Informatics in Medicine Unlocked 63 (2026) 101758

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

AILMENT: A novel ML framework for prediction and analysis of microbiota associations in colorectal cancer

N. Strepis [a,*,1], Z. Lu [b,1], W. de Koning [a,c], B.J.M. Rijvers [a], A.A. de Souza [d], C. Verhoef [e], B. Fosso [f], M. Doukas [a], D.E. Hilling [e,g], T. Abeel [h,i], J.P. Hays [j,1], A.P. Stubbs [a,1]

[a] Department of Pathology and Clinical Bioinformatics, Erasmus University Medical Centre (Erasmus MC), Rotterdam, the Netherlands
[b] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, the Netherlands
[c] Department of Pulmonary Medicine, Erasmus MC Cancer Institute, Erasmus University Medical Centre (Erasmus MC), Rotterdam, the Netherlands
[d] Department of Internal Medicine, Erasmus MC Transplant Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands
[e] Department of Surgical Oncology and Gastrointestinal Surgery, Erasmus MC Cancer Institute, the Netherlands
[f] Department of Biosciences, Biotechnologies and Environment, University of Bari, Bari, Italy
[g] Erasmus MC Datahub, University Medical Center Rotterdam, the Netherlands
[h] Delft Bioinformatics Lab, TU Delft, 2628 XE, Delft, the Netherlands
[i] The Netherlands University Medical Center Rotterdam, the Netherlands
[j] Department of Medical Microbiology and Infectious Diseases, Erasmus University Medical Centre (Erasmus MC), Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning framework
Colorectal cancer
Tissue microbiota
Faecal microbiota

ABSTRACT

Objective: Colorectal cancer (CRC) is one of the most common cancers in the world, with research suggesting a potential association with the human microbiota. However, simply comparing relative microbial abundances could overlook connections between microbes and specific clinical characteristics of CRC.
Methods: Here, we present the machine learning (ML) framework 'AILMENT' (AI-linked Microbiota Exploration of Nascent Tumours) that efficiently associates microbiota profiles with CRC metadata.
The Random Forest and Extreme Gradient Boosting machine learning methods incorporated in AILMENT were used to identify associations between the microbiota and CRC phenotypes relating to clinical outcomes.
Results: Sixteen ML models were generated from public data of 778 individuals using AILMENT, indicating associations between the microbiota and several different clinical characteristics of CRC, including microsatellite instability (MSI) and *BRAF* mutations (median AUROC and F1 scores of the ML models reached up to 0.90 and 0.85, respectively). Additionally, associations between *Odoribacter, Leptotrichia, Granulicella, Parvimonas, Fusobacterium* and other genera with CRC were observed. With respect to sample type, distinct microbial compositions were observed between tissue and faecal samples, indicating fundamental differences in microbiota composition between these sample types. The AILMENT framework pinpointed an association between pathogens such as *Porphyromonas* and *Parvimonas* and CRC, confirming their role as microbial signatures in the disease. Moreover, the framework could indicate microbes linked to a healthy gut distinct from the CRC state, such as the butyrate-producers *Lactobacillus, Eubacterium and Ruminococcus.* To validate the performance and utility of AILMENT, we applied it to a publicly available dataset of bacterial species abundance and associated metadata, successfully replicating the key findings.
Conclusion: The AILMENT framework can efficiently predict associations between different clinical characteristics of CRC and complex microbial relative abundance data. AILMENT enables the identification of specific microbes at the genus level for detailed clinical characterisation of CRC, demonstrating its potential as a tool for a better understanding of cancer-microbiota interactions.

1. Introduction

Cancer is one of the major public health concerns throughout the world. Although new developments in cancer research and new therapies have increased the survival rate from cancer in the last few decades, there remains significant global morbidity and mortality associated with

* Corresponding author.
[1] Authors contributed equally to the work.

https://doi.org/10.1016/j.imu.2026.101758
Received 15 December 2025; Received in revised form 31 March 2026; Accepted 12 April 2026
Available online 15 April 2026

cancer [1]. Colorectal cancer (CRC) - consisting of colon adenocarcinoma and rectum adenocarcinoma, is one of the most common cancers globally, having been reported as the leading cause of cancer-related mortality worldwide [2–4]. The 5-year survival rate for CRC patients is ~90% when diagnosed early, versus 13% at later stages, emphasising the value of early detection [5].

Cancer-based diseases can result from various causes [6,7]. However, in recent years, the microbiota has been increasingly recognised as playing an important role in the development and promotion of cancer through inflammation, infection, and diet [8–10]. For example, commensal microbiota could influence cancer via the production of metabolites and inflammatory genotoxins [11]. In this respect, scientists are currently investigating microbiota-based therapies for a range of cancers, including prostate, pancreatic, and lung cancer [12–14], showing that microbes could be both novel treatment targets and a means to enhance existing cancer therapies. While the link between the human microbiota and CRC is well-studied, its role in specific aspects, such as the formation of genetic mutations and the effect on tumour location, remains unclear.

In CRC-microbiota research, faecal samples and tumour/tumour environment tissue samples are widely used, with faecal samples having the advantage that they can be collected in a non-invasive way, offering a convenientfor CRC diagnosis [15]. However, tissue samples are more effective in identifying specific microbial involvement [16,17]. For example, CRC patients can be distinguished from healthy controls using mucosal bacterial profiles, which are more accurately determined using tissue, rather than faecal, samples for microbial identification [18]. Technologies such as omics studies and artificial intelligence (AI) can be a source for studying the impact of the human microbiota on CRC. In humans, the Human Genome Project [19] has generated vast amounts of omics data for analysis. For the microbiota, metagenomics analyses have been used to investigate the composition and functions of microbial communities, focusing on genes, coding regions, and reference genomes [20]. A subset of techniques, e.g., 16*S* rRNA sequencing, uses meta-genotyping to examine taxonomic diversity [21]. Meanwhile, metatranscriptomics explores microbial RNA transcripts, providing insights into microbiota gene expression [22]. However, the effective use of these complex data for CRC diagnosis and treatment remains limited, although AI approaches could play a crucial role in CRC disease diagnostics and biomarker discovery. For example, AI has been used with high-throughput proteomic technology to explore cancer biomarkers [23,24].

Machine Learning (ML), a branch of AI, has become increasingly important in cancer research, including cancer diagnosis, therapy, and microbiota analysis [25–27]. The discovery of links between gut microbiota and diseases such as CRC has driven the development of ML tools for microbiota research. These tools enable deeper insights into complex microbial data, supporting tasks such as identifying microbial associations and signatures [28–30]. Popular ML models, including Logistic Regression, Support Vector Machines, Random Forest (RF), and Extreme Gradient Boosting (XGB), were used in gut microbiota studies, with RF and XGB demonstrating high accuracy and robustness when applied to different microbial datasets [27,31–34]. Their success highlights the potential of ML to advance cancer microbiota research. During the current research, an ML framework called AILMENT (AI-Linked Microbiota Exploration of Nascent Tumours) was built to predict and analyse clinical information from CRC patients, including MSI status, tumour stage, tumour location, B-Raf proto-oncogene, serine/threonine kinase *BRAF* status, tumour protein P53 *TP53* status, and tissue sample types, and associated relative abundance microbiota data. ML outcomes identified important microbiota genera that are potentially associated with specific clinical characteristics of CRC. AILMENT is an effective ML framework that uses microbial abundances from metatranscriptomics and metagenomics to predict cancer characteristics, providing another step towards diagnosis and treatment decisions using AI.

2. Methods

2.1. Data description

In this study, two previously published datasets relating to CRC and microbiota research were used (Table 1). The first dataset comprised human transcriptomics data obtained from 162 Singapore colorectal cancer (CRC) patients (SG-Bulk), which was generated from collected colon tissue samples [35]. Corresponding human gene expression data and metadata for these 162 CRC patients were also included in the SG-Bulk dataset, which was also used to extend the application of the AILMENT framework to identify human genes significantly associated with CRC in these tissues.

The second dataset is shotgun metagenomics sequencing data of faecal samples from a cohort of 616 participants (Faecal-MG), including 365 CRC patients and 251 healthy individuals [29]. In this dataset, 8367 species and 1941 genera were previously identified using a data filtering pipeline. Relative abundances of each microbiota species were calculated, and the sum of the relative abundance for each sample was calculated. The relative abundance of each genus was calculated as the sum of all species belonging to a specific genus.

The framework's performance was validated using an independent dataset of shotgun metagenomic sequences from a cohort of 1262 samples (Table 2). For this validation, genus-level abundances were calculated by aggregating the MetaPhlAn3-derived species-level relative abundances for the 191 microbes reported in the source study [36].

2.2. Data preprocessing

The transcriptomics data from SG-Bulk were processed through quality trimming and filtering with Fastp (version 0.23.4) [37,38]. Host depletion was performed by mapping the reads to the human reference genome using BWA-MEM2 (version 2.2.1) [39]. Taxonomic classification was conducted with Kraken2 (version 2.1.3) and the standard database [40], and the relative microbial abundances were re-estimated using Bracken (version 2.9) [41]. The relative abundance of each species in each sample was calculated by dividing the read count for each species by the total number of reads in the sample. To refine the dataset, we selected bacterial taxa present in more than 50% of samples. Further filtering was applied using a “blacklist” of potential microbial contaminants present in laboratory reagents [42]. The relative abundance of each genus was then calculated by summing the abundances of all species belonging to that genus. Samples with missing clinical diagnoses were removed from SG-Bulk.

The Faecal-MG dataset was normalised at the genus level using the same method that was used for the SG-Bulk dataset; therefore, we directly selected microbes present in more than 50% of samples. Microbes in the “blacklist” were also removed from the Faecal-MG dataset [42]. Additionally, unclassified species at the genus level were discarded to ensure data quality. After processing, a total of 449 species from 225 genera were included for further analysis. As with SG-Bulk, the relative abundance of a genus was calculated by summing all the species belonging to that genus. To ensure label consistency of clinical traits, we relabelled tumour stages into “Early stage”, “Late stage” and “Others”, and tumour location into “Left colon”, “Right colon” and “Multi-site”

Table 1
Overview of datasets included in this study.

Datasets	No. Samples	Data Types	Sample Types	Sources
SG-Bulk	162	Human meta-transcriptome	Tissue samples	[35]
Faecal-MG	616	Human metagenome	Faecal samples	[29]
Validation	1262	Human metagenome	Faecal samples	[36]

Table 2
A comparative summary of the three datasets analysed in this study: the internal SG-Bulk and Faecal-MG cohorts, and the large-scale Validation dataset (n = 1262).

	SG-Bulk	Faecal-MG	Validation
Age Group, n (%)			
18-64	73 (45.06%)	322 (52.27%)	689 (54.60%)
≥65	86 (53.09%)	294 (47.73%)	573 (45.40%)
NA	3 (1.85%)	-	-
Sex, n (%)			
Male	94 (58.02%)	358 (58.12%)	749 (59.35%)
Female	65 (40.12%)	258 (41.88%)	513 (40.65%)
NA	3 (1.85%)	-	-
Disease Status, n (%)			
Healthy	0 (0%)	251 (40.75%)	662 (52.46%)
CRC	162 (100%)	365 (59.25%)	600 (47.54%)
Tumor Stage (CRC), n (%)			
Early Stage	60 (37.04%)	111 (30.41%)	12 (2%)
Late Stage	99 (61.11%)	74 (20.27%)	18 (3%)
Pre-invasive stages	-	180 (49.32%)	-
NA	3 (1.85%)	-	1232(97.6%)
Tumor Location, n (%)			
Left Colon	107 (66.05%)	167 (27.11%)	75 (5.94%)
Right Colon	51 (31.48%)	83 (13.47%)	11 (0.87%)
Multi-Site		8 (1.30%)	-
Non-specified/ NA	4 (2.47%)	358 (58.12%)	9 (7.13%)
No. Microbes after filtering			
Genera Level	212	258	230
Species Level	-	-	859
Primary Reference	Singapore EGAD00001008512	Japan	PRJEB7774, PRJNA531273, PRJNA447983, PRJDB4176, PRJEB12449, PRJEB27928, PRJDB4176, PRJEB10878, and PRJEB6070

based on clinical classes. Samples with missing clinical diagnoses were removed from Faecal-MG.

The validation dataset, derived from 1262 samples (Table 2), was preprocessed to ensure data quality for the analysis, as were the other databases. Genus-level relative abundances were first calculated by summing the MetaPhlAn3-derived abundances of all constituent species. Following this, a prevalence filter was applied, and genera present in less than 50% of the samples were removed from the dataset. To further refine the data, any taxa that remained unclassified at the genus level were also discarded. Ultimately, a total of 230 genera were retained for the final validation analysis. To ensure label consistency of clinical traits, we relabelled tumour stages into "Early stage", "Late stage" and "Others", and tumour location into "Left colon", "Right colon" and "Multi-site" based on clinical classes. Samples with missing clinical diagnoses were removed from the validation dataset.

2.3. Assessing complexity

Principal Coordinates Analysis (PCA), alpha diversity analysis using the Shannon index, and beta diversity analysis using Bray-Curtis dissimilarity were applied to each pre-processed dataset in R (v.4.4.0).

2.4. AILMENT framework

The AILMENT framework was designed to provide a simple framework for feature selection of microbiota and human genes that are predictive for classifying cancer patients based on diagnostic (e.g. MSI status) or clinical outcome. The AILMENT framework is composed of four stages: (1) Machine learning Model construction, (2) Model Performance and Evaluation, and (3) Integrative Feature Analysis, (4) Validation Data Set Analysis as outlined below and in Fig. 1.

1. Model Training: The entire modelling process was embedded within a bootstrap framework consisting of 20 iterations to ensure robust performance evaluation. Specifically, to prevent data leakage and ensure class balance, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training dataset during each iteration. Using the Python (v.3.11.7) SMOTE function, synthetic samples were generated for the minority class based on the 3-nearest neighbours of existing samples, ensuring that the test sets remained entirely independent and representative of real-world class distributions. Each iteration began with a random shuffle of the dataset, followed by an 80:20 split into training and test data. All preprocessing steps, including feature selection, class balancing, data scaling, 3-fold cross-validation and hyperparameter tuning, were applied strictly to the training partition. The Random Forest (RF) model and Extreme Gradient Boosting (XGB) models were trained on the training data in Python (v.3.11.7) using the functions "*RandomForestClassifier*" and "*xgb.XGBClassifier*", respectively. For hyperparameter selection, we performed a grid search with the function "*GridSearchCV*", which was based on the accuracy of predictions. The grid search included a 3-fold stratified cross-validation via the function "*StratifiedKFold*" in Python (v.3.11.7), ensuring that model selection remained independent of the held-out test set. For RF models, hyperparameters included the number of trees, and the maximum number of features that yielded the best split at each node, being fine-tuned by the grid search. For XGB models, hyperparameters including learning rate, maximum depth of the tree, and size of subsample were fine-tuned by the grid search. Final performance metrics were calculated solely on the 20% test data, which remained entirely unseen by the models during the training and hyperparameter tuning phases. A detailed schematic and algorithmic pseudocode of the full AILMENT pipeline, illustrating the strict isolation of the independent test set during preprocessing and model selection, are provided in the Supplementary Information (Supplementary Pseudocode).
2. *Model Evaluation and Assessment of Model Robustness*: The optimised hyperparameters were applied to the RF and XGB models to generate predictions from the test data. To evaluate performance, accuracy, precision, recall, and F1 score were calculated for the test set (Equations (1)–(4)). The area under the receiver operating characteristic curve (AUROC) was also used to assess the discriminative ability of the RF and XGB models, evaluating their performance in distinguishing between classes across all classification thresholds. An AUROC close to 1.0 indicates strong discriminative ability, while an AUROC of 0.5 suggests performance no better than random chance. For multi-class classification models, indicators of the overall performance were macro-averaging of precision, recall, F1 score, and AUROC (Equation (5)).

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

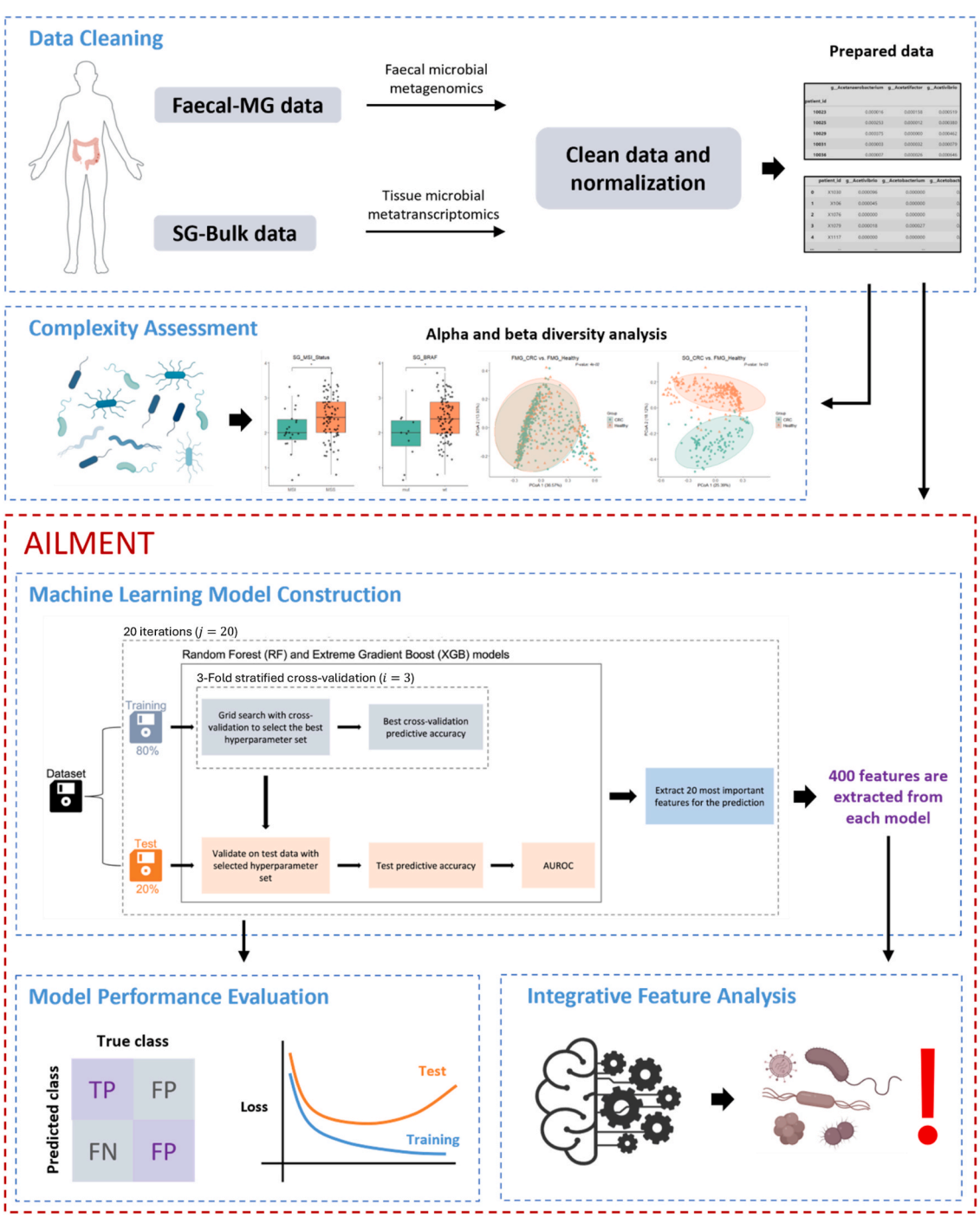

Fig. 1. General scheme of the AILMENT framework. All three datasets contain microbiota relative abundance and metadata. SG-Bulk has an extra data matrix that includes human gene expression profiles. Details of the three datasets and ML framework are described in Methods.

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

$$MaA_X = \frac{\sum_{i=1}^{n} X_i}{n} \tag{5}$$

Equations (1)–(5). TP, TN, FN, and FP represent true positives, true negatives, false negatives, and false positives, respectively. The F1 score, ranging from 0 to 1, is the harmonic mean of precision and recall, with higher values indicating better performance [43]. Macro-average (MaAX) values, including precision, recall, F1 score, and AUROC, were calculated across all classes, where n denotes the number of classes in the model. Although a 3-fold stratified cross-validation was performed during the training process, the split of training data and test data remained identical. To ensure a robust assessment of our framework, we repeated the entire modelling process 20 times (j = 20) on different random samples of the data. In each of these iterations, the data was split into a training and a test set. The model training process itself involved a grid search to optimise hyperparameters, using 3-fold stratified cross-validation (i = 3) on the training data only. The performance

of the resulting tuned model was then evaluated on the corresponding test set, with final metrics averaged across all 20 iterations (Fig. 3). Therefore, there were 60 runs of either the RF or XGB model in the framework ($i \times j = 60$). The performance metrics from cross-validation and testing across 20 iterations were extracted from a box plot in Python (v.3.11.7) and compared to see if there was a significant fluctuation along iterations and if the test accuracy was significantly lower than the training accuracy with respect to overfitting.

3. *Integrative Feature Importance Analysis*: In ML models, 'feature importance' indicates the weight of each feature in a prediction. Therefore, we utilised the built-in function' feature importance' with default settings, in both RF and XGB models, to extract the 20 most important genera from each model (Fig. 3). Each iteration of the AILMENT framework extracted 400 important genera. To ensure the selection of only the most robust biological markers, we implemented a stability-selection protocol requiring a consensus

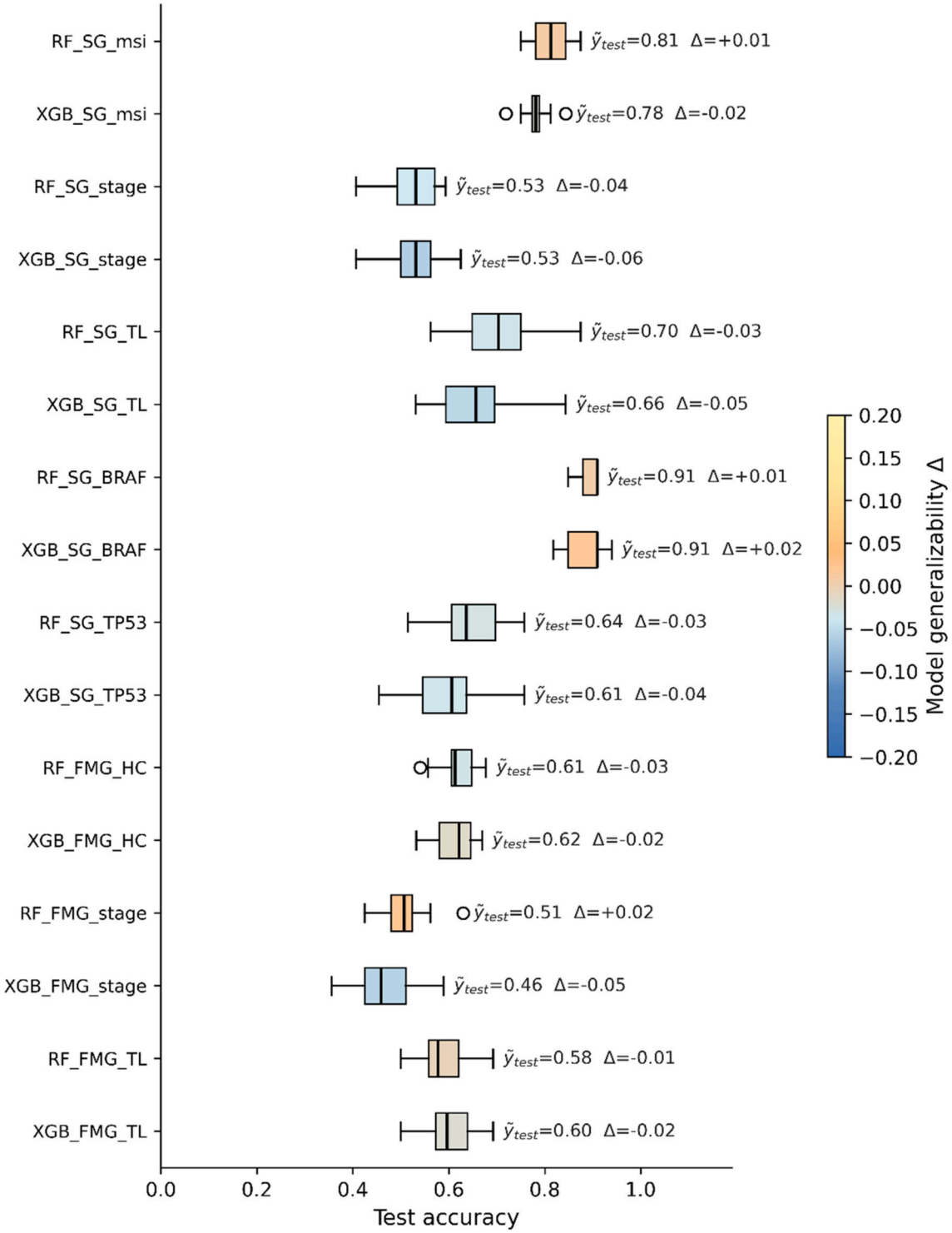

Fig. 2. *General classification performance of RF and XGB models within our AILMENT framework using test accuracy (y_test) and the gap between median test accuracy and median training accuracy (Δ) for model overfitting assessment. SG and FMG refer to datasets SG-Bulk and Faecal-MG, respectively. TL, stage, BRAF, TP53, HC, and MSI refer to predictive clinical characteristics of CRC, relating to classes of tumour location, tumour stage, BRAF status, TP53 status, healthy condition, and MSI status, respectively.*

frequency of $\geq = 40$. This threshold represents a 100% consensus requirement across all 20 bootstrap iterations and both ensemble models (20 iterations $\times$ 2 models $= 40$). This stringent criterion ensures that a genus is only retained if it is consistently prioritized regardless of the data partition or the specific optimisation strategy (bagging-based vs. boosting-based). Features meeting this 100% consensus were identified as highly involved microbes, while those prioritized by both RF and XGB engines were classified as critically involved signatures for specific CRC clinical characteristics.

4. *Validation Data Set Analysis:* For cross-validating the AILMENT, we implemented the validation dataset in the framework with the same process and settings. The framework's performance was evaluated by correlating genus-level abundances with patient age, health condition, tumour stage, and tumour location on an independent dataset of 1262 CRC samples. Age-specific correlation was also evaluated at the species level with our framework.

3. Results

3.1. Assessing the complexity of microbiome datasets

To investigate the diversity and complexity of the datasets studied, we performed PCA, alpha diversity, and beta diversity calculations on each dataset based on clinical characteristics that included MSI status, tumour stage, tumour location, etc. In the PCA analysis, a significant difference was observed only between microbial profiles from tumour tissue and faecal samples Supplementary Fig. 1, while no separation was observed for other characteristics (Supplementary Fig. 1). Alpha diversity analysis, however, revealed substantial differences among groups for several clinical characteristics (Supplementary Fig. 3; $P < 0.05$). Similarly, beta diversity demonstrated significant variation between groups for three clinical characteristics across both datasets (Supplementary Fig. 4; $p < 0.05$). Notably, healthy participants in the Faecal-MG dataset exhibited significantly higher microbial alpha diversity compared to CRC patients in the SG-Bulk dataset (Supplementary Fig. 3c; $P < 0.001$), with beta diversity also showing significant differences (Supplementary Fig. 4i; $P = 0.001$). All other clinical characteristics did not display significant differences between microbial groups from either alpha diversity or beta diversity (Supplementary Figs. 3–4; $p > 0.05$). Together, these analyses highlighted the complexity of the datasets studied, suggesting that advanced ML methods are best suited to uncovering the complex patterns potentially hidden in these datasets.

3.2. Robust prediction of clinical characteristics in CRC

To explore associations between complex microbial data and CRC, we built 16 ML models in the AILMENT framework, with the resultant predictive accuracy and model generalizability being calculated to evaluate the predictive performance of these models (Fig. 2). We observed no significant overfitting problem in our ML models when combining cross-validation with the resampling of training and test data, shown in Fig. 2. Further, AUROC and F1 scores did not show significant differences in predictive performance between RF and XGB models (Supplementary Table 2). These results suggested that AILMENT can robustly predict the clinical characteristics of CRC without overfitting the training data.

For SG-Bulk patients, the performance metrics of MSI status based on tissue microbial relative abundance data were 0.81 and 0.78 using the RF and XGB models, respectively (Fig. 2; Supplementary Table 2); and the median AUROC of these two models were approximately 0.71 (Supplementary Table 2). These findings indicated considerable associations between tissue microbiota and MSI status in CRC. Also, predictions of *BRAF* status demonstrated good performance almost reaching values above 0.9 for accuracy and F1-score, suggesting profound associations between the microbial composition in the tissues of CRC and *BRAF* status. On the other hand, *TP53* status showed fewer associations with the microbial composition in the tissues of CRC (Fig. 2). The median accuracy values for the prediction of tumour stage and tumour location in SG-Bulk patients ranged from 0.53 to 0.7 (Fig. 2; Supplementary Table 2), revealing associations between tissue microbiota and tumour stage or tumour location in CRC. The AUROC predictions on tumour location from SG-Bulk data and RF and XGB models were 0.65 and 0.7 respectively, suggesting the models can distinguish between the classes (Fig. 2; Supplementary Table 2).

Compared to the SG-Bulk dataset, the median accuracy values of tumour stage and tumour location based on the microbial relative abundance of the Faecal-MG dataset were generally lower, ranging from 0.45 to 0.60 (Fig. 2). This indicated much weaker potential associations between the gut microbiota and tumour stage or tumour location. As the median AUROC of Faecal-MG data was higher than 0.61 in both RF and XGB models, the faecal microbiota in healthy people and CRC patients was notably different (Supplementary Table 2).

As AUROC can occasionally present an overly optimistic evaluation of model performance when applied to imbalanced datasets, we also calculated precision, recall, and F1 scores to ensure a more rigorous and comprehensive assessment of the models (Supplementary Table 2). The precision and recall scores of models of tissue samples were generally higher than 0.6, compared to those of faecal samples, ranging from 0.45 to 0.62 (Fig. 2). The F1 score, as a balance between precision and recall, provided a statistical indicator of the general performance of the ML models in AILMENT (Supplementary Table 2). The median values for precision and recall, and the F1 score for each ML model, across 20 iterations, are provided in Supplementary Table 2, and the representative confusion matrix output is available in Supplementary Fig. 5. The significance between predictive performances was determined by the P-value from a two-tailed *t*-test between individual predictive performances, where the null hypothesis was that two independent predictive performances were equal.

Together with accuracy and AUROC, we showed that microbiota composition may be significantly associated with CRC, indicating the potential usefulness of AILMENT in making predictions about specific clinical characteristics and the human microbiota in CRC patients.

3.3. Identification of microbial involvement in CRC

To identify microbial genera that are associated with CRC, we selected the 20 most important genera from each of the 16 models from each iteration of AILMENT, which resulted in 400 important genera. Of those features (genera), we kept important features that were present in the identification to a value of $\geq$40, whereby microbes presenting in both RF and XGB models were identified as critically involved microbes for the specific clinical characteristics of CRC. These results were used to generate a heat map that included the identified genera and their relative feature importance within each ML model (Fig. 3). Next, we compared the genera identified using the AILMENT framework with several previous CRC microbiota studies [29,30,44–49], and found that many of the ML-identified genera could also be reported in the previously published studies, including *Lactobacillus, Selenomonas, Leptotrichia, Ruminococcus, Parvimonas, Fusobacterium Peptostreptococcus, Acetobacterium, Lachnospira, Anaerostripes* and others (Fig. 3). Here, AILMENT demonstrates its robustness by validating features consistently identified in previous studies, confirming its reliability, while also uncovering potentially overlooked microbes that may have been underrepresented in prior research. This dual capability highlights AILMENT's ability to both reinforce established findings and provide novel biological insights. Notably, while certain genera such as *Parvimonas* and *Peptostreptococcus* served as robust, cross-niche biomarkers, the tissue-specific models identified a more diverse array of high-importance features, such as *Leptotrichia* and *Acetobacterium* for MSI status, that were largely absent in the feacal signal (Fig. 3). This divergence underscores the framework's ability to resolve niche-specific signatures that may otherwise be obscured by the high volume of

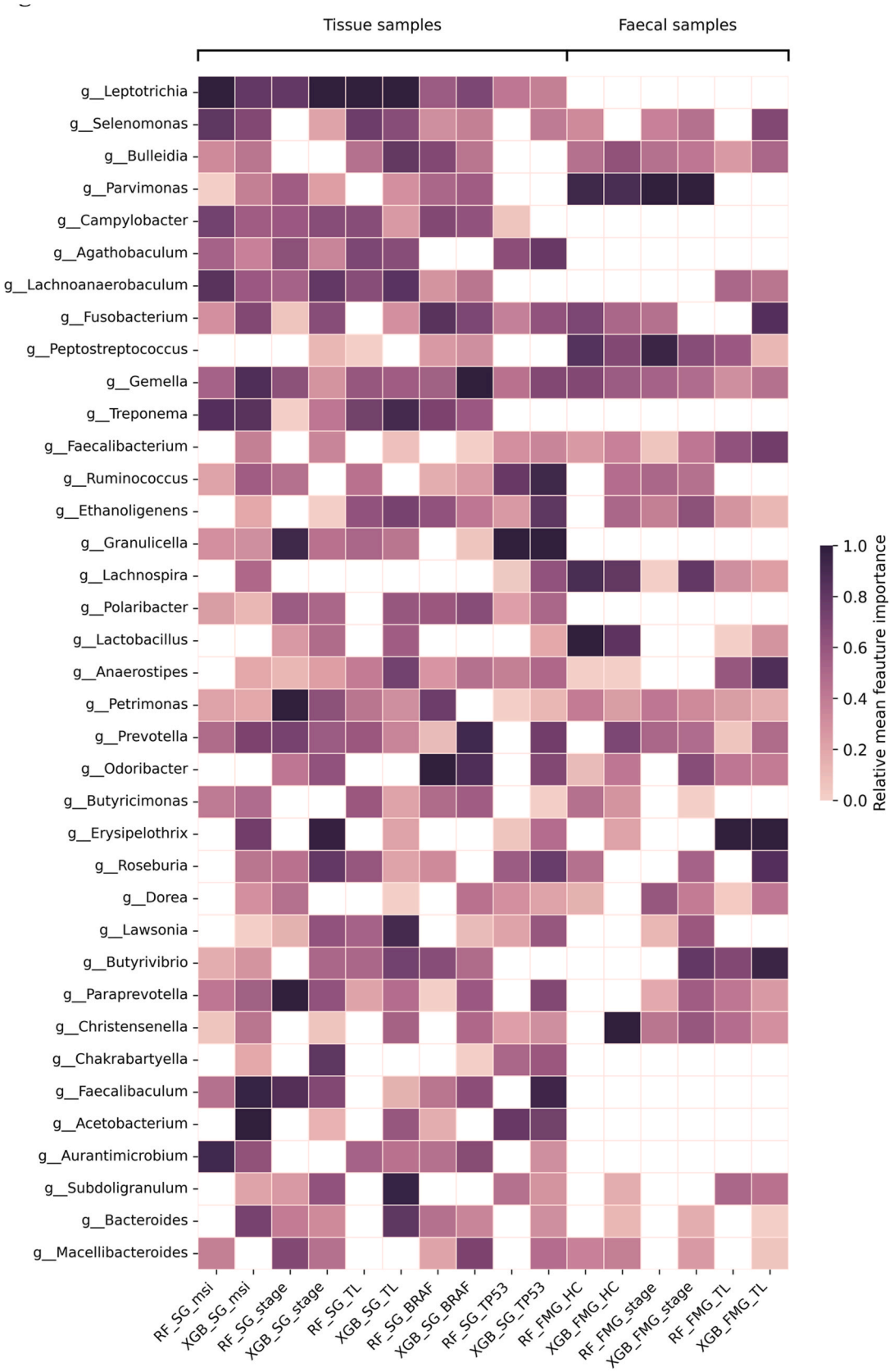

Fig. 3. Integrative feature analysis from AILMENT. The X-axis represents ML models predicting clinical information of CRC research participants. The Y-axis represents the genera identified from the integrative feature analysis. The cells were coloured according to their frequency in the framework, where a deeper colour refers to a higher frequency - indicating a connection between the identified genera and CRC. ML models were labelled into three groups: 1) tissue samples, 2) faecal samples, depending on their sample sources.

gut-derived microbiota in feacal samples.

In addition, certain genera were also found to be associated with specific clinical characteristics of CRC, including MSI status, tumour stage, tumour location, *BRAF,* and *TP53* status, potentially indicating that those may play a role in the development of CRC (Fig. 4). For example, 20 genera were identified as being important features of MSI status in both RF and XGB models (Fig. 4a). Additionally, *Leptotrichia* and *Treponema* had relative mean feature importance of >0.8, which implied these two genera could be significantly involved in changes of MSI status in CRC. For the tumour stage, 29 genera such as *Leptotrichia*, *Petrimonas*, *Paraprevotella*, and others were identified as important features in both models from tissue samples (Fig. 4b), while only 10 genera, including *Parvimonas* and *Peptostreptococcus*, were identified in both models from faecal samples. In faecal samples, *Parvimonas* displayed the highest relative importance, reaching a median value of 1.0.

Moreover, we found that the microbial composition from tumour tissue samples was distinct compared to that from faecal samples for the tumour stage (Fig. 4b and g). This implies that the tissue and faecal microbiota could have different influences on the tumour stage, or that tissue-specific microbiota may be masked by the sheer volume of gut–derived microbiota i.e., the faecal microbiota only partially reflects the mucosal microbiota. For tumour location, 21 genera were identified as important features in both models from tissue samples (Fig. 4c), whereas 14 genera were identified from faecal samples (Fig. 4h). The genus *Leptotrichia* showed the highest relative importance from tissue samples, while *Erysipelothrix* showed an association with tumour location from faecal samples (median ~ 1.0). This potentially indicated that these genera could be involved in changing the tissue microenvironment at different tumour locations. Also, we observed significantly different microbial compositions for two important genetic mutations, where

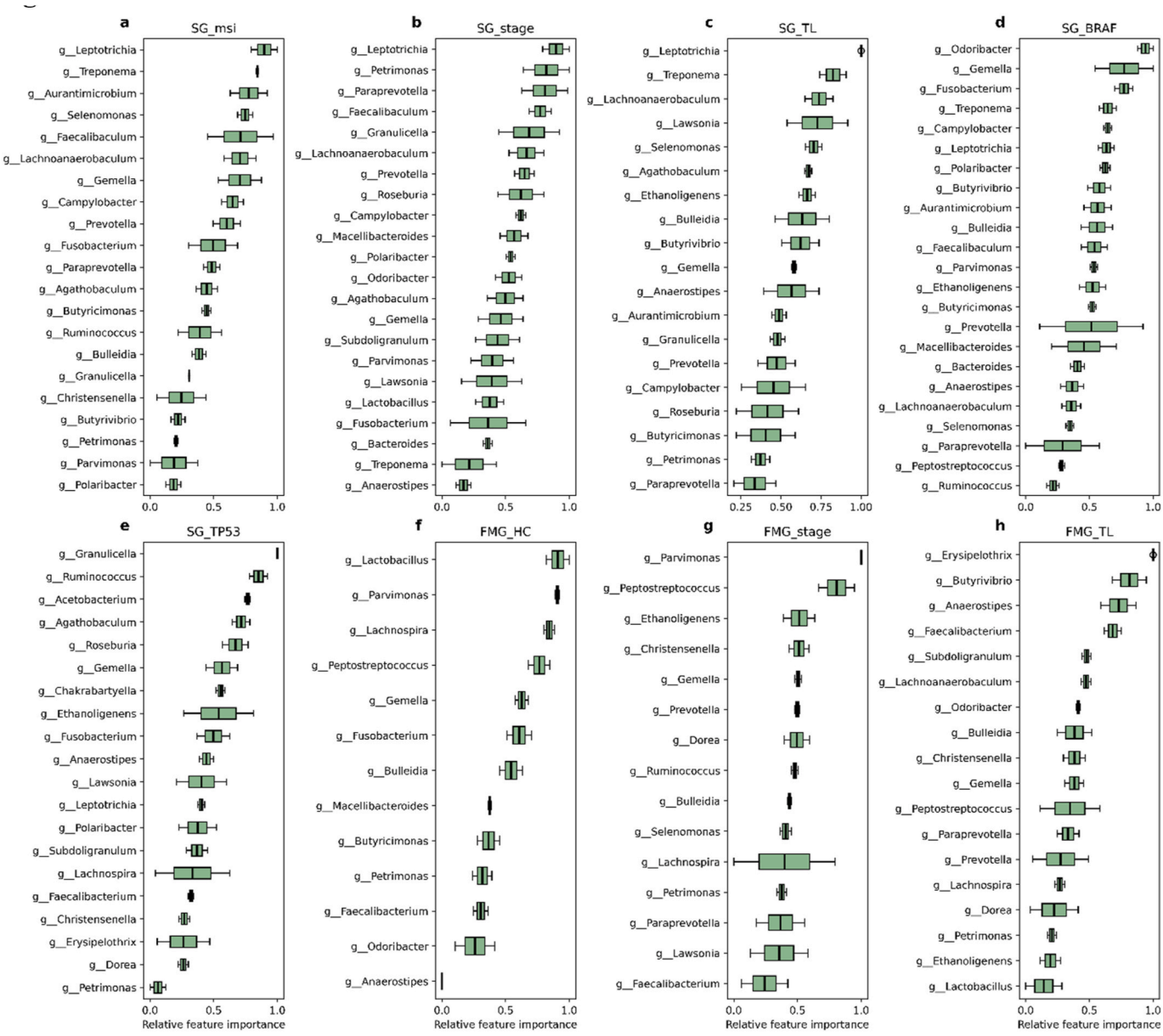

Fig. 4. Genera that were associated with different clinical characteristics of CRC using both RF and XGB models. Boxes were created for each clinical characteristic based on relative mean feature importance values from both models. a: Stably identified genera in MSI status from SG-Bulk. b: Stably identified genera in tumour stage from SG-Bulk. c: Stably identified genera in tumour location from SG-Bulk. d: Stably identified genera in *BRAF* status from SG-Bulk. e: Stably identified genera in *TP53* status from SG-Bulk. f: Stably identified genera in healthy conditions from Faecal-MG. g: Stably identified genera in tumour stage from Faecal-MG. h: Stably identified genera in tumour location from Faecal-MG.

Odoribacter was identified as the leading genus associated with *BRAF* status and *Granulicella* as dominant for *TP53* status (Fig. 4d and e). Additionally, several genera were associated with both genetic mutations, including *Selenomonas, Fusobacterium,* and *Leptotrichia.*

For the healthy condition, several well-known differential genera, including *Lactobacillus*, *Parvimonas*, *Lachnospira*, *Peptostreptococcus*, *Gemella*, and *Fusobacterium*, were identified among 16 genera that potentially distinguished between healthy individuals and CRC patients using faecal samples (Fig. 4f). In addition, we identified 19 genera that were important for distinguishing healthy individuals using faecal samples or tissue samples from CRC patients (Fig. 4i). Genera present in both conditions, such as *Eubacterium*, *Phocaeicola, Ruminococcus, Clostridium* and *Pseudobutyrivibrio* may indicate an association with the absence of CRC.

External validation on an independent 1262-sample dataset confirmed the AILMENT framework's robust predictive performance and biomarker discovery capabilities (Figs. 5–7). The framework excelled at predicting multiple clinical characteristics, most notably tumour stage (median AUROC ~ 0.69, F1-score ~ 0.70), as well as patient age and tumour location (median AUROC > 0.80). In the age-prediction task, when analysing the data at the species level, the framework identified *Bifidobacterium dentium* as the most important and stable predictive species, reaching a median importance value of 1.0 (Fig. 7b). Significantly, this finding corroborates the results of the original study [36], demonstrating the framework's ability to successfully replicate published work at high resolution. This validation on an external dataset, combining high predictive accuracy with the replication of known species-level findings, confirms that AILMENT is a reliable analytical tool for microbiome research.

Essentially, our findings highlight that our analytic method identifies several dominant genus signatures associated with CRC. However, it should be noted that focusing on genera reduces the sensitivity of the findings compared to the majority of publications that associate CRC with specific bacterial species. Consequently, while our analysis identified key genera such as *Bacteroides*, *Escherichia*, *Streptococcus*, and *Enterococcus*, the data did not allow for the discrimination of specific species or strains previously cited in CRC literature, such as *B. fragilis*, *E. coli*, *S. gallolyticus*, or *E. faecalis*. However, Bacteroides spp. (including *B. fragilis*) and *E. coli* are also present in the majority of healthy individuals, with enterotoxigenic producing and non-enterotoxin-producing strains of *B. fragilis* and genotoxic colibactin (pks+) strains of *E. coli* being present in the human gut [50]. Further, recent research indicates that CRC-associated *B. fragilis* may be more significantly infected with *Caudoviricetes* prophages. It is possible that the distribution of these genotoxin and/or prophage-carrying bacteria may vary within different CRC cohorts [51]. For *S. gallolyticus*, significant genomic diversity has been shown in CRC strains, and although *E. faecalis* produces biliverdin (promoting angiogenesis and cell proliferation) and reactive oxygen species (ROS), different strains of these bacteria exist, and their role in CRC may be controversial [52].

4. Discussion

Research suggests that the human microbiota plays a key role in CRC initiation and progression, with alterations in gut or tissue-specific microbial composition linked to inflammation and genetic changes [46]. However, microbiota profiles may vary across populations due to variations in extrinsic factors (e.g. lifestyle), or intrinsic factors (e.g. immune repertoire), which complicate the identification of specific microbes or groups of microbes involved in CRC. Additionally, traditional statistical methods such as PCA, alpha diversity, and beta diversity analyses may fail to capture the complexity of microbiota data, as indicated in this study (Supplementary Figs. 1–4). The development of advanced supervised ML models may be better suited to analyse such complex data, potentially identifying novel key microbes or human genes involved in CRC. However, ML applications in this field are still emerging. Therefore, we developed AILMENT, an ML framework designed to identify key microbes (genera) against a range of CRC clinical characteristics, including MSI status, *BRAF* and *TP53* mutations, tumour stage, tumour location, and tissue sample type. MSI, is a key factor in DNA replication and repair that influences genetic mutation risks and intratumoural microbiota diversity, affecting the tumour microenvironment [35, 53–58]. However, only a small proportion of CRC are related to MSI [59] Genetic mutations in *BRAF* and *TP53* also play significant roles in CRC,

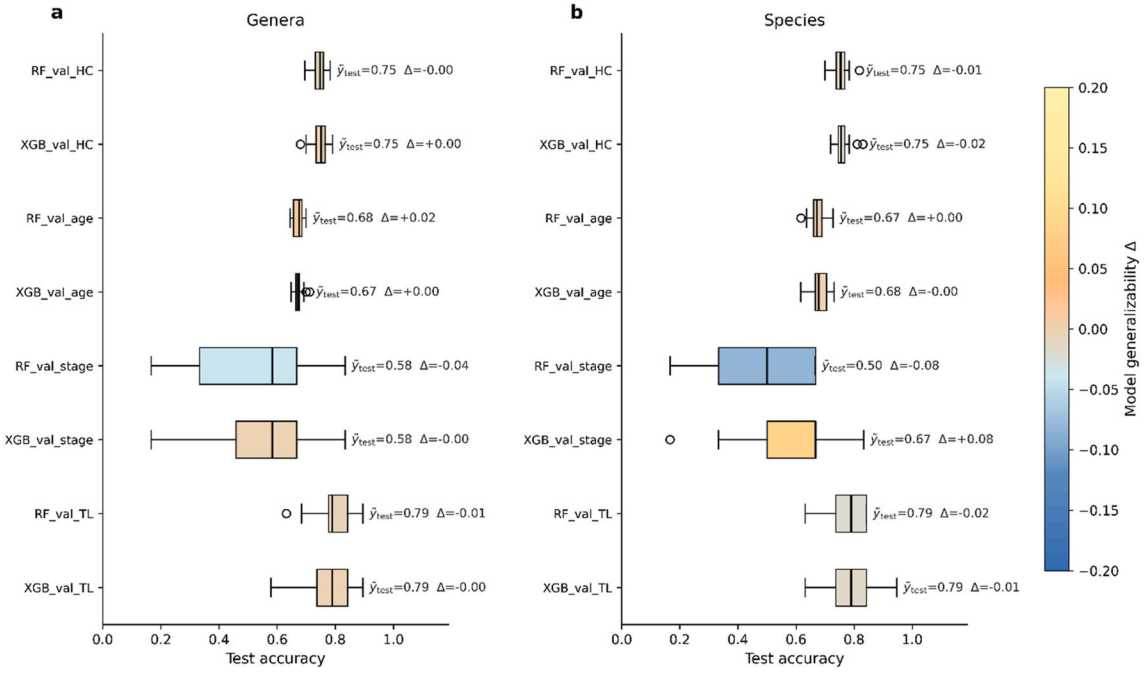

Fig. 5. Predictive performance of the RF and XGB models on the validation dataset from either genera or species level. Model performance was evaluated for the prediction of healthy controls (HC), age, tumour stage, and tumour location using multiple standard metrics. Test accuracy (y_test) was used to assess model predictive performance and the gap between median test accuracy and median training accuracy (Δ) was used for model overfitting assessment.

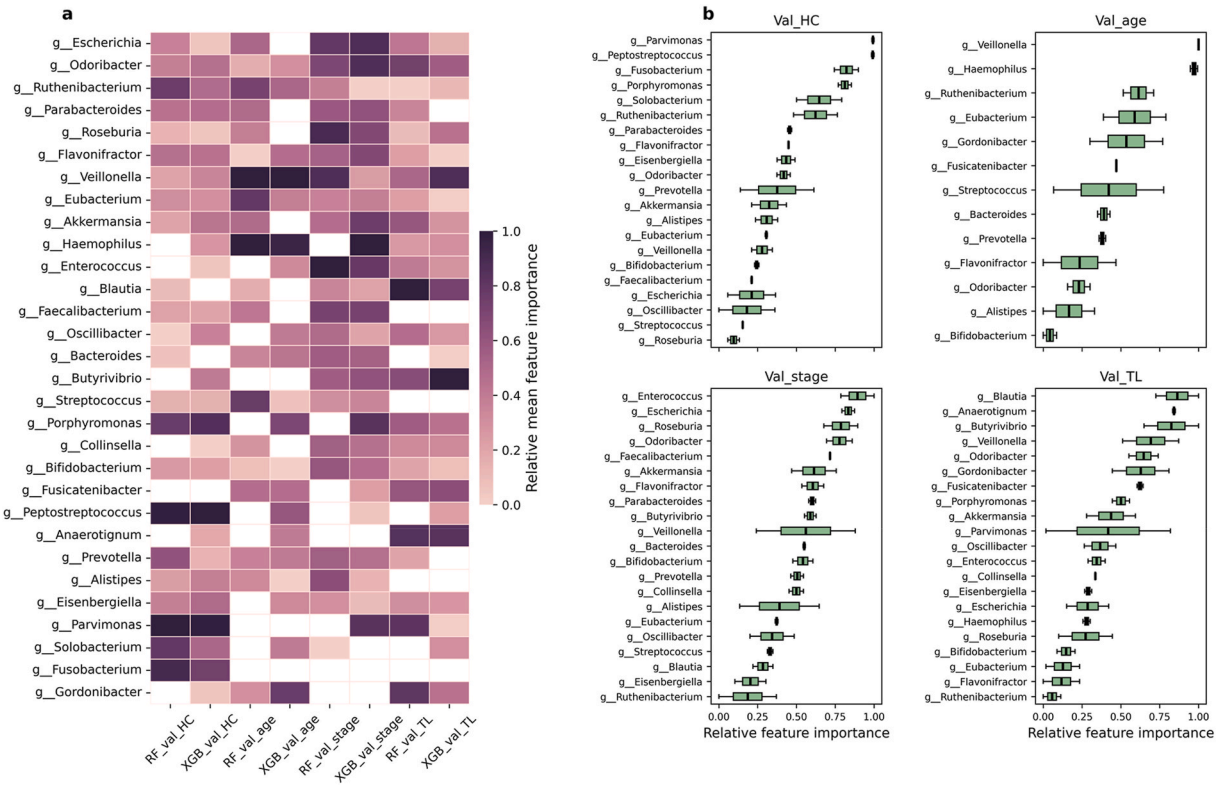

Fig. 6. Feature importance of bacterial genera for the prediction of patient age in the validation dataset. The relative mean feature importance values from both the RF and XGB models were used to identify genera predictive of healthy conditions, age, tumour stages, and tumour locations. a: Heatmap illustrating the relative mean feature importance of all genera identified by the models. b: Boxplot showing the distribution of relative feature importance scores for the top genera predictive of patient healthy conditions (HC), age, tumour stages, and tumour locations (TL).

while tumour stage and tumour location may affect the prognosis of disease progression [59]. Because these characteristics dictate both disease prognosis and therapeutic efficacy, the ability of AILMENT to robustly link specific microbial signatures to these distinct genetic and clinical profiles offers significant potential for non-invasive patient stratification.

Traditional methods for identifying microbes involved in CRC often rely on differences in relative abundance, which may not directly relate to CRC [46,60]. In contrast, our ML framework, AILMENT, employs an integrative feature importance analysis. This approach identified numerous genera consistently linked to CRC, including *Fusobacterium*, *Parvimonas*, and *Peptostreptococcus* (Fig. 3). AILMENT's primary strength, however, is its ability to move beyond these known associations to uncover potentially overlooked microbes linked to specific clinical and genetic characteristics of CRC (Fig. 4). The AILMENT framework successfully identified a strong association between the genera *Porphyromonas* and *Parvimonas* and the CRC tumour microenvironment (Figs. 3 and 4). This data-driven finding is particularly compelling as it corroborates a growing body of clinical research that has implicated these two common oral bacteria in the development and progression of CRC. Studies consistently show a higher abundance of these microorganisms within colorectal tumours and have directly linked their presence to disease progression and poorer patient survival rates. Crucially, research has noted the frequent co-occurrence of *Porphyromonas* and *Parvimonas* within tumours, suggesting a potential synergistic relationship. AILMENT's ability to independently pinpoint genera with relevance for CRC underscores its power to extract complex, actionable signatures from biological data and indicate the critical link between disease [61–63].

Beyond just identifying disease markers, the AILMENT framework demonstrated its sensitivity by pinpointing several genera associated with a healthy gut, effectively distinguishing them from the CRC state. It successfully highlighted potential beneficial microbes, including *Eubacterium*, *Ruminococcus*, and specific clusters of *Clostridium* [64,65]. These bacteria are cornerstones of a balanced microbiome, renowned for their crucial role in fermenting dietary fibre to produce anti-inflammatory compounds like butyrate. AILMENT's ability to detect these positive, health-associated signals confirms its capacity to analyse the full spectrum of microbial states, not just signatures of disease.

During integrative feature importance analysis, we used a threshold of 40 (i.e., 20 x 2), which was determined by the number of iterations and ML models used in the study. However, it should be noted that changing this threshold could lead to different identification outputs. Further, the inclusion of more data and ML models into AILMENT results in a more expansive set of identified microbes associated with the same threshold (Fig. 7a and c). This indicates that the microbial association is more complex at the species level. Our requirement for 100% consensus across 40 independent bootstrap runs and two distinct ensemble algorithms is a highly conservative approach. While this stringency successfully limits false positives and ensures the identification of highly stable biomarkers, an adaptive thresholding method could be explored in future iterations to improve sensitivity for rarer microbial signatures.

In this study, we observed that the predictive performance from tissue samples was generally higher than that from faecal samples (Fig. 2). It was also noticeable that there were many fewer faecal sample-specific genera identified compared to tissue sample-specific genera (Supplementary Table 3). This finding can be supported by the conclusion from previous research that faecal microbiota only partially reflect the mucosal microbiota from which tissue biopsies are collected [19]. Such findings further enhanced our understanding of the differences in microbiota composition between faecal samples and tumour tissue samples in cancer microbiota research.

A key finding from the external validation was the framework's predictive performance, especially for identifying tumour stage, patient

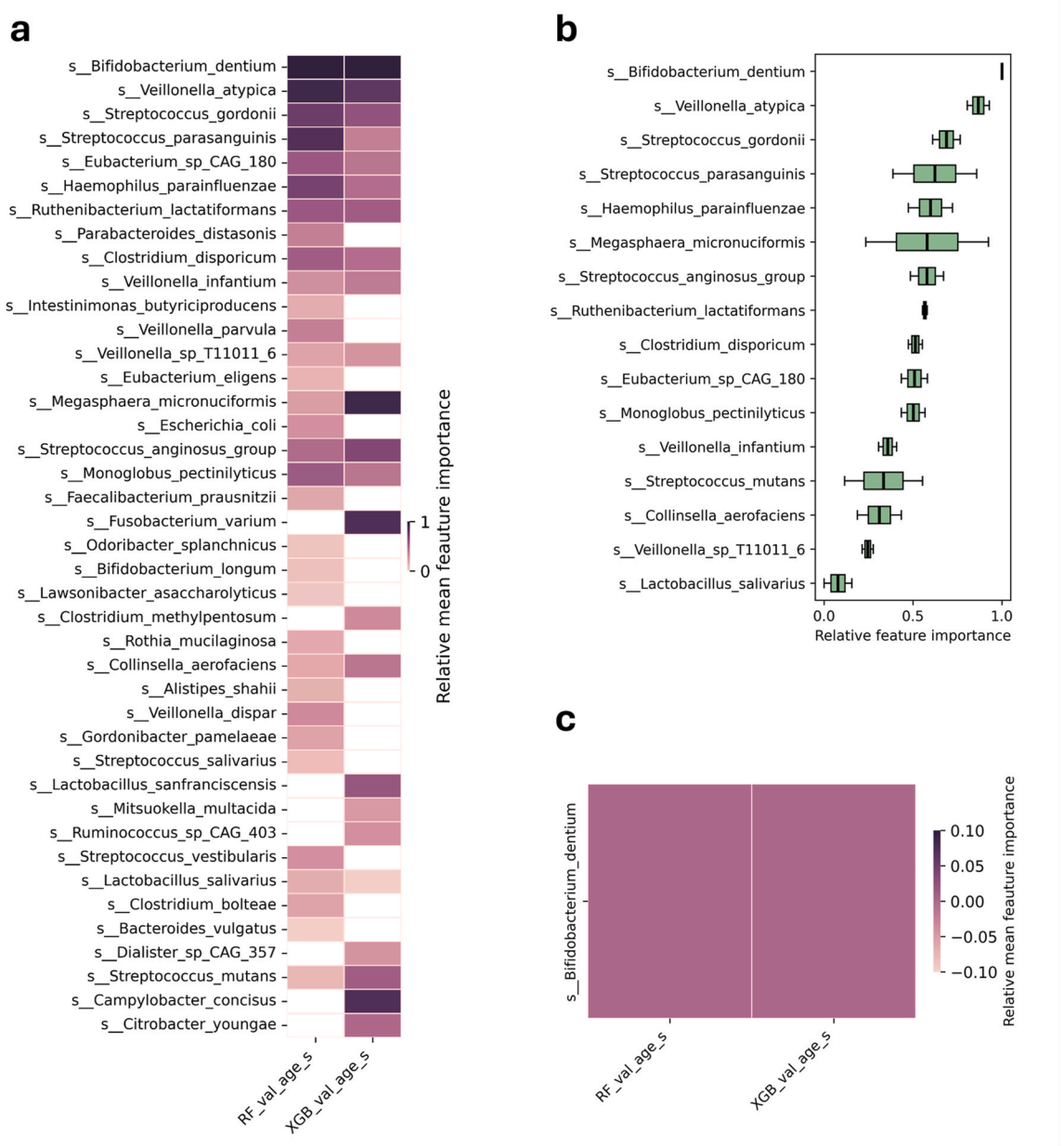

Fig. 7. Feature importance of bacterial species for the prediction of patient age in the validation dataset. The relative mean feature importance values from both the RF and XGB models were used to identify species predictive of age. a: Heatmap illustrating the relative mean feature importance of all species identified by the models. b: Boxplot showing the distribution of relative feature importance scores for the top species predictive of patient age. c: Detailed view of the relative mean feature importance for *Bifidobacterium dentium*, the top-ranked species associated with patient age.

age, and tumour location. However, the predictive performance for tumour stage (median AUROC~0.69) must be interpreted with caution. These results must be interpreted in the context of the original data structure; while AILMENT utilizes SMOTE to balance training sets, the held-out test sets remain representative of the original, imbalanced real-world distributions. For clinical characteristics where the minority class is significantly underrepresented, this data sparsity in the evaluation partition can lead to increased sensitivity and lower median AUROC values. Furthermore, these results generated the largest error bars, suggesting model instability, which is likely a direct consequence of the small sample size available for this specific task (~50 samples), resulting in extensive missing metadata in the validation set. Additionally, a consistent trend was observed where cross-validation accuracy was slightly higher than the performance on the final test set for all predictive tasks. Ultimately, despite these specific considerations, the AILMENT framework's predictive accuracy across multiple tasks and its ability to replicate findings from the original study were confirmed in the external validation analysis.

The high dimensionality of microbiome data, where combinatorial interactions create a vast feature space, poses significant risks of model instability and overfitting, a common challenge in ML where models fail to generalise beyond the training set. To mitigate this, the AILMENT framework utilizes ensemble learning techniques (RF and XGB) rather than single-estimator models. These engines, recently identified by the MiDx (2025) [66] benchmark as superior for CRC detection, effectively capture the synergistic interactions that linear tools like SIAMCAT (LASSO-based) [67] may overlook. Furthermore, AILMENT ensures numerical robustness through a 100% consensus protocol requiring signatures to be consistently identified across 40 independent bootstrap runs. This differentiates the framework from deep learning approaches

like DeepMicro (autoencoder-based) [68], which can be sensitive to stochastic instability or the ill-conditioned optimisation landscapes often encountered in high-feature-to-sample ratios. By utilising a dual-algorithm consensus between bagging-based and boosting-based engines, the AILMENT framework reduces the risk of entrapment in local minima, as distinct optimisation trajectories are unlikely to converge on the same sub-optimal solution. Beyond algorithmic design, AILMENT's primary novelty is the systematic integration of multi-modal clinical metadata, addressing a critical limitation in current literature. This approach is reinforced by a multi-stage validation on an independent cohort of 1262 samples, demonstrating that the captured microbial signatures remain predictive across diverse populations and are not idiosyncratic to the discovery set. By combining these rigorous pre-processing steps with an ensemble consensus, AILMENT offers a potentially robust and interpretable perspective on the tumour-microbiome interface, seeking to maintain model stability within the inherent constraints of high-dimensional biological data.

Despite the robustness of the AILMENT framework, we recognise that cross-cohort variations in sample type (tissue vs. faecal) and patient demographics can introduce confounding effects. By maintaining stratified models for distinct sample types, we ensure that identified associations are specific to their respective biological niches. Furthermore, the framework's ability to generate stable predictions at the individual sample level, validated across an external cohort, suggests that the identified microbial signatures are conserved biological features of CRC progression rather than artefacts of study-specific batches. However, we acknowledge the lack of granular metadata such as BMI, sex, and dietary habits in certain public datasets as a limitation. While the current study prioritises the integration of primary clinical traits like MSI status and tumour stage, future prospective studies with comprehensive demographic tracking will be essential to further refine the specificity of these microbial markers.

AILMENT performed with higher prediction and accuracy when tumour stage and location were compared between tissue samples and faecal samples, likely due to associations between tissue (mucosal) microbiota and resultant tumour microenvironments (Fig. 2). However, the performance of AILMENT will require further evaluation using larger datasets in order to further improve its accuracy. While AILMENT models were generally robust, predictions for tumour stage and *TP53* status from the SG-Bulk dataset showed instability, possibly due to a small sample size or limited hyperparameter optimisation. While typical ML techniques are powerful, their 'black box' nature can make results difficult to interpret without extensive bioinformatics expertise [69,70]. AILMENT addresses this barrier by providing clear, rank-ordered feature importance (explainable AI), offering researchers a more transparent and interpretable tool for biomarker discovery.

5. Conclusions

In conclusion, the AILMENT framework was developed to identify statistical associations between the gut microbiota and diverse clinical characteristics of CRC patients across multiple discovery datasets. Our results demonstrate high predictive performance and feature stability, which were further confirmed through external validation on an independent 1262-sample dataset. In this validation, the framework successfully identified significant correlations between microbial abundance and patient clinical profiles, including age. AILMENT identified genera previously associated with CRC in the literature, demonstrating its capacity to integrate and analyse both metatranscriptomics and metagenomics data. While the framework provides potentially novel insights into the relationship between the microbiota and specific clinical characteristics of CRC, these findings should be viewed as a platform for biological hypothesis generation. As AILMENT also identified associations between several human genes and CRC, it represents a versatile tool for multi-omics biomarker discovery; however, prospective, multi-centre clinical studies are mandatory to validate these signatures before any diagnostic or clinical application can be considered.

CRediT authorship contribution statement

N. Strepis: Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **Z. Lu:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **W. de Koning:** Writing – review & editing, Methodology. **B.J.M. Rijvers:** Writing – review & editing, Formal analysis. **A.A. de Souza:** Writing – review & editing, Methodology. **C. Verhoef:** Writing – review & editing, Resources. **B. Fosso:** Writing – review & editing, Investigation. **M. Doukas:** Writing – review & editing, Data curation. **D.E. Hilling:** Writing – review & editing, Data curation. **T. Abeel:** Writing – review & editing, Supervision. **J.P. Hays:** Writing – review & editing, Supervision, Methodology, Conceptualization. **A.P. Stubbs:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Availability of data and code

All code can be found on GitHub: https://github.com/ErasmusMC-Bioinformatics/AILMENT_ML_GUT.

The datasets generated and analysed during the current study are available from the European Genome Archive (EGA) with the codes EGAS00001005978 for SG-Bulk and EGAS00001005978 for Faecal-MG.

Ethics declarations

This study did not receive nor require ethics approval, as it reused the publicly available data.

Ethics declarations

This study did not receive nor require ethics approval, as it reused the publicly available data.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Google Gemini and ChatGPT to search for relevant literature, make corrections in the grammar of human-written text, and improve the clarity of writing. After using this tool/service, Nikolaos Strepis and Zhongyuan Lu reviewed and edited the content as needed and took full responsibility for the content of the publication.

Funding

This research is supported by the Convergence Health and Technology funding from Erasmus MC, Delft and Erasmus University (https://convergence.nl/).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our sincere gratitude to the Department of Pathology and Clinical Bioinformatics at the Erasmus University Medical Center (Erasmus MC), Rotterdam, the Netherlands, for providing resources and an environment conducive to this research. We are grateful to the members contributing to this research and their feedback

on the development and implementation of AILMENT and their comments on the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2026.101758.

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72:7–33. https://doi.org/10.3322/caac.21708.

[2] Xia C, Dong X, Li H, Cao M, Sun D, He S, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (Engl) 2022;135:584–90. https://doi.org/10.1097/cm9.0000000000002108.

[3] Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin 2020;70:145–64. https://doi.org/10.3322/caac.21601.

[4] Siegel RL, Torre LA, Soerjomataram I, Hayes RB, Bray F, Weber TK, et al. Global patterns and trends in colorectal cancer incidence in young adults. Gut 2019;68:2179–85. https://doi.org/10.1136/gutjnl-2019-319511.

[5] Cai P, Xiong J, Sha H, Dai X, Lu J. Tumor bacterial markers diagnose the initiation and four stages of colorectal cancer. Front Cell Infect Microbiol 2023;13. https://doi.org/10.3389/fcimb.2023.1123544.

[6] Ponder BAJ. Cancer genetics. Nature 2001;411:336–41. https://doi.org/10.1038/35077207.

[7] Lewandowska A, Rudzki M, Rudzki S, Lewandowski T, Laskowska B. Environmental risk factors for cancer – review paper. Ann Agric Environ Med 2019;26:1–7. https://doi.org/10.26444/aaem/94299.

[8] Singh N, Baby D, Rajguru J, Patil P, Thakkannavar S, Pujari V. Inflammation and cancer. Ann Afr Med 2019;18:121. https://doi.org/10.4103/aam.aam_56_18.

[9] Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome and human cancer. Science 1979;2021:371. https://doi.org/10.1126/science.abc4552.

[10] Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer 2013;13:800–12. https://doi.org/10.1038/nrc3610.

[11] Bhatt AP, Redinbo MR, Bultman SJ. The role of the microbiome in cancer development and therapy. CA Cancer J Clin 2017;67:326–44. https://doi.org/10.3322/caac.21398.

[12] Kustrimovic N, Bombelli R, Baci D, Mortara L. Microbiome and prostate cancer: a novel target for prevention and treatment. Int J Mol Sci 2023;24:1511. https://doi.org/10.3390/ijms24021511.

[13] Bangolo AI, Trivedi C, Jani I, Pender S, Khalid H, Alqinai B, et al. Impact of gut microbiome in the development and treatment of pancreatic cancer: newer insights. World J Gastroenterol 2023;29:3984–98. https://doi.org/10.3748/wjg.v29.i25.3984.

[14] Sun Y, Wen M, Liu Y, Wang Y, Jing P, Gu Z, et al. The human microbiome: a promising target for lung cancer treatment. Front Immunol 2023;14. https://doi.org/10.3389/fimmu.2023.1091165.

[15] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 2014;10. https://doi.org/10.15252/msb.20145645.

[16] Rezasoltani S, Dabiri H, Asadzadeh-Aghdaei H, Akhavan Sepahi A, Modarressi MH, Nazemalhosseini-Mojarad E. The gut microflora assay in patients with colorectal cancer: in feces or tissue samples? Iran J Microbiol 2019. https://doi.org/10.18502/ijm.v11i1.696.

[17] Valciukiene J, Strupas K, Poskus T. Tissue vs. fecal-derived bacterial dysbiosis in precancerous colorectal lesions: a systematic review. Cancers (Basel) 2023;15:1602. https://doi.org/10.3390/cancers15051602.

[18] Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ, et al. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut 2016;66:633–43. https://doi.org/10.1136/gutjnl-2015-309595.

[19] Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45. https://doi.org/10.1038/nature03001.

[20] New FN, Brito IL. What is metagenomics teaching Us, and what is missed? Annu Rev Microbiol 2020;74:117–35. https://doi.org/10.1146/annurev-micro-012520-072314.

[21] Morgan XC, Huttenhower C. Chapter 12: human microbiome analysis. PLoS Comput Biol 2012;8:e1002808. https://doi.org/10.1371/journal.pcbi.1002808.

[22] Bikel S, Valdez-Lara A, Cornejo-Granados F, Rico K, Canizales-Quinteros S, Soberón X, et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. Comput Struct Biotechnol J 2015;13:390–401. https://doi.org/10.1016/j.csbj.2015.06.001.

[23] Tai M-T. The impact of artificial intelligence on human society and bioethics. Tzu Chi Med J 2020;32:339. https://doi.org/10.4103/tcmj.tcmj_71_20.

[24] Xiao Q, Zhang F, Xu L, Yue L, Kon OL, Zhu Y, et al. High-throughput proteomics and AI for cancer biomarker discovery. Adv Drug Deliv Rev 2021;176:113844. https://doi.org/10.1016/j.addr.2021.113844.

[25] Elemento O, Leslie C, Lundin J, Tourassi G. Artificial intelligence in cancer research, diagnosis and therapy. Nat Rev Cancer 2021;21:747–52. https://doi.org/10.1038/s41568-021-00399-1.

[26] Mitsala A, Tsalikidis C, Pitiakoudis M, Simopoulos C, Tsaroucha AK. Artificial intelligence in colorectal cancer screening, diagnosis and treatment. A new era. Curr Oncol 2021;28:1581–607. https://doi.org/10.3390/curroncol28030149.

[27] Loganathan T, Priya Doss CG. The influence of machine learning technologies in gut microbiome research and cancer studies - a review. Life Sci 2022;311:121118. https://doi.org/10.1016/j.lfs.2022.121118.

[28] Huang K, Duan J, Wang R, Ying H, Feng Q, Zhu B, et al. Landscape of gut microbiota and metabolites and their interaction in comorbid heart failure and depressive symptoms: a random forest analysis study. mSystems 2023;8. https://doi.org/10.1128/msystems.00515-23.

[29] Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 2019;25:968–76. https://doi.org/10.1038/s41591-019-0458-7.

[30] Wu Y, Jiao N, Zhu R, Zhang Y, Wu D, Wang A-J, et al. Identification of microbial markers across populations in early detection of colorectal cancer. Nat Commun 2021;12. https://doi.org/10.1038/s41467-021-23265-y.

[31] Parmar A, Katariya R, Patel V. A review on random forest: an ensemble classifier. Lecture Notes on Data Engineering and Communications Technologies 2019;26:758–63. https://doi.org/10.1007/978-3-030-03146-6_86.

[32] Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. A framework for effective application of machine learning to microbiome-based classification problems. mBio 2020;11. https://doi.org/10.1128/mbio.00434-20.

[33] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining 2016. 13-17-August-2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[34] Novielli P, Romano D, Magarelli M, Bitonto P Di, Diacono D, Chiatante A, et al. Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification. Front Microbiol 2024;15. https://doi.org/10.3389/fmicb.2024.1348974.

[35] Joanito I, Wirapati P, Zhao N, Nawaz Z, Yeo G, Lee F, et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nat Genet 2022;54:963–75. https://doi.org/10.1038/s41588-022-01100-4.

[36] Qin Y, Tong X, Mei WJ, Cheng Y, Zou Y, Han K, et al. Consistent signatures in the human gut microbiome of old- and young-onset colorectal cancer. Nat Commun 2024;15(1):3396. https://doi.org/10.1038/s41467-024-47523-x. 2024;15.

[37] Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560.

[38] Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2023;2. https://doi.org/10.1002/imt2.107.

[39] Md V, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE international parallel and distributed processing symposium (IPDPS). 2019. p. 314–24. https://doi.org/10.1109/IPDPS.2019.00041.

[40] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20. https://doi.org/10.1186/s13059-019-1891-0.

[41] Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci 2017;3:e104. https://doi.org/10.7717/peerj-cs.104.

[42] Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;12. https://doi.org/10.1186/s12915-014-0087-z.

[43] Opitz J, Burst S. Macro F1 and macro F1. 2021.

[44] Mouradov D, Greenfield P, Li S, In E-J, Storey C, Sakthianandeswaren A, et al. Oncomicrobial community profiling identifies clinicomolecular and prognostic subtypes of colorectal cancer. Gastroenterology 2023;165:104–20. https://doi.org/10.1053/j.gastro.2023.03.205.

[45] Cao C, Yue S, Lu A, Liang C. Host-gut microbiota metabolic interactions and their role in precision diagnosis and treatment of gastrointestinal cancers. Pharmacol Res 2024;207:107321. https://doi.org/10.1016/j.phrs.2024.107321.

[46] Liang Y, Zhang Q, Yu J, Hu W, Xu S, Xiao Y, et al. Tumour-associated and non-tumour-associated bacteria co-abundance groups in colorectal cancer. BMC Microbiol 2024;24. https://doi.org/10.1186/s12866-024-03402-5.

[47] Gao R, Wu C, Zhu Y, Kong C, Zhu Y, Gao Y, et al. Integrated analysis of colorectal cancer reveals cross-cohort gut microbial signatures and associated serum metabolites. Gastroenterology 2022;163:1024–1037.e9. https://doi.org/10.1053/j.gastro.2022.06.069.

[48] Loftus M, Hassouneh SA-D, Yooseph S. Bacterial community structure alterations within the colorectal cancer gut microbiome. BMC Microbiol 2021;21. https://doi.org/10.1186/s12866-021-02153-x.

[49] Yang Y, Han Z, Gao Z, Chen J, Song C, Xu J, et al. Metagenomic and targeted metabolomic analyses reveal distinct phenotypes of the gut microbiota in patients with colorectal cancer and type 2 diabetes mellitus. Chin Med J (Engl) 2023;136:2847–56. https://doi.org/10.1097/cm9.0000000000002421.

[50] Yu Y, Zhao W, Yang M, Wu B, Yuan X. Tumor-promoting gut microbes in colorectal cancer: mechanisms and translational perspectives. Int J Med Sci 2026;23:63–75. https://doi.org/10.7150/ijms.123494.

[51] Damgaard F, Jespersen MG, Møller JK, Coia JE, Dessau RB, Sydenham TV, et al. Distinct prophage infections in colorectal cancer-associated Bacteroides fragilis. Commun Med 2026 2026. https://doi.org/10.1038/s43856-026-01403-1.

[52] De Almeida CV, Lulli M, Di Pilato V, Schiavone N, Russo E, Nannini G, et al. Differential responses of colorectal cancer cell lines to enterococcus faecalis' strains isolated from healthy donors and colorectal cancer patients. J Clin Med 2019;8. https://doi.org/10.3390/jcm8030388. 2019;8.

[53] Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology 2010;138:2073–2087.e3. https://doi.org/10.1053/j.gastro.2009.12.064.
[54] Lin A, Zhang J, Luo P. Crosstalk between the MSI status and tumor microenvironment in colorectal cancer. Front Immunol 2020;11. https://doi.org/10.3389/fimmu.2020.02039.
[55] Zheng K, Wan H, Zhang J, Shan G, Chai N, Li D, et al. A novel NGS-based microsatellite instability (MSI) status classifier with 9 loci for colorectal cancer patients. J Transl Med 2020;18. https://doi.org/10.1186/s12967-020-02373-1.
[56] Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite instability: a review of what the oncologist should know. Cancer Cell Int 2020;20. https://doi.org/10.1186/s12935-019-1091-8.
[57] Byrd DA, Fan W, Greathouse KL, Wu MC, Xie H, Wang X. The intratumor microbiome is associated with microsatellite instability. J Natl Cancer Inst: J Natl Cancer Inst 2023;115:989–93. https://doi.org/10.1093/jnci/djad083.
[58] Li L, Chandra V, McAllister F. Tumor-resident microbes: the new kids on the microenvironment block. Trends Cancer 2024;10:347–55. https://doi.org/10.1016/j.trecan.2023.12.002.
[59] Ng C, Li H, Wu WKK, Wong SH, Yu J. Genomics and metagenomics of colorectal cancer. J Gastrointest Oncol 2019;10:1164–70. https://doi.org/10.21037/jgo.2019.06.04.
[60] Yao Q, Tang M, Zeng L, Chu Z, Sheng H, Zhang Y, et al. Potential of fecal microbiota for detection and postoperative surveillance of colorectal cancer. BMC Microbiol 2021;21. https://doi.org/10.1186/s12866-021-02182-6.
[61] Löwenmark T, Löfgren-Burström A, Zingmark C, Ljuslinder I, Dahlberg M, Edin S, et al. Tumour colonisation of Parvimonas micra is associated with decreased survival in colorectal cancer patients. Cancers (Basel) 2022;14:5937. https://doi.org/10.3390/cancers14235937.
[62] Kerdreux M, Edin S, Löwenmark T, Bronnec V, Löfgren-Burström A, Zingmark C, et al. *Porphyromonas gingivalis* in colorectal cancer and its association to patient prognosis. J Cancer 2023;14:1479–85. https://doi.org/10.7150/jca.83395.
[63] Zhou Y, Luo G-H. *Porphyromonas gingivalis*and digestive system cancers. World J Clin Cases 2019;7:819–29. https://doi.org/10.12998/wjcc.v7.i7.819.
[64] Mukherjee A, Lordan C, Ross RP, Cotter PD. Gut microbes from the phylogenetically diverse genus*Eubacterium*and their various contributions to gut health. Gut Microbes 2020;12:1802866. https://doi.org/10.1080/19490976.2020.1802866.
[65] Sulaiman JE, Thompson J, Cheung PLK, Qian Y, Mill J, James I, et al. Phocaeicola vulgatus shapes the long-term growth dynamics and evolutionary adaptations of Clostridioides difficile. Cell Host Microbe 2025;33:42–58.e10. https://doi.org/10.1016/j.chom.2024.12.001.
[66] Sun Y, Huang Y, Li R, Zhang J, Fan X, Su X. Benchmarking and optimizing microbiome-based bioinformatics workflow for non-invasive detection of intestinal tumors. Microbiome Res Rep 2025;4(43). https://doi.org/10.20517/mrr.2025.75. 2025;4:N/A-N/A.
[67] Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol 2021;22(1):93. https://doi.org/10.1186/s13059-021-02306-1. 2021;22.
[68] Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep 2020;10(1):6026. https://doi.org/10.1038/s41598-020-63159-5. 2020;10.
[69] Moreno-Indias I, Zomer AL, Gómez-Cabrero D, Claesson MJ. Editorial: microbiome and machine learning. Front Microbiol 2022;13. https://doi.org/10.3389/fmicb.2022.964921.
[70] Papoutsoglou G, Tarazona S, Lopes MB, Klammsteiner T, Ibrahimi E, Eckenberger J, et al. Machine learning approaches in microbiome research: challenges and best practices. Front Microbiol 2023;14. https://doi.org/10.3389/fmicb.2023.1261889.

Abbreviations

CRC: Colorectal cancer
ML: Machine learning
RF: Random Forest
XGB: Extreme gradient boosting
PCA: Principal component analysis
PCoA: Principal Coordinates Analysis
MSI: Microsatellite instable (instability)
MSS: Microsatellite stable
TL: Tumour location(s)
HC: Healthy condition(s)
AUROC: Area under receiver operating characteristic curve
TP: True positives
TN: True negatives
FP: False positives
FN: False negatives
MaA: Macro-average values