

AI-DRIVEN MICROBIOME-BASED BIOMARKERS FOR EARLY DIAGNOSIS OF CANCER AND AUTOIMMUNE DISEASES: CURRENT ADVANCES AND FUTURE PERSPECTIVES

Dr. Károly Szili^{*1}, Anza Ahmad², Dr. Viktor Gulyás-Oldal³, Bahaaeldin Anwer⁴

^{*1,3}Department of Obstetrics and Gynecology, Faculty of Health and Sport Sciences, Széchenyi István University in Győr, Hungary.

²Lecturer, Department of Pharmacy Practice, Akhter Saeed College of Pharmaceutical Sciences.

⁴Department of Microbiology and Immunology, Faculty of Pharmacy, Al-Azhar University, Assiut branch.

Corresponding Author: *

Dr. Károly Szili

DOI: <https://doi.org/10.5281/zenodo.21389095>

Received	Accepted	Published
01 January 2026	17 March 2026	31 March 2026

ABSTRACT

The human microbiome encompassing the gut, oral, and mucosal microbial communities has emerged as a rich, minimally invasive source of biomarkers for diseases that remain difficult to diagnose at an early, treatable stage, including gastrointestinal and non-gastrointestinal cancers and a broad spectrum of autoimmune conditions. Advances in high-throughput sequencing (16S rRNA amplicon and shotgun metagenomics) have generated large, high-dimensional, compositional datasets that exceed the analytic capacity of conventional statistics. Artificial intelligence (AI) and machine learning (ML) spanning random forests, gradient boosting, support vector machines, deep neural networks, graph neural networks, and, most recently, transformer-based genomic foundation models have become central to extracting reproducible diagnostic signals from this complexity. This review synthesizes recent (2021–2026) evidence on AI-driven microbiome biomarker discovery across colorectal, gastric, pancreatic, hepatocellular, oral/head-and-neck, and breast cancers, and across autoimmune diseases including systemic lupus erythematosus, rheumatoid and juvenile idiopathic arthritis, inflammatory bowel disease, multiple sclerosis, and myasthenia gravis. We summarize commonly used algorithms and representative diagnostic performance, describe the growing role of explainable AI and multi-omics integration in improving biological interpretability, and critically appraise persistent barriers to clinical translation including batch effects, cohort heterogeneity, low-biomass contamination, data leakage, limited external validation, and regulatory uncertainty. We conclude by outlining future directions, including federated learning, foundation models for metagenomics, and prospective multicentre validation, that are needed to move AI-microbiome diagnostics from research cohorts into routine clinical screening.

Keywords: Artificial Intelligence, Machine Learning, Gut Microbiome, Biomarkers, Cancer Diagnosis, Autoimmune Disease, Metagenomics, Explainable AI, Precision Medicine

1. INTRODUCTION

Early diagnosis remains the single most important determinant of outcome in both oncology and autoimmune disease management, yet the biomarkers currently used in routine clinical practice serum tumour markers, autoantibody

panels, endoscopy, and histopathology are frequently insensitive, invasive, or detect disease only after substantial tissue damage has occurred (Das et al., 2023). Over the past decade, the human microbiome has been established as an active participant in carcinogenesis and immune

dysregulation rather than a passive bystander, through mechanisms including chronic mucosal inflammation, genotoxin and carcinogenic metabolite production, barrier disruption, molecular mimicry, and modulation of systemic immune tone (Li et al., 2025).

Because microbial community composition can be sampled non-invasively (stool, saliva, tongue coating) and reflects both local and systemic pathophysiology, it has attracted intense interest as a diagnostic substrate (Radu et al., 2025). However, microbiome data are compositional, sparse, high-dimensional, and confounded by diet, geography, medication use, and sequencing platform properties that render conventional univariate statistics poorly suited to biomarker discovery (Radu et al., 2025). Artificial intelligence and machine learning offer a natural analytic complement, capable of modelling non-linear, high-order interactions among thousands of microbial features simultaneously while performing embedded feature selection to yield parsimonious, clinically actionable biomarker panels (DiMucci, 2020).

This review synthesizes advances published predominantly between 2021 and 2026 on AI-driven microbiome biomarkers for cancer and autoimmune disease diagnosis. We describe the analytic toolkit now in routine use, summarize disease-specific evidence, present two comparative tables and an integrated workflow figure, and critically examine the methodological and translational barriers that must be overcome before these tools can be deployed at the point of care.

2. The Human Microbiome in Cancer and Autoimmune Pathogenesis

Dysbiosis—an imbalance in the diversity, composition, or function of resident microbial communities—has been implicated across the cancer-autoimmunity spectrum (Lazar et al., 2018). In colorectal carcinogenesis, expansion of pro-inflammatory and genotoxic taxa such as *Fusobacterium nucleatum* and *Parvimonas micra* promotes chronic inflammation, disrupts the mucosal barrier, and generates carcinogenic metabolites that accelerate the adenoma-to-carcinoma sequence (Singh et al., 2025). In the

stomach, *Helicobacter pylori* infection reshapes the broader gastric and intestinal microbiota in ways that extend beyond its direct genotoxic effects, influencing both gastric and colorectal cancer risk through immune-microbiota crosstalk (Nobels et al., 2025).

In autoimmune disease, gut dysbiosis is thought to contribute to loss of immune tolerance through mechanisms such as Th17 pathway activation, molecular mimicry between microbial and self-antigens, and altered short-chain fatty acid production (Sadeghpour Heravi, 2024). Specific taxa—for example *Subdoligranulum* in rheumatoid arthritis, *Eggerthella* in multiple autoimmune conditions, and expanded *Veillonella* in autoimmune hepatitis—have been repeatedly implicated across independent cohorts, suggesting shared, cross-disease microbial signatures that AI-based multi-class classifiers are now beginning to exploit (An et al., 2025).

3. AI and Machine Learning Approaches for Microbiome Data Analysis

The analytic pipeline underlying most AI-driven microbiome biomarker studies proceeds from biospecimen collection and sequencing, through preprocessing and feature selection, to model training, interpretation, and ideally external validation (Figure 1). Classical supervised ML methods, particularly random forest and gradient-boosted trees (XGBoost, LightGBM), remain the most widely used owing to their robustness to compositional, high-dimensional data and their capacity to rank feature importance. Regularized linear models (LASSO, elastic net) are frequently used for sparse biomarker panel selection, especially in smaller cohorts such as juvenile idiopathic arthritis or systemic lupus erythematosus studies (Cavalca et al., 2025).

Deep learning approaches—convolutional neural networks operating on restructured two-dimensional microbiome representations, graph neural networks that explicitly model microbial co-occurrence networks, and autoencoder-based dimensionality reduction—have improved predictive performance in larger, pooled multi-cohort datasets, though at the cost of interpretability and increased data requirements (Li et al., 2026). Most recently, transformer-based

genomic foundation models pretrained on large sequence corpora have been proposed to generate transferable metagenome representations that generalize across heterogeneous cohorts and sequencing platforms, addressing a long-standing weakness of earlier ML pipelines: poor cross-cohort generalizability (Khan et al., 2026). A parallel and increasingly essential development is explainable AI (XAI). Methods such as SHapley

Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) allow researchers to trace a classifier's output back to specific contributing taxa or functional pathways, improving both biological plausibility and clinician trust – a prerequisite for regulatory approval and clinical adoption of any microbiome-based diagnostic (Das and Rad, 2020; Ortigossa et al., 2024).

AI-Driven Microbiome Biomarker Discovery Workflow

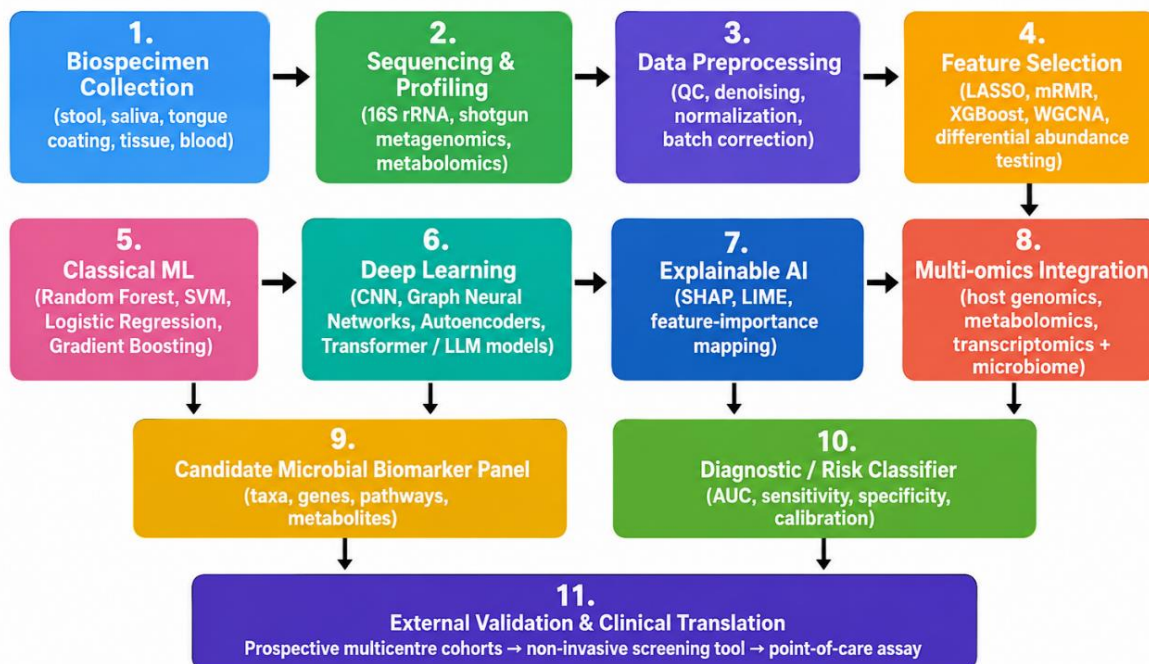


Figure 1. Generalized AI-driven microbiome biomarker discovery workflow, from biospecimen collection through sequencing, preprocessing, feature selection, classical ML and deep learning classification, explainable AI, multi-omics integration, and external clinical validation.

Table 1. Overview of AI/ML algorithm classes commonly applied to microbiome-based biomarker discovery.

Algorithm / Class	Type	Typical Application	Strengths	Limitations
Random Forest (RF)	Ensemble Machine Learning	Taxonomic classification of cancer vs. healthy stool/saliva microbiomes; feature-importance ranking of	Robust to overfitting; handles high-dimensional compositional data; provides built-in feature importance	Limited ability to capture complex microbe-microbe interactions; performance may decline with very small cohorts

			microbial biomarkers		
Gradient Boosting (XGBoost/LightGBM)	Ensemble Machine Learning		Multi-class classification of autoimmune diseases; microbial biomarker panel construction	High predictive accuracy; efficient with sparse datasets; supports regularized feature selection	Can overfit without careful hyperparameter tuning; less interpretable than individual decision trees
Support Vector Machine (SVM)	Kernel-based Machine Learning		Binary classification of inflammatory bowel disease (IBD) and colorectal cancer (CRC) using 16S rRNA abundance profiles	Effective for high-dimensional, low-sample-size datasets; strong generalization	Performance depends on kernel selection; limited scalability to very large cohorts
LASSO / Elastic-Net Regression	Regularized Linear Model		Sparse biomarker selection from thousands of microbial features (e.g., JIA and SLE studies)	Produces compact and interpretable biomarker panels; reduces multicollinearity	Assumes approximately linear relationships; may exclude correlated but biologically relevant taxa
Convolutional Neural Network (CNN)	Deep Learning		Disease prediction using image-like or structured microbiome representations	Learns complex spatial and hierarchical microbial patterns; high predictive performance with large datasets	Requires extensive training data; limited model interpretability
Graph Neural Network (GNN)	Deep Learning		Modeling microbial co-occurrence networks and host-microbe interaction networks	Naturally captures ecological and relational structures within microbiomes	Computationally demanding; still an emerging approach in microbiome research
Autoencoder Embedding	Deep Learning		Dimensionality reduction and denoising of sparse metagenomic datasets	Learns informative latent representations without manual feature engineering	Learned latent features are often difficult to interpret biologically
Transformer Genomic Models	Deep Learning (LLM-based)		Metagenomic sequence representation	Leverages pretrained sequence	High computational requirements;

		and cross-cohort disease prediction	knowledge; performs well across heterogeneous datasets	limited clinical validation to date
Explainable AI (SHAP, LIME)	Post-hoc Interpretability	Identifying influential taxa and pathways contributing to AI predictions	Enhances model transparency, clinical trust, and biological interpretability	Explanations may vary across models and do not replace experimental validation

4. AI-Driven Microbiome Biomarkers in Cancer

4.1 Colorectal and Gastric Cancer

Colorectal cancer (CRC) remains the most extensively studied cancer in the AI-microbiome literature (Ortigossa et al., 2024). Meta-analyses pooling metagenomic data from multiple international cohorts have used interpretable ML applied to functional gene-family profiles, rather than taxonomy alone, to improve prediction of the adenoma-to-carcinoma transition and reveal candidate molecular mechanisms underlying microbiome-driven tumourigenesis (Peterson et al., 2023). Explainable AI frameworks applied to fecal microbiome data have further refined biomarker identification for CRC by ranking the contribution of individual taxa to model output. In gastric cancer, machine learning applied to tongue-coating microbiota and tongue-image data has shown promise as a low-cost, non-invasive prospective screening approach validated in multicentre cohorts, while AI-mapped gut-oral microbiota interactions are being used to model gastric cancer risk more broadly (Canakcimaksutoglu et al., 2026; Wang et al., 2026).

4.2 Pancreatic and Hepatocellular Cancer

Pancreatic cancer, typically diagnosed at an advanced, unresectable stage, has benefited from fecal and oral/salivary microbiome-based ML classifiers. A faecal microbiota signature identified through ensemble ML demonstrated high specificity for pancreatic ductal adenocarcinoma across multiple cohorts, and combined salivary microbiota circulating miRNA panels have further improved non-invasive detection (Zhao and Liu, 2020). For hepatocellular carcinoma (HCC), a

distinct gut microbiota signature has been shown to precede clinical diagnosis in longitudinal case-control data, raising the possibility of pre-symptomatic microbiome-based risk stratification in cirrhotic and high-risk populations (Haque, 2025).

4.3 Oral, Head-and-Neck, and Other Cancers

Oral and salivary microbiome profiling, analysed with supervised ML classifiers, has yielded diagnostic models for head and neck squamous cell carcinoma with reported discriminative accuracy (AUC) in the range of 0.78–0.89 (Wang et al., 2025). Tongue-coating microbiota-based ML classifiers have also been applied across a panel of digestive system tumours (HCC, pancreatic ductal adenocarcinoma, cholangiocarcinoma, and CRC), distinguishing tumour patients from healthy controls using microbial diversity and evenness metrics (Ma et al., 2025). AI-based analysis of breast cancer, while historically dominated by imaging and pathology biomarkers, is increasingly incorporating tissue and fecal microbial signatures as complementary features within multimodal risk models (Chen et al., 2019).

5. AI-Driven Microbiome Biomarkers in Autoimmune Diseases

5.1 Systemic Lupus Erythematosus and Rheumatoid/Juvenile Idiopathic Arthritis

Machine learning applied to fecal microbiota in female patients with systemic lupus erythematosus (SLE) has identified candidate genus-level biomarkers using feature-selection-guided classifiers, reflecting the disproportionate burden of SLE in women and the clinical need for earlier, non-invasive diagnostic tools (An et al., 2025). In

rheumatoid arthritis (RA), integrative analyses combining gut microbiome metagenomics with host gene-expression data (via weighted correlation network analysis and ML) have identified shared RA-inflammatory bowel disease molecular hubs, while a dedicated RA gut microbial biobank has characterized core arthritogenic species associated with host inflammatory indices (Guan et al., 2023). In juvenile idiopathic arthritis, LASSO-guided feature selection applied to fecal metagenomic data has produced a validated auxiliary ML diagnostic tool (Tu et al., 2024).

5.2 Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is the autoimmune/inflammatory condition with the most mature AI-microbiome diagnostic literature. Large benchmarking efforts spanning multiple North American cohorts (thousands of samples) have systematically compared data-processing pipelines and ML algorithms (random forest, logistic regression, neural networks) for IBD classification, while other studies have applied multiple feature-selection algorithms (mRMR, XGBoost, information gain) to generate sparse, interpretable IBD biomarker panels and patient subgroup clustering (Faisal et al., 2026; Tu et al., 2024). Recently reported gut-derived biomarker models have achieved area-under-the-curve values as high as 0.97–0.999 for distinguishing IBD from

other gastrointestinal and rheumatologic conditions in non-IBD comparator cohorts, although questions remain regarding the influence of prior antibiotic and immunosuppressive treatment on these signatures (Niu et al., 2026).

5.3 Multiple Sclerosis, Myasthenia Gravis, and Multi-Autoimmune Classification

Beyond gastrointestinal autoimmunity, supervised ML applied to gut microbiota amplicon sequence variant (ASV) data has demonstrated feasibility for diagnostic screening in myasthenia gravis, a neuromuscular autoimmune disorder previously unexplored from a microbiome perspective (Faisal et al., 2026). A notable methodological advance is the development of multi-class XGBoost classifiers trained on aggregated public cohorts spanning eight autoimmune diseases (rheumatoid arthritis, ankylosing spondylitis, Crohn's disease, ulcerative colitis, celiac disease, multiple sclerosis, myasthenia gravis, and SLE), which identified shared microbial genera including *Eggerthella*, *Subdoligranulum*, *Ruminococcus*, *Parabacteroides*, and *Shigella* with cross-disease relevance, pointing toward a common dysbiotic 'autoimmune signature' that AI models can help resolve into disease-specific components. (Lazar et al., 2018; Li et al., 2025)

Table 2. Representative AI-driven microbiome biomarker studies across cancer and autoimmune diseases (2021–2026).

Disease	Specimen	Cohort Size	AI/ML Method	Key Microbial Biomarkers	Reported Performance
Colorectal cancer	Fecal (shotgun metagenomics)	~1,042 samples; 7 cohorts (meta-analysis)	Interpretable ML on functional (gene-family) profiles	<i>Fusobacterium nucleatum</i> , <i>Parvimonas micra</i> , functional pathway shifts	Improved precision over taxonomy-only models (Casimiro-Soriguer et al., 2022)
Gastric cancer	Tongue coating image + microbiota	Prospective multicentre cohort	ML tongue-image classifier + microbiome profiling	Altered tongue-coating bacterial diversity and composition	High discriminative accuracy in prospective validation

						(Yuan et al., 2023)
Pancreatic cancer	Fecal		Multi-cohort case-control	Random Forest / Gradient Boosting	Fecal microbial signature with high specificity for pancreatic ductal adenocarcinoma	High specificity for early detection (Kartal et al., 2022)
Hepatocellular carcinoma	Fecal (pre-diagnostic)		Longitudinal case-control	ML classifier on pre-diagnostic microbiota profiles	Distinct microbiota signature preceding clinical diagnosis	Signature detectable before clinical HCC diagnosis (Yang et al., 2023)
Head and neck squamous carcinoma	Oral saliva (16S rRNA + metagenomics)		Case-control; single centre	Supervised ML diagnostic classifier	Microbial-metabolic dysbiosis signatures	AUC = 0.78–0.89 (2025 study)
Digestive system tumours (HCC, PDAC, cholangiocarcinoma, CRC)	Tongue coating		710 patients vs. 489 controls	ML classifier on tongue-coating microbiota	Increased microbial evenness (Pielou index)	Distinguished tumour vs. healthy individuals with clinically useful accuracy (2025)
Systemic lupus erythematosus (SLE)	Fecal (female cohort)		Case-control	Feature selection + Supervised ML	Altered genus-level gut taxa associated with disease activity	Candidate diagnostic biomarker panel identified (Song et al., 2023)
Myasthenia gravis	Fecal		Case-control	Supervised ML (ASV-based taxonomic classification)	Distinct gut microbiota structure compared with controls	Demonstrated feasibility for non-invasive diagnostic screening (Kang et al., 2023)
Inflammatory bowel disease (Crohn's disease/Ulcerative colitis)	Fecal (multi-cohort)		15 cohorts; 7,707 samples	Random Forest, Logistic Regression,	Reduced microbial diversity, dysbiotic taxa,	AUC up to 0.97–0.999 using optimized pipelines

				Neural Networks	batch-corrected features	(Kubinski et al., 2022; 2025)
Juvenile idiopathic arthritis	Fecal metagenomics	Case-control	LASSO feature selection + ML classifier	Selected microbial taxa	Validated as an auxiliary diagnostic tool (2024)	
Multi-autoimmune disease panel (RA, SpA, CD, UC, CeD, MS, MG, SLE)	Fecal (aggregated public cohorts)	Multi-cohort aggregation	XGBoost multi-class classifier	<i>Eggerthella</i> , <i>Subdoligranulum</i> , <i>Ruminococcus</i> , <i>Parabacteroides</i> , <i>Shigella</i> , and related taxa	Demonstrate feasibility of multi-disease classification (2025)	

6. Multi-Omics Integration and Explainable AI

A clear trajectory in the recent literature is the shift from single-omics (taxonomic) microbiome classifiers toward integrated multi-omics frameworks that combine metagenomics with metatranscriptomics, metabolomics, host genomics, and clinical metadata. Such integration substantially improves mechanistic interpretability for example, linking a taxonomic shift to a specific metabolite (short-chain fatty acids, secondary bile acids, or genotoxins) and, in turn, to a host immune or epithelial response. AI frameworks incorporating graph neural networks and reinforcement learning have been proposed to model these multi-layered host-microbe interactions and to support causal, rather than purely correlative, biomarker discovery (Faisal et al., 2026; Guan et al., 2023).

Explainable AI is increasingly positioned not as an optional add-on but as a translational requirement: regulatory bodies and clinicians are unlikely to accept 'black-box' microbiome classifiers without transparent, taxon-level or pathway-level justification for a diagnostic call. Studies applying SHAP-based feature attribution to colorectal cancer classifiers, for instance, have demonstrated that model-derived biomarker rankings can recapitulate independently validated microbial risk factors, strengthening confidence in AI-nominated candidates prior to wet-lab validation (DiMucci, 2020; Nobels et al., 2025).

7. Challenges and Limitations

Despite rapid progress, several methodological barriers constrain clinical translation. First, batch effects arising from differences in DNA extraction, sequencing platform, and bioinformatic pipeline can introduce spurious signals that mimic disease association; systematic reanalyses have shown that some published cancer-microbiome findings are vulnerable to such technical confounding, and that low-biomass samples are particularly susceptible to contamination artefacts that can be misread as biological signal, as illustrated by controversies surrounding claims of an intratumoural or fetal microbiome.

Second, most existing models are trained and validated within single cohorts or single countries, limiting generalizability; benchmarking studies applying identical pipelines across 15–83 independent cohorts have shown that model performance can degrade substantially upon external validation, underscoring the need for standardized preprocessing and reporting practices. Third, privacy and re-identification risks are non-trivial human genomic reads embedded within microbiome shotgun sequencing data can, in principle, be used to reconstruct identifiable host genetic information, raising data-governance concerns for large, pooled microbiome datasets. Fourth, deep learning and foundation models, while powerful, require large, well-annotated training sets that remain scarce for many rarer autoimmune conditions, and their interpretability, even with post-hoc XAI methods, remains imperfect. Finally, regulatory pathways for

microbiome-based in vitro diagnostics are still immature relative to those for genomic or proteomic biomarkers, and prospective, pre-registered clinical validation studies remain the exception rather than the rule.

8. Future Perspectives

Several developments are likely to shape the next phase of AI-microbiome diagnostics. Federated learning architectures, which allow model training across multiple institutions without centralizing raw sequencing data, offer a route to larger, more diverse training sets while addressing privacy and data-governance concerns. Transformer-based genomic foundation models, pretrained on large, heterogeneous metagenomic corpora, may substantially improve cross-cohort and cross-population generalizability compared with cohort-specific classical ML models. Greater emphasis on multi-omics and multimodal integration combining microbiome, metabolome, host transcriptome, imaging, and electronic health record data is expected to yield biomarker panels with improved mechanistic grounding and clinical robustness.

Prospective, multicentre validation studies, ideally pre-registered and adhering to standardized reporting guidelines, are needed to move candidate biomarkers from retrospective case-control discovery toward regulatory-grade diagnostic tools. Point-of-care and home-collection sampling technologies (stool, saliva, tongue coating), paired with rapid, low-cost sequencing or targeted qPCR panels derived from AI-nominated biomarkers, could enable population-scale, non-invasive screening programmes for both cancer and autoimmune disease, analogous to existing fecal immunochemical testing for colorectal cancer. Finally, continued investment in explainable and causally-informed AI will be essential to ensure that microbiome-derived diagnostics are not only accurate but also biologically interpretable, clinically trustworthy, and equitable across diverse populations.

9. Conclusion

AI and machine learning have transformed the human microbiome from a scientifically intriguing but analytically intractable dataset into an

increasingly actionable source of non-invasive diagnostic biomarkers for cancer and autoimmune disease. Across colorectal, gastric, pancreatic, hepatocellular, and head-and-neck cancers, and across systemic lupus erythematosus, rheumatoid and juvenile idiopathic arthritis, inflammatory bowel disease, multiple sclerosis, and myasthenia gravis, AI-driven classifiers have consistently identified reproducible microbial signatures with clinically meaningful discriminative performance. Realizing the full clinical potential of these advances will require sustained attention to methodological rigor, cross-cohort validation, explainability, and data governance but the trajectory of the past five years suggests that AI-microbiome diagnostics are approaching, rather than merely aspiring toward, clinical relevance.

REFERENCES

- Kang Y, Li L, Kang X, Zhao Y, Cai Y. Machine learning strategy for identifying altered gut microbiomes for diagnostic screening in myasthenia gravis. *Front Cell Infect Microbiol.* 2023;13:1249381.
- Song X, et al. Gut microbiota landscape and potential biomarker identification in female patients with systemic lupus erythematosus using machine learning. *Front Cell Infect Microbiol.* 2023;13:1289124.
- An, T., S. Zhang, J. Li, H. Wang, L. Chen, Y. Shi, J. Wang, S. Han, R. Wang, and L. Wang. 2025. Gut microbiota analysis reveals microbial signature for multi-autoimmune diseases based on machine learning model. *Frontiers in Microbiology.* 16:1660775.
- Canakcimaksutoglu, B., N.S. Ersoz, B. Bakir-Gungor, and M. Yousef. 2026. Discovering Potential Taxonomic Biomarkers of Gastrointestinal Cancers from Various Human Microbiota via GSM Machine Learning Approach. *Applied Sciences.* 16:6879.

- Cavalca, G., M. Vergani, D. Cangelosi, A. Consolaro, M. Gattorno, A. Ravelli, J. Munro, B. Novakovic, A. Duncan, and P. Uva. 2025. Stochastic epigenetic mutation profiles as biomarkers of clinical activity in juvenile idiopathic arthritis: a multi-omic machine learning approach for gene prioritization. *Molecular Medicine*. 31:289.
- Chen, Y., X. Chen, H. Yu, H. Zhou, and S. Xu. 2019. Oral microbiota as promising diagnostic biomarkers for gastrointestinal cancer: a systematic review. *OncoTargets and therapy*:11131-11144.
- Das, A., and P. Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.
- Das, S., M.K. Dey, R. Devireddy, and M.R. Gartia. 2023. Biomarkers in cancer detection, diagnosis, and prognosis. *Sensors*. 24:37.
- DiMucci, D.M. 2020. Machine learning for microbial ecology: predicting interactions and identifying their putative mechanisms. Boston University.
- Faisal, S., I. Ullah, P.A. Kambey, A. Malik, M.A. Ejaz, S.A. Shah, and Y.-X. Li. 2026. Nurrl deficiency orchestrates a coupled liver-gut pathological axis revealed by multi-omics and deep-learning histopathology. *Frontiers in Immunology*. 17:1861669.
- Guan, Y., Y. Zhang, Y. Zhu, and Y. Wang. 2023. CXCL10 as a shared specific marker in rheumatoid arthritis and inflammatory bowel disease and a clue involved in the mechanism of intestinal flora in rheumatoid arthritis. *Scientific Reports*. 13:9754.
- Haque, M. 2025. PNPLA3I148M-associated MASLD susceptibility: the pivotal role of microbiota-driven metabolic dysregulation. Dissertation, RWTH Aachen University, 2025.
- Khan, S.A., X. Martinez-deMorentin, A.R. Alsabbagh, A. Maillo, V. Lagani, D. Gomez-Cabrero, R. Lehmann, and J. Tegner. 2026. Multimodal foundation transformer models for multiscale genomics. *Nature Methods*. 23:299-311.
- Lazar, V., L.-M. Ditu, G.G. Pircalabioru, I. Gheorghe, C. Curutiu, A.M. Holban, A. Picu, L. Petcu, and M.C. Chifiriuc. 2018. Aspects of gut microbiota and immune system interactions in infectious diseases, immunopathology, and cancer. *Frontiers in immunology*. 9:1830.
- Li, H., Y. Bai, Y. Lv, S. Fang, R. Zhao, B. Xu, J. Fan, and Z. Tang. 2026. DVMMHGNN: A Dual View Multi-Modal Heterogeneous Graph Neural Network with Contrastive Learning for Microbe-Informed Drug Repurposing. *Journal of Chemical Information and Modeling*. 66:2663-2682.
- Li, J., A. Wang, H. Guo, W. Zheng, R. Chen, C. Miao, D. Zheng, J. Peng, J. Wang, and Z. Chen. 2025. Exosomes: innovative biomarkers leading the charge in non-invasive cancer diagnostics. *Theranostics*. 15:5277.
- Ma, Y., Z. Jiang, Y. Wang, L. Pan, K. Liu, R. Xia, L. Yuan, and X. Cheng. 2025. Tongue coating microbiota-based machine learning for diagnosing digestive system tumours. *Journal of oral microbiology*. 17:2487645.
- Niu, M., Y. Luo, C. Yangyang, A.F. Almulla, A.F. Carvalho, J. Li, Y. Zhang, and M. Maes. 2026. A Shift Toward Proteolytic Gut Fermentation Links Systemic Inflammation to Clinical Phenotypes in Major Depressive Disorder. *medRxiv:2026.2003.2022.26348995*.
- Nobels, A., C. van Marcke, B.F. Jordan, M. Van Hul, and P.D. Cani. 2025. The gut microbiome and cancer: from tumorigenesis to therapy. *Nature metabolism*. 7:895-917.
- Ortigossa, E.S., T. Gonçalves, and L.G. Nonato. 2024. Explainable artificial intelligence (xai)—from theory to methods and applications. *IEEE Access*. 12:80799-80846.
- Peterson, S.N., A.M. Eroshkin, P.Z. Kozbial, E. Florio, F. Fouladi, N. Strom, Y. Valdes, G. Kuehn, G. Casaburi, and T. Kuehn. 2023. Meta-analysis of Multi-functional Biomarkers for Discovery and Predictive Modeling of Colorectal Adenoma and Carcinoma.

- Radu, C.-M., C.C. Radu, and D.C. Zaha. 2025. Salivary and Microbiome Biomarkers in Periodontitis: Advances in Diagnosis and Therapy—A Narrative Review. *Medicina*. 61:1818.
- Sadeghpour Heravi, F. 2024. Gut microbiota and autoimmune diseases: mechanisms, treatment, challenges, and future recommendations. *Current Clinical Microbiology Reports*. 11:18-33.
- Singh, G., Z. Chaudhry, A. Boyadzhyan, K. Sasaninia, and V. Rai. 2025. Dysbiosis and colorectal cancer: Conducive factors, biological and molecular role, and therapeutic prospectives. *Exploration of Targeted Anti-tumor Therapy*. 6:1002329.
- Tu, J.-B., W.-J. Liao, S.-P. Long, M.-P. Li, and X.-H. Gao. 2024. Construction and validation of a machine learning model for the diagnosis of juvenile idiopathic arthritis based on fecal microbiota. *Frontiers in Cellular and Infection Microbiology*. 14:1371371.
- Wang, M., Y. Sun, and W. Gong. 2025. Machine learning models diagnose oral squamous cell carcinoma based on cross-cohort oral microbial signatures. *Frontiers in Microbiology*. 16:1692186.
- Wang, X., D. Xiong, S. Cui, B. Duan, G. Ding, Y. Huang, and Q. Wang. 2026. Artificial intelligence-enabled multi-omics biomarkers for immune checkpoint blockade: mechanisms, predictive modeling, and clinical translation. *Frontiers in Immunology*. 17:1732079.
- Zhao, Z., and W. Liu. 2020. Pancreatic cancer: a review of risk factors, diagnosis, and treatment. *Technology in cancer research & treatment*. 19:1533033820962117.