models and improving robustness across datasets [5], [8]. In microbiome analysis, it is common to rank microbial taxa by

*Corresponding author's e-mail: nthai.cit@ctu.edu.vn
Article received 2026-04-07; accepted 2026-05-26

statistical significance (based on p-values from differential abundance tests) to select features associated with phenotypes of interest, and combining or adjusting these p-values has been proposed as an effective way to determine important taxa before downstream modelling [9].

Despite the promise of metagenomics as a non-invasive tool for disease diagnosis ‘geographical bias’ remains a formidable barrier to the development of robust clinical models. Current machine learning approaches often suffer from limited generalizability; they are frequently ‘brittle’ – excelling within their discovery cohorts but faltering when confronted with the vast microbial heterogeneity of global populations. This lack of transferability is largely driven by stark contrasts in lifestyle and dietary patterns, particularly between Western populations – whose fat-rich diets promote carcinogenic metabolites – and Asian populations, where fibre-rich diets support protective microbial functions. Such variability necessitates a shift from simple model training to a more statistically-principled framework.

Motivated by these challenges, this study introduces a unified computational framework that integrates p-value-based statistical filtering with ensemble learning architectures to address the inherent noise and high dimensionality of metagenomic data. The core contribution lies in an automated feature selection pipeline that systematically prunes the metagenomic feature space – reducing it by over 90 % (from approximately 2000 to a compact core of 170 species) – while actually enhancing predictive accuracy. For instance, our framework enabled the Gradient Boosting Classifier (GBC) to achieve a peak Area Under the Receiver Operating Characteristic Curve (AUC) of 0.88, significantly outperforming the 0.80 baseline using full-dimensional data. Furthermore, we provide a comprehensive cross-cohort evaluation across four global datasets (Europe, Asia, and North America), establishing the Zeller cohort as a “Gold Standard” architecture for extracting biologically conserved and transferable microbial signatures. By demonstrating the superior resilience of ensemble models (GBC and Random Forest) over Support Vector Machines (SVM) under distribution shifts, this work offers a computationally lean and scalable system design. Our findings provide a practical, high-throughput solution to colorectal cancer screening that remains robust across heterogeneous, real-world populations.

The remainder of this paper is organised as follows. Section II reviews related work relevant to metagenomic analysis and machine learning-based disease diagnosis. The research questions addressed in this study are presented in Section III. Section IV provides a detailed description of the datasets considered in this work, together with the learning models and the proposed feature selection methodology. Section V reports and discusses the experimental results, including AUC evaluation, p-value analysis, and comparative performance assessments. Finally, Section VI concludes the paper and summarises the main findings of the study.

II. RELATED WORK

The application of machine learning to metagenomic species abundance data has shown considerable promise in disease prediction, particularly for CRC. Recent studies have demonstrated that abundance-based representations can effectively capture disease-associated microbial shifts and support accurate classification without relying on functional annotations or pathway-level features [1], [10], [11]. Large-scale benchmarking efforts further confirm that species-level abundance profiles provide a strong basis for non-invasive CRC prediction when combined with appropriate ML models and evaluation strategies [4], [5].

Importantly, accumulating evidence suggests that gut microbiome composition is strongly influenced by geographical and demographic factors, particularly dietary patterns that differ between Western and Asian populations. These differences can lead to distinct microbial community structures and CRC-associated signatures across cohorts, thereby affecting the reproducibility and transferability of machine learning models [2], [3]. For example, Western diets rich in fat and red meat have been associated with microbiome profiles that promote carcinogenic metabolites, whereas fibre-rich diets more common in Asian populations support protective microbial functions [12].

Nevertheless, metagenomic abundance data pose several inherent challenges for machine learning analysis, including high dimensionality, sparsity, compositionality, and strong cohort-specific variability. In most microbiome studies, the number of microbial taxa greatly exceeds the number of samples, which increases the risk of overfitting and limits the generalizability of learned models across independent datasets [7], [13]. These challenges are further exacerbated by cohort heterogeneity, including differences in population background, sequencing protocols, and environmental factors, which have been shown to substantially affect microbiome-based disease prediction and cross-cohort transferability [3], [14].

In particular, recent cross-population studies have highlighted that variations in diet, lifestyle, and host demographics contribute significantly to microbiome heterogeneity, making cross-cohort prediction a non-trivial task [3], [15]. Such population-specific effects may lead to reduced model performance when features learned from one geographic region are applied to another, emphasising the importance of identifying conserved microbial biomarkers that are robust across diverse populations.

To address these issues, feature selection has become a central component of microbiome ML pipelines. Statistical filtering approaches based on univariate hypothesis testing are widely adopted to rank microbial taxa according to their discriminative power between disease and control groups. Differential abundance analyses that rely on p-value-based criteria have been shown to substantially reduce feature dimensionality while preserving informative microbial signals, thereby improving model interpretability and computational efficiency in downstream classification tasks [6], [16], [17]. Such statistical pre-processing is especially attractive for

abundance-based data due to its simplicity, transparency, and compatibility with a wide range of classifiers.

In addition to feature pre-processing, the choice of machine learning classifiers strongly influences predictive performance on abundance-based metagenomic data. Support Vector Machines (SVM) have been extensively applied due to their effectiveness in high-dimensional feature spaces [14]. At the same time, ensemble-based methods such as Random Forest (RF) and GBC are increasingly favoured for their robustness to noise, ability to capture nonlinear relationships, and relative stability under feature reduction [5], [11], [18]. Comparative studies consistently report that ensemble models achieve competitive or superior performance when combined with prior feature selection, particularly in CRC classification tasks [3], [4].

In particular, most existing studies focus on optimising classification performance within individual cohorts or rely on model-specific importance measures, rather than examining the generalizability of statistically selected features across heterogeneous metagenomic datasets [5]. This limitation becomes more critical in the presence of geographical and demographic variability, where microbiome composition and CRC-associated patterns may differ substantially across populations due to dietary and environmental factors [2], [3]. Although prior studies have explored statistical feature filtering and machine learning models for metagenomic disease prediction, there remains a lack of unified frameworks that systematically evaluate p-value-based feature selection across multiple abundance-based datasets and cross-cohort experimental settings. This limitation highlights the need for a principled pre-processing strategy that can identify compact yet informative microbial feature subsets while maintaining predictive robustness [7], [8].

III. RESEARCH QUESTIONS

To advance the development of metagenomics-based approaches for disease diagnosis, this study proposes a specially designed architecture that integrates statistical analysis and machine learning techniques for feature selection from high-dimensional metagenomic data. Within this framework, a set of research questions (RQs) is formulated to systematically investigate the effectiveness of the proposed method in supporting colorectal cancer detection.

- RQ1: How do different training strategies and learning models influence classification performance in colorectal cancer detection using metagenomic data?

To address this question, multiple training strategies and classification models are implemented and compared to enhance the predictive performance of the proposed framework. The details of these strategies are described in Section IV, while a comprehensive performance evaluation and comparative analysis are provided in Section V.

- RQ2: To what extent do demographic and geographic variations, particularly diet-driven differences in gut microbiome composition, affect the generalizability and cross-cohort performance of metagenomics-based models for colorectal cancer detection?

This question examines whether CRC classification models should explicitly account for population heterogeneity. The experimental results indicate the presence of notable demographic and geographic differences, highlighting their potential impact on model generalizability and diagnostic performance.

- RQ3: Can statistically selected feature subsets identify microbial biomarkers that remain robust across geographically and demographically diverse cohorts, thereby improving both interpretability and cross-cohort generalizability of CRC prediction models?

To address this question, the proposed p-value-based feature selection strategy is evaluated in terms of its ability to identify compact and informative microbial taxa that are consistently associated with CRC across multiple cohorts. The analysis focuses on whether these statistically derived feature subsets capture conserved microbial patterns that are less sensitive to population-specific variability, thereby enhancing both the interpretability of the selected biomarkers and the robustness of classification models under cross-cohort settings.

IV. METHODOLOGY

Motivated by recent advances in machine learning-based analysis of metagenomic data, this study proposes a feature selection-driven framework for disease diagnosis. As illustrated in Fig. 1, the proposed methodology consists of several sequential stages, including data pre-processing, statistical feature selection, model training, and performance evaluation. Experiments are conducted on multiple publicly available colorectal cancer metagenomic datasets to ensure robustness and generalizability across heterogeneous data sources. The core contribution of this work lies in the development of a p-value-based feature selection strategy that aims to identify a compact yet informative subset of metagenomic features, enabling effective model training while mitigating the challenges posed by the extremely high dimensionality of the original data. Feature relevance is assessed by statistically comparing feature distributions between disease and control groups. During pre-processing and model construction, features with zero variance, insufficient observations, or undefined statistical measures are handled to ensure numerical stability. Furthermore, for cross-dataset evaluations where selected features may be absent in the target dataset, a zero-imputation strategy is adopted to preserve feature matrix consistency and to address the inherent sparsity of metagenomic profiles. Detailed descriptions of each stage of the proposed framework are provided in the subsequent sections.

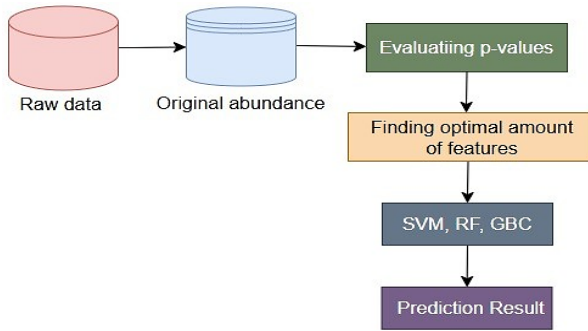


Fig. 1. Workflow of the proposed feature selection and classification framework based on p-values and machine learning for metagenomic data.

TABLE I
SUMMARY OF COLORECTAL CANCER METAGENOMIC DATASETS USED IN THIS STUDY

Dataset	Region (Year)	Healthy	CRC	Total	Avg. Age (H/CRC)	Features
Feng	Austria (2014)	63	46	109	67.1 / 67.1	1981
Vogtmann	USA (2016)	52	48	100	61.23 / 60.96	1976
Yu	China (2015)	92	73	165	58.51 / 65.90	1932
Zeller	Germany & France (2014)	64	88	152	58.75 / 68.44	1980

A. Data Description

This study employs four publicly available metagenomic datasets related to CRC, namely the Feng dataset, Vogtmann dataset, Yu dataset, and Zeller dataset. These datasets were collected independently across different geographic regions and population groups, enabling a comprehensive evaluation of the proposed framework under heterogeneous conditions.

A summary of the datasets is presented in Table I, including the number of healthy controls and CRC patients, total sample size, average age, and the number of microbial features. The cohorts vary in size and composition: the Yu dataset contains the largest number of samples, while the Feng and Vogtmann datasets are relatively smaller. Each dataset consists of high-dimensional microbial profiles, with approximately two thousand features derived from Operational Taxonomic Units (OTUs) obtained via 16S rRNA sequencing. Such high dimensionality reflects the complexity of metagenomic data and poses significant challenges for downstream analysis, particularly in feature selection and classification tasks.

To further characterise the datasets, Fig. 2 illustrates the demographic information, including sample size, average age, and gender distribution for both healthy and CRC groups. Noticeable variability can be observed across the cohorts. In particular, the Yu dataset shows a substantially larger sample size, especially among healthy individuals, whereas the Feng and Vogtmann datasets exhibit more limited cohort sizes. The average age across datasets is generally comparable between healthy and CRC groups, although certain differences are evident, such as in the Zeller dataset where CRC patients tend to be older. In terms of gender distribution, male participants are predominant in most datasets; however, the degree of imbalance varies across cohorts.

These demographic and cohort-level differences highlight the inherent heterogeneity among datasets collected from distinct geographic regions. Such variability provides a suitable testbed for assessing the robustness and generalizability of the proposed method, particularly in cross-dataset and real-world

scenarios where population characteristics may differ significantly.

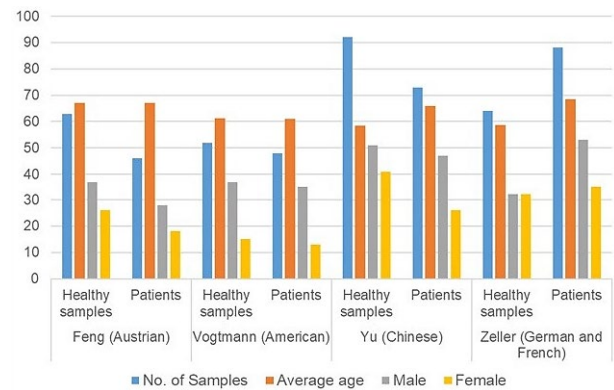


Fig. 2. Demographic information across the four metagenomic datasets, including sample size, average age, and gender distribution for healthy controls and CRC patients.

B. Feature Selection

An independent two-sample t-test was applied to each dataset to identify features exhibiting statistically significant differences between patient and non-patient groups. Features were ranked based on their p-values, and a significance threshold of $\alpha = 0.05$ was adopted. Features with zero variance or insufficient observations were excluded to ensure reliable statistical inference.

T-Test Analyzing: After performing the T-test on each original dataset, all features are ranked in ascending order according to their corresponding p-values. This ranking process allows the identification of features with stronger statistical significance between disease and control groups. The results are visually summarised using p-value distribution plots, which provide an intuitive comparison of how p-values increase across different datasets and facilitate the assessment of feature discriminability within each dataset.

Determination of the Optimal Number of Features: Representative features for disease classification are selected by focusing on those with p-values below 0.05, indicating statistically significant differences between classes. These features are further ranked by their p-values within each dataset. Comparative analysis across datasets identifies subsets of features that consistently display low p-values, which are prioritized as candidate biomarkers. This approach facilitates the construction of reduced feature sets that retain discriminative information while substantially decreasing the number of features compared to the original datasets.

C. Classification Models

The selected feature subsets were subsequently used as inputs for disease classification and performance evaluation. Three machine learning classifiers, namely Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Classifier (GBC), were employed to assess their predictive capability. The RF classifier was implemented with 500 decision trees ($n_{\text{estimators}} = 500$). To evaluate the effectiveness of feature selection, the same classifiers were also trained on the corresponding original datasets, and their performance was compared with that obtained using the reduced feature subsets.

To evaluate the generalizability of the selected biomarkers across datasets, a cross-dataset feature alignment strategy is employed. Specifically, the top-ranked feature subset obtained from one dataset is used as a reference feature list. Based on this feature list, the corresponding features are extracted from the remaining datasets by matching feature names. This procedure is repeated in a round-robin manner, where each dataset in turn serves as the reference for feature selection, and the aligned feature sets from all datasets are subsequently used for classification.

In cases where a selected feature from the reference dataset is absent in a target dataset, the corresponding feature values are assigned to zero, indicating the absence of that microbial taxon in the target dataset. This strategy ensures a consistent feature dimensionality across datasets while preserving biologically meaningful absence information, which may arise from geographical or population-specific variations.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Evaluation Metrics

The performance of the proposed classification framework was evaluated using the AUC, which serves as the primary evaluation metric for comparing different machine learning models, including SVM, RF, and GBC. AUC provides a threshold-independent assessment of a model's discriminative capability by measuring its ability to distinguish between disease and control samples across all possible decision thresholds. An AUC value of 0.5 corresponds to random classification, whereas a value of 1.0 indicates perfect discrimination. This metric is particularly suitable for metagenomic datasets, where class imbalance and high-dimensional feature spaces are common. By employing AUC as a unified evaluation criterion, a fair and consistent

comparison of the classification performance among SVM, RF, and GBC models was ensured across different datasets and experimental settings.

B. Statistical Feature Filtering and Cross-Dataset Discrepancy

After applying a two-sample t-test to each original dataset independently, the resulting p-value distributions indicate that only a subset of features exhibits statistically significant differences between colorectal cancer and control samples (Fig. 3). Using a significance threshold of $p < 0.05$, approximately 170 features were retained from the Feng dataset (Fig. 3a), around 260 features from the Vogtmann dataset (Fig. 3b), about 250 features from the Yu dataset (Fig. 3c), and a substantially larger set of approximately 600 features from the Zeller dataset (Fig. 3d). This pronounced variation in both the number and identity of statistically significant features across datasets highlights the strong heterogeneity of metagenomic profiles derived from different cohorts and experimental settings.

In addition, a small number of features in the Feng and Zeller datasets failed to yield valid p-values due to insufficient variance and were therefore excluded during pre-processing. Consequently, T-test-based filtering resulted in dataset-specific feature subsets with markedly different dimensionalities and limited overlap. Because the Feng dataset yielded the smallest number of statistically significant features, this set size was regarded as the most restrictive yet feasible choice and was therefore adopted as the optimal reference dimensionality for subsequent cross-dataset experiments. Such discrepancies in the selected feature spaces pose a challenge for direct cross-dataset model transfer, as classifiers trained on features derived from one dataset cannot be directly applied to others without reconciling feature mismatches.

Motivated by this observation, we subsequently introduce a cross-dataset feature alignment strategy in which the top 170 features with the lowest p-values identified from a source dataset are used as a reference feature set, and the corresponding features in the remaining datasets are selected strictly based on matched OTU identifiers. This procedure enforces a consistent feature space across datasets while preserving the statistical ranking derived from the source domain, thereby enabling reliable cross-dataset evaluation.

C. Internal Model Optimisation via Feature Selection

The initial phase of the analysis examined the impact of p-value-based feature selection within each cohort. Integrating p-value-based feature selection significantly improved predictive performance across most studied cohorts. Comparisons between the original full-feature sets and the top 170 subsets revealed that dimensionality reduction alleviates computational burden and enhances model discriminative power.

Superiority of Ensemble Learning with Feature Selection: The ensemble models, particularly RF and GBC, demonstrated strong compatibility with the feature selection process. Across the Zeller (Germany/France) and Feng (Austria) datasets, these models consistently outperformed their full-feature counterparts. As shown in Fig. 4d, the GBC model achieved an AUC of approximately 0.88 on the Zeller cohort when trained

on the selected feature subset, representing a substantial improvement over the baseline AUC of 0.80 obtained using the original feature set. Similarly, as illustrated in Fig. 4a, the RF model on the Feng dataset improved from a baseline AUC of 0.77 to 0.85 after feature selection. These results highlight the effectiveness of feature selection in reducing noise and emphasising informative microbial biomarkers.

Model-Specific Sensitivity and Dataset Heterogeneity: While ensemble methods consistently benefited from feature

selection, the SVM exhibited pronounced sensitivity to feature density. In the Yu (China) cohort (Fig. 4c), SVM performance increased from an AUC of 0.61 to 0.69 as the number of selected features exceeded the Top 80, indicating that SVM required a critical mass of informative features to construct an effective decision boundary in high-dimensional metagenomic spaces.

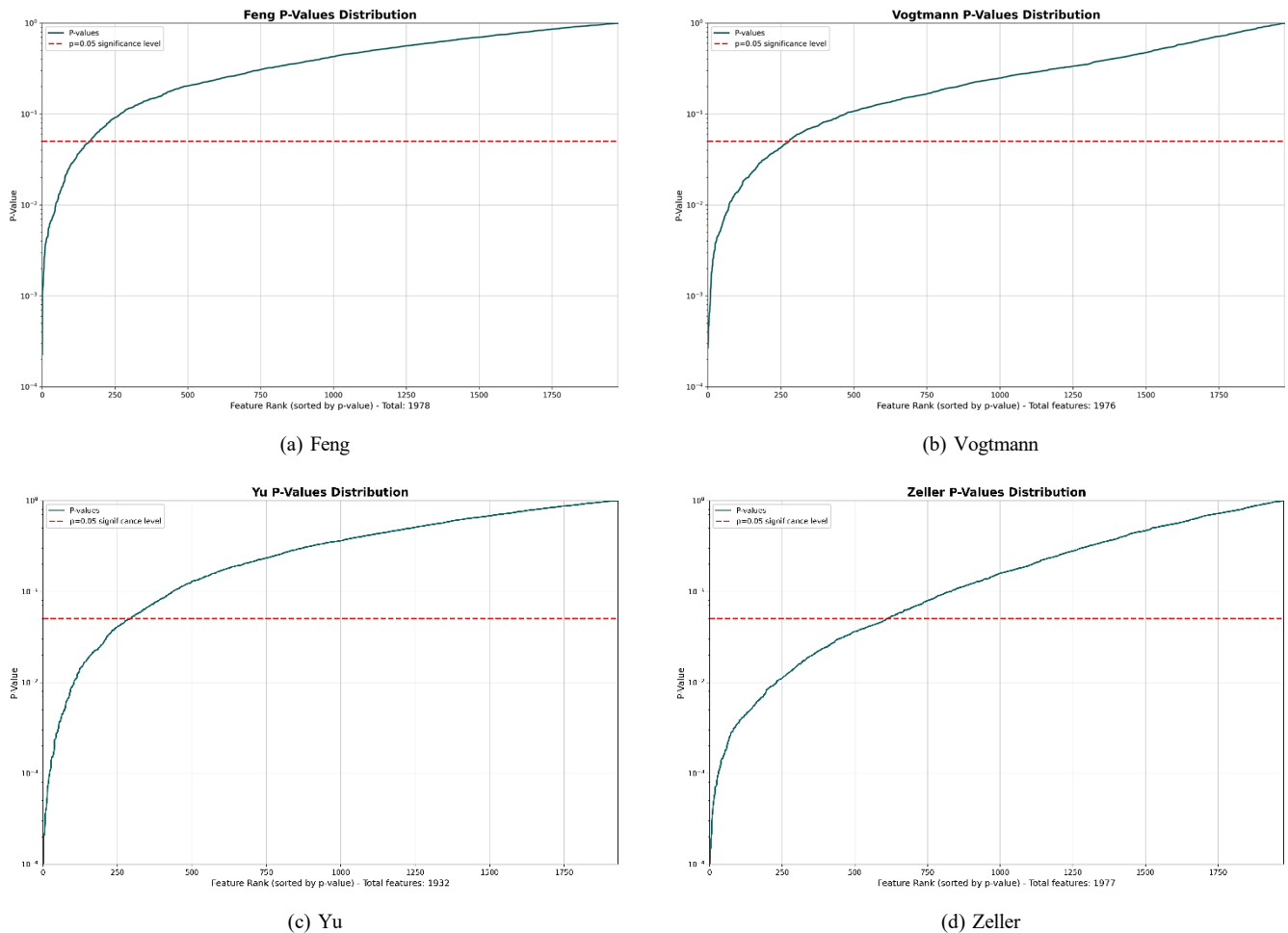


Fig. 3. Ranked feature p-value distributions across four colorectal cancer cohorts.

In contrast, the Vogtmann (USA) dataset (Fig. 4b) posed a distinct challenge, as SVM performance deteriorated markedly under aggressive feature reduction. This observation suggests that in populations characterised by greater microbial heterogeneity, the full feature set may capture subtle yet discriminative signals that are lost during pruning. Notably, even in this complex cohort, RF and GBC demonstrated robust stability at lower feature counts (Top 10–30), consistently matching or surpassing the baseline performance achieved using the original feature set.

Synthesis of Findings: Ensemble methods (RF and GBC) emerged as the primary beneficiaries of this approach (Fig. 4). In the Zeller and Feng datasets, the “Top 170” models significantly outperformed the original baselines, with GBC in

the Zeller cohort reaching an AUC of 0.88 compared to the 0.80 original. This confirms that a substantial portion of the metagenomic features in the original sets acts as statistical noise, which, when pruned, allows tree-based models to capture more robust biological signals. These findings advocate for a lean, biomarker-focused diagnostic approach which is both more accurate and computationally efficient for large-scale clinical applications.

D. Cross-Dataset Feature Robustness and “Gold Standard” Signatures

To evaluate the universal applicability of the identified biomarkers, a cross-dataset feature selection procedure was implemented. Models were trained using the top features

derived from one source dataset and tested on the remaining three target datasets.

Tables II–V summarize the cross-dataset AUC performance of SVM, RF, and GBC models trained on the Feng, Vogtmann, Yu, and Zeller datasets and evaluated on external cohorts. Overall, RF and GBC consistently outperformed SVM, indicating better generalization across datasets. GBC achieved the highest average AUC in most training–testing scenarios (Tables II–V), while RF also demonstrated stable and competitive performance. In general, predictive performance improved as the number of selected features increased, reaching its best results around Top130–Top170 features. These findings suggest that ensemble-based models combined with an appropriate feature subset can effectively enhance cross-dataset prediction performance.

Zeller cohort as the “Gold Standard”: Features derived from the Zeller (Germany/France) cohort demonstrated superior stability and generalizability. As shown in Fig. 5, when applied to the Yu (China) and Feng (Austria) datasets, Zeller’s signatures achieved peak AUCs (0.82 and 0.80, respectively), consistently outperforming features sourced from other cohorts. The steady increase in AUC as the feature count rose toward 170 suggests that the Zeller dataset contains a high density of high-information biomarkers that are conserved across different geographical populations.

The Vogtmann Heterogeneity: Conversely, the Vogtmann (USA) dataset exhibited the weakest generalization. Features sourced from this cohort resulted in near-random performance (AUC 0.5–0.6) when tested on other populations. This lack of transferability points to significant population-specific “noise” or “batch effects” inherent in the North American samples.

TABLE II

CROSS-DATASET AUC PERFORMANCE OF MODELS TRAINED ON FENG AND TESTED ON VOGTMANN, YU, AND ZELLER DATASETS

Model	Test cohort	Top10	Top30	Top50	Top70	Top90	Top110	Top130	Top150	Top170	Average
SVM	2.vogtmann	0.3790	0.3912	0.4397	0.4355	0.4645	0.4482	0.4482	0.4645	0.4642	0.4372
	3.yu	0.4068	0.3762	0.6683	0.6607	0.4982	0.5183	0.5183	0.5652	0.5652	0.5308
	4.zeller	0.4849	0.4891	0.4837	0.4225	0.4517	0.4413	0.4370	0.6468	0.6452	0.5002
RF	2.vogtmann	0.4540	0.5743	0.5668	0.5485	0.6153	0.6157	0.6197	0.6177	0.6281	0.5822
	3.yu	0.5900	0.6430	0.6265	0.6796	0.7041	0.6936	0.7070	0.7113	0.7163	0.6746
	4.zeller	0.7078	0.7216	0.7115	0.7595	0.7935	0.8012	0.8083	0.8066	0.8230	0.7703
GBC	2.vogtmann	0.4245	0.5163	0.4837	0.4885	0.5418	0.5733	0.5610	0.5812	0.5945	0.5294
	3.yu	0.5059	0.5833	0.5626	0.6429	0.6637	0.6709	0.6593	0.7317	0.7075	0.6364
	4.zeller	0.7140	0.6734	0.6834	0.7477	0.8044	0.8142	0.8066	0.7997	0.8033	0.7608

TABLE III

CROSS-DATASET AUC PERFORMANCE OF MODELS TRAINED ON VOGTMANN AND TESTED ON FENG, YU, AND ZELLER DATASETS

Model	Test cohort	Top10	Top30	Top50	Top70	Top90	Top110	Top130	Top150	Top170	Average
SVM	1.feng	0.6256	0.5751	0.5992	0.4969	0.4519	0.4093	0.5748	0.6685	0.3690	0.5300
	3.yu	0.6235	0.3541	0.3487	0.3827	0.3950	0.5381	0.7356	0.7400	0.7360	0.5393
	4.zeller	0.6364	0.6378	0.4929	0.5242	0.4686	0.5526	0.6843	0.6303	0.5849	0.5791
RF	1.feng	0.5727	0.6181	0.6374	0.7120	0.7001	0.6330	0.6694	0.7312	0.7431	0.6686
	3.yu	0.5534	0.6192	0.6282	0.6468	0.6702	0.6889	0.7134	0.7146	0.7086	0.6604
	4.zeller	0.6028	0.6781	0.7272	0.7739	0.7394	0.7364	0.7282	0.7934	0.8202	0.7333
GBC	1.feng	0.5617	0.5542	0.6612	0.6529	0.6644	0.6437	0.6935	0.7173	0.7098	0.6509
	3.yu	0.4850	0.5878	0.5435	0.5976	0.6419	0.6659	0.7330	0.7183	0.7086	0.6313
	4.zeller	0.5501	0.6306	0.7286	0.7522	0.7305	0.6937	0.6909	0.7872	0.7980	0.7069

TABLE IV

CROSS-DATASET AUC PERFORMANCE OF MODELS TRAINED ON YU AND TESTED ON FENG, VOGTMANN, AND ZELLER DATASETS

Model	Test cohort	Top10	Top30	Top50	Top70	Top90	Top110	Top130	Top150	Top170	Average
SVM	1.feng	0.4744	0.4650	0.4364	0.5135	0.5550	0.5617	0.5583	0.5548	0.5583	0.5197
	2.vogtmann	0.3665	0.5485	0.4282	0.5455	0.4600	0.4640	0.4640	0.4348	0.4602	0.4635
	4.zeller	0.6817	0.4569	0.4312	0.4945	0.4700	0.4646	0.4439	0.4646	0.4700	0.4864
RF	1.feng	0.5313	0.6371	0.7498	0.7464	0.7452	0.7424	0.7573	0.7004	0.7189	0.7032
	2.vogtmann	0.5497	0.5090	0.5783	0.5999	0.6409	0.6128	0.6347	0.5925	0.6147	0.5925
	4.zeller	0.6611	0.6726	0.7851	0.8015	0.7904	0.7890	0.7989	0.8085	0.8062	0.7681
GBC	1.feng	0.5040	0.6849	0.7198	0.7058	0.7405	0.7292	0.7388	0.7119	0.7261	0.6957
	2.vogtmann	0.4742	0.5388	0.5142	0.5092	0.4973	0.5987	0.5947	0.6352	0.5745	0.5485
	4.zeller	0.6422	0.6564	0.8004	0.8309	0.8184	0.7759	0.7479	0.8127	0.8227	0.7675

TABLE V

CROSS-DATASET AUC PERFORMANCE OF MODELS TRAINED ON ZELLER AND TESTED ON FENG, VOGTMANN, AND YU DATASETS

Model	Test cohort	Top10	Top30	Top50	Top70	Top90	Top110	Top130	Top150	Top170	Average
SVM	1.feng	0.5535	0.5044	0.5433	0.5071	0.4339	0.3967	0.3967	0.6737	0.6951	0.5227
	2.vogtmann	0.4717	0.4277	0.4277	0.4277	0.4277	0.4237	0.4237	0.4237	0.4237	0.4308
	3.yu	0.3714	0.4450	0.4589	0.4327	0.4549	0.4474	0.4093	0.4738	0.4752	0.4409
RF	1.feng	0.6133	0.5044	0.5433	0.5071	0.4339	0.3967	0.3967	0.6737	0.6951	0.5294
	2.vogtmann	0.4640	0.4277	0.4277	0.4277	0.4277	0.4237	0.4237	0.4237	0.4237	0.4299
	3.yu	0.5776	0.4450	0.4589	0.4327	0.4549	0.4474	0.4093	0.4738	0.4752	0.4639
GBC	1.feng	0.6238	0.6276	0.5565	0.5754	0.7345	0.7393	0.7330	0.7535	0.7274	0.6746
	2.vogtmann	0.4487	0.4727	0.5053	0.5213	0.4848	0.5737	0.5497	0.5660	0.5462	0.5187
	3.yu	0.5962	0.6747	0.6603	0.6513	0.6988	0.7029	0.6914	0.6763	0.6555	0.6675

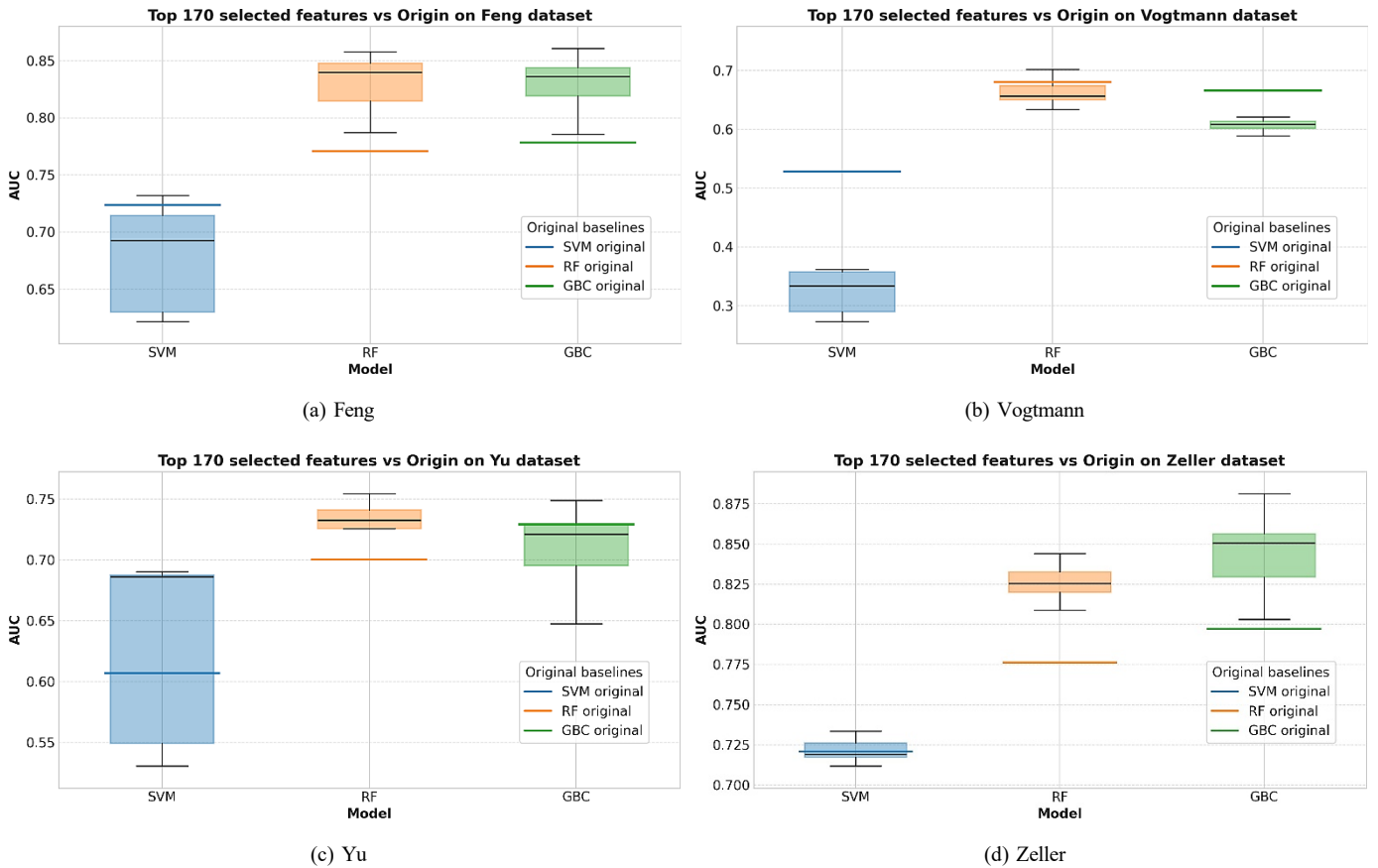


Fig. 4. Performance comparison (AUC) between models trained on the top 170 selected features and the original feature space across four colorectal cancer cohorts.

E. Comparative Analysis of Classifiers in Cross-Study Scenarios

The performance of machine learning algorithms varied significantly under the stress of cross-dataset distribution shifts:

- **Ensemble Dominance:** GBC and RF remained the most resilient classifiers. GBC, in particular, showed superior capacity in handling complex feature interactions at higher feature counts (100–170), while RF maintained high stability even with minimal feature inputs. These tree-based architectures are evidently more suited for the sparse and high-dimensional nature of cross-cohort metagenomic data.

- **SVM Instability:** SVM showed severe performance degradation in cross-dataset applications. In several instances, SVM's AUC dropped below 0.4, indicating it was highly sensitive to the distribution shifts between datasets. Without cohort-specific hyperparameter tuning or rigorous normalization, SVM struggled to maintain a stable decision boundary when features were transferred between disparate populations.

Overall, the experimental results consistently demonstrate that p-value-based feature selection, when combined with ensemble learning models, leads to improved classification

performance and robustness across both within-dataset and cross-dataset evaluation settings. Random Forest and Gradient Boosting Classifier show superior adaptability to feature space reduction and distributional shifts between cohorts. In contrast, SVM exhibits pronounced sensitivity to feature sparsity and cross-cohort discrepancies. These findings highlight the importance of dataset-aware feature selection strategies for reliable CRC metagenomic classification. The detailed numerical results corresponding to these comparisons are reported in Tables VI–IX.

TABLE VI

EVALUATION RESULTS ON THE FENG DATASET (ZELLER PERFORMS BEST, VOGTMANN SHOWS THE WEAKEST PERFORMANCE)

Dataset	Classifier	Top10	Top50	Top100	Top170	Average
Zeller	GBC	0.7140	0.6834	0.75	0.8033	0.7481
Zeller	RF	0.7078	0.7115	0.7913	0.8030	0.7225
Yu	RF	0.5900	0.6265	0.6984	0.7163	0.6757
Vogtmann	GBC	0.4245	0.4837	0.5453	0.5945	0.5294
Vogtmann	SVM	0.3790	0.4397	0.4277	0.4642	0.4359

TABLE VII

EVALUATION RESULTS ON THE YU DATASET (ZELLER CONTINUES TO DOMINATE WITH A LARGE PERFORMANCE GAP COMPARED TO VOGTMANN)

Dataset	Classifier	Top10	Top50	Top100	Top170	Average
Zeller	GBC	0.6422	0.7904	0.7926	0.8027	0.7493
Feng	GBC	0.5040	0.7198	0.7598	0.7261	0.6999
Vogtmann	RF	0.5497	0.5783	0.6207	0.6147	0.5912
Vogtmann	SVM	0.3665	0.4282	0.4560	0.4602	0.4729

TABLE VIII

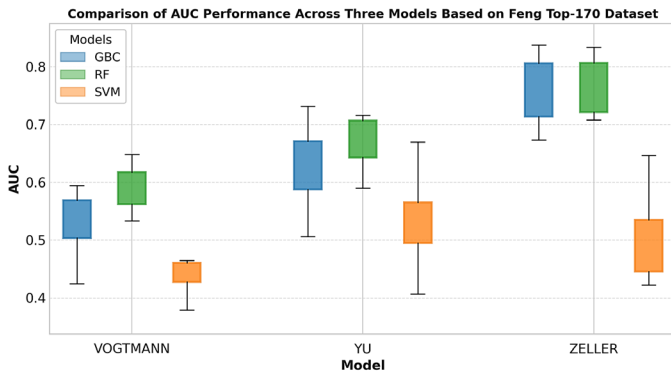
EVALUATION RESULTS ON THE ZELLER DATASET (FENG AND YU EXHIBIT COMPARABLE PERFORMANCE)

Dataset	Classifier	Top10	Top50	Top100	Top170	Average
Feng	GBC	0.6238	0.5565	0.7483	0.7274	0.6734
Yu	RF	0.5776	0.6511	0.6915	0.6979	0.6797
Vogtmann	SVM	0.4717	0.4277	0.4237	0.4237	0.4269

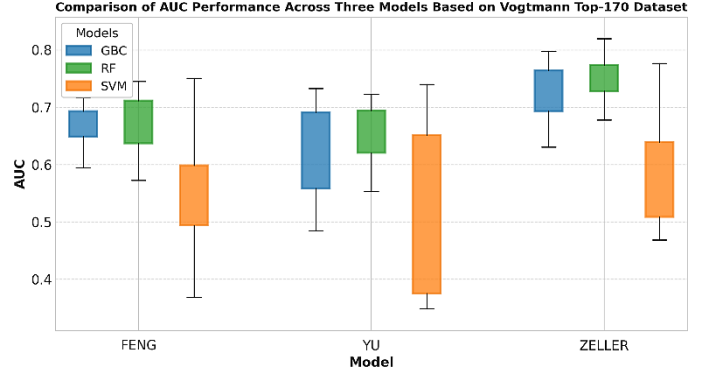
TABLE IX

EVALUATION RESULTS ON THE VOGTMANN DATASET (A CHALLENGING DATASET, YET ZELLER YIELDS RELATIVELY COMPETITIVE PERFORMANCE)

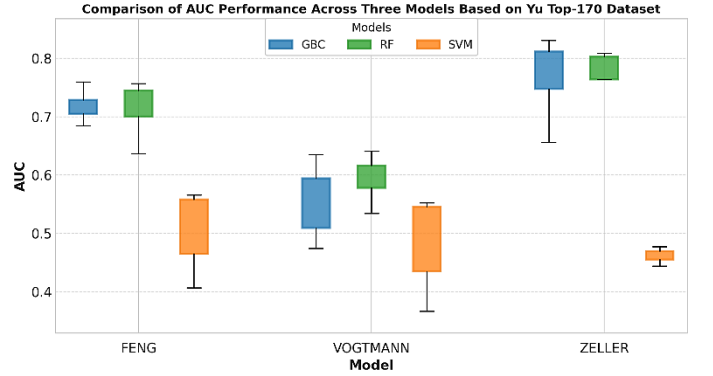
Dataset	Classifier	Top10	Top50	Top100	Top170	Average
Zeller	RF	0.6028	0.7272	0.7467	0.8202	0.7377
Feng	GBC	0.5617	0.6612	0.6590	0.7098	0.6631
Yu	SVM	0.6235	0.3487	0.3934	0.7359	0.5254



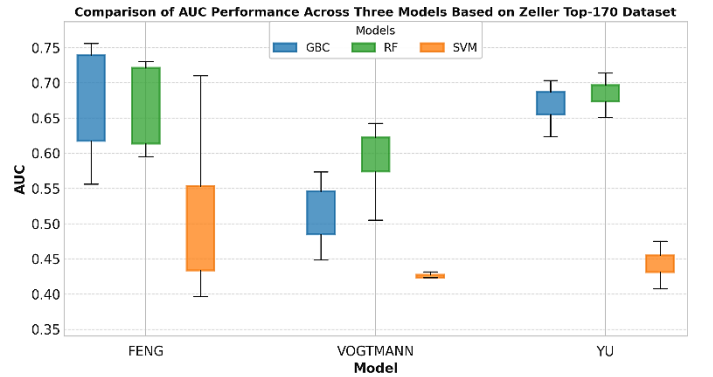
(a) Feng cohort



(b) Vogtmann cohort



(c) Yu cohort



(d) Zeller cohort

Fig. 5. Performance comparison of three classifiers (GBC, RF, and SVM) across four cohorts using the Top 170 feature subsets.

F. Discussion

The experimental results demonstrate that the Zeller cohort, representing Western populations from Germany and France, provides the most reliable reference for feature extraction, consistently achieving superior cross-dataset predictive performance. This finding is consistent with previous large-scale metagenomic studies, which show that Western cohorts often capture diverse yet reproducible microbial signatures associated with colorectal cancer (CRC), thereby improving cross-population generalizability [2], [3].

Feature subsets derived from the Feng and Yu cohorts achieve moderately high AUC values, indicating partial transferability across datasets. This observation aligns with

prior studies reporting that although Asian populations exhibit distinct gut microbiome profiles – largely shaped by traditional high-fibre, plant-based diets – there remains a shared core of CRC-associated microbial markers across regions [19]. The comparatively lower generalization performance suggests that regional dietary patterns and environmental exposures contribute to microbiome variability.

In contrast, the limited robustness of Vogtmann-derived features under cross-cohort evaluation likely reflects stronger cohort-specific effects, including demographic, dietary, and technical variability. It has been well established that gut microbiome composition is highly sensitive to geographical and lifestyle differences, particularly dietary patterns distinguishing Western and Asian populations [12], [15]. Western diets rich in fat and red meat are associated with carcinogenic metabolites such as secondary bile acids, whereas fibre-rich diets promote protective metabolites like short-chain fatty acids, thereby influencing CRC risk [12].

Across all experimental settings, ensemble-based classifiers consistently outperform Support Vector Machine, particularly under feature reduction and cross-dataset distribution shifts, confirming their robustness in high-dimensional metagenomic data analysis [14]. Overall, these findings highlight the importance of dataset-aware feature selection and support the broader consensus that geographical and dietary differences across populations significantly shape gut microbiome composition and modulate CRC risk, providing a robust framework for cross-cohort CRC prediction.

VI. CONCLUSION AND FUTURE WORK

In this study, we propose a novel classification framework for CRC based on metagenomic data collected from multiple geographic regions, including Asia (China), Europe (Austria, Germany, and France), and North America (United States). The proposed approach integrates statistical feature filtering using the T-test with machine learning-based classification, and systematically compares GBC, RF, and SVM models. The framework comprises data pre-processing, feature selection, model training, and performance evaluation, with extensive cross-cohort experiments conducted to assess robustness and generalizability.

Experimental results demonstrate that feature subsets derived from the Zeller cohort (Germany and France) consistently yield the highest predictive performance, indicating that this cohort serves as a de facto “Gold Standard” for feature extraction and cross-dataset generalization. This finding suggests the presence of transferable and biologically conserved microbial signatures across populations. Such transferability may be attributed to the cohort’s ability to capture diverse microbiome patterns shaped by Western lifestyles, which have been shown to differ significantly from those in Asian populations due to variations in dietary habits and environmental exposures. These geographical and demographic differences are known to influence gut microbiome composition and, consequently, modulate CRC risk, thereby impacting cross-cohort model performance and generalizability.

Moreover, comparative analyses reveal that GBC and RF outperform SVM across all experimental settings, particularly when trained on reduced feature subsets. These findings highlight the suitability of ensemble tree-based methods for metagenomic CRC classification. Their robustness under cross-dataset distribution shifts further suggests an advantage in handling microbiome variability arising from population-specific factors such as diet and lifestyle.

The main contribution of this work lies in the development of a p-value-driven feature selection strategy, which enables the identification of a compact yet informative subset of metagenomic features. This strategy achieves predictive performance comparable to that of the original high-dimensional feature space while substantially reducing dimensionality, computational cost, and model complexity. As a result, the proposed framework offers an efficient and effective solution for CRC diagnosis support, with strong potential for application in large-scale and heterogeneous metagenomic studies. Importantly, the findings also emphasise the need to incorporate demographic and geographical context into microbiome-based predictive modelling to ensure robust cross-population applicability.

Future work will focus on validating the proposed framework on larger and more diverse metagenomic cohorts to assess its robustness and generalizability further. In addition, integrating advanced feature selection strategies and incorporating multi-omics data may enhance predictive performance and provide deeper biological insights into CRC-associated microbial signatures. Further investigation into the interaction between diet, microbiome, and CRC across populations may also improve the interpretability and clinical relevance of such models.

DECLARATIONS

Availability of Data

The data supporting the research findings are available from the study “Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers” [20].

Competing Interests

The authors have no conflicts of interest to declare.

REFERENCES

- [1] G. Zeller *et al.*, "Potential of fecal microbiota for early-stage detection of colorectal cancer," *Molecular Systems Biology*, vol. 10, Nov. 2014. <https://doi.org/10.15252/msb.20145645>
- [2] 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, pp. 667–678, Apr. 2019. <https://doi.org/10.1038/s41591-019-0405-7>
- [3] J. Wirbel *et al.*, "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer," *Nature Medicine*, vol. 25, pp. 679–689, Apr. 2019. <https://doi.org/10.1038/s41591-019-0406-6>
- [4] Q. Su *et al.*, "Faecal microbiome-based machine learning for multi-class disease diagnosis," *Nature Communications*, vol. 13, Nov. 2022, Art. no. 6818. <https://doi.org/10.1038/s41467-022-34405-3>
- [5] S. Lee and I. Lee, "Comprehensive assessment of machine learning methods for diagnosing gastrointestinal diseases through whole metagenome sequencing data," *Gut Microbes*, vol. 16, no. 1, July 2024. <https://doi.org/10.1080/19490976.2024.2375679>
- [6] S. Weiss *et al.*, "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, Mar. 2017, Art. no. 27. <https://doi.org/10.1186/s40168-017-0237-y>
- [7] G. Papoutsoglou *et al.*, "Machine learning approaches in microbiome research: challenges and best practices," *Frontiers in Microbiology*, vol. 14, Sep. 2023. <https://doi.org/10.3389/fmicb.2023.1261889>
- [8] M. Teixeira, F. Silva, R. M. Ferreira, T. Pereira, C. Figueiredo, and H. P. Oliveira, "A review of machine learning methods for cancer characterization from microbiome data," *npj Precision Oncology*, vol. 8, May 2024, Art. no. 123. <https://doi.org/10.1038/s41698-024-00617-7>
- [9] H. Ham and T. Park, "Combining p-values from various statistical methods for microbiome data," *Frontiers in Microbiology*, vol. 13, Nov. 2022. <https://doi.org/10.3389/fmicb.2022.990870>
- [10] Q. Feng *et al.*, "Gut microbiome development along the colorectal adenoma-carcinoma sequence," *Nature Communications*, vol. 6, Mar. 2015, Art. no. 6528. <https://doi.org/10.1038/ncomms7528>
- [11] 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, Sep. 2017. <https://doi.org/10.1136/gutjnl-2015-309800>
- [12] S. J. D. O'Keefe *et al.*, "Fat, fibre and cancer risk in African Americans and rural Africans," *Nature Communications*, vol. 6, Apr. 2015, Art. no. 6342. <https://doi.org/10.1038/ncomms7342>
- [13] T. P. Quinn, I. Erb, G. Gloor, C. Notredame, M. F. Richardson, and T. M. Crowley, "A field guide for the compositional analysis of any-omics data," *GigaScience*, vol. 8, no. 9, Sep. 2019. <https://doi.org/10.1093/gigascience/giz107>
- [14] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights," *PLOS Computational Biology*, vol. 12, July 2016, Art. no. e1004977. <https://doi.org/10.1371/journal.pcbi.1004977>
- [15] L. A. David *et al.*, "Diet rapidly and reproducibly alters the human gut microbiome," *Nature*, vol. 505, pp. 559–563, Dec. 2013. <https://doi.org/10.1038/nature12820>
- [16] J. T. Nearing *et al.*, "Microbiome differential abundance methods produce different results across 38 datasets," *Nature Communications*, vol. 13, Jan. 2022, Art. no. 342. <https://doi.org/10.1038/s41467-022-28034-z>
- [17] Y. Sun *et al.*, "Benchmarking and optimizing microbiome-based bioinformatics workflows for non-invasive detection of intestinal tumors," *Microbiome Research Reports*, vol. 4, Dec. 2025, Art. no. 43. <https://doi.org/10.20517/mrr.2025.75>
- [18] P. Freitas, F. Silva, J. V. Sousa, R. M. Ferreira, C. Figueiredo, T. Pereira, and H. P. Oliveira, "Machine learning-based approaches for cancer prediction using microbiome data," *Scientific Reports*, vol. 13, July 2023, Art. no. 11821. <https://doi.org/10.1038/s41598-023-38670-0>
- [19] S. Yachida *et al.*, "Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer," *Nature Medicine*, vol. 25, pp. 968–976, June 2019. <https://doi.org/10.1038/s41591-019-0458-7>
- [20] Z. Dai *et al.*, "Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers," *Microbiome*, vol. 6, Apr. 2018. <https://doi.org/10.1186/s40168-018-0451-2>



Hien Thi Thanh Nguyen, MSc, is a Lecturer at the College of Information and Communication Technology at CanTho University, Vietnam. She graduated with a master's degree from National Central University, Taiwan. Her research interests include machine learning applications for data analysis in healthcare systems, web applications, education technology and smart devices.

E-mail: ntthien@ctu.edu.vn

ORCID iD: <https://orcid.org/0009-0007-1838-913X>



Hai Thanh Nguyen is an Associate Professor at the College of Information and Communication Technology, Can Tho University, Vietnam. He received his Master's degree in Computer Science and Engineering from National Chiao Tung University, Taiwan, and his Ph.D. degree in Computer Science from Sorbonne University, France. His current research includes healthcare systems and applications using computer vision, machine learning, and data analysis.

E-mail: nthai.cit@ctu.edu.vn

ORCID iD: <https://orcid.org/0000-0002-1386-1390>