



Explainable Serum Microbiome Signatures for Fatty Liver Screening and Early Risk Stratification

Julien Bernard, Hazel Fleming, Francesca Greco

Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract

Nonalcoholic fatty liver disease represents one of the most prevalent chronic liver conditions worldwide, characterized by the progressive accumulation of hepatic steatosis without significant alcohol consumption. Despite its growing epidemiological burden, early and non-invasive screening remains a significant clinical challenge. Traditional diagnostic modalities, such as liver biopsy, are inherently invasive, whereas conventional imaging and standard serological markers frequently lack the requisite sensitivity to detect early-stage metabolic alterations. Recent advancements in multi-omics technologies have illuminated the complex bidirectional communication within the gut-liver axis, suggesting that host systemic metabolism and gut microbiota composition are inextricably linked during the pathogenesis of hepatic steatosis. This paper presents a comprehensive framework for identifying explainable serum-microbiome signatures to facilitate the early screening of nonalcoholic fatty liver disease. By integrating high-throughput microbiome sequencing data with serum metabolomic profiles, we aim to capture the multidimensional biological perturbations that precede structural hepatic changes. Furthermore, the application of explainable artificial intelligence techniques ensures that the predictive models transcend traditional black-box paradigms, providing clinicians with transparent, biologically interpretable feature importances. Through the systematic evaluation of interacting microbial taxa and circulating metabolites, this approach not only achieves superior predictive performance but also elucidates the underlying pathophysiological mechanisms, thereby paving the way for targeted therapeutic interventions and personalized clinical management strategies.

Keywords: *Nonalcoholic fatty liver disease, microbiome, metabolomics, explainable artificial intelligence, diagnostic signatures*

1. Introduction

1.1 The Global Burden of Hepatic Steatosis

Nonalcoholic fatty liver disease has rapidly emerged as the most common chronic liver disorder globally, functioning as the hepatic manifestation of the metabolic syndrome. The pathophysiological spectrum of this condition ranges from simple isolated steatosis to nonalcoholic steatohepatitis, which is further characterized by hepatocellular ballooning, lobular inflammation, and progressive fibrosis. Over the past two decades, the global prevalence has escalated in parallel with rising rates of obesity, type 2 diabetes mellitus, and sedentary lifestyle patterns, currently affecting approximately one-quarter of the global adult population [1]. This unprecedented epidemiological trajectory places an immense economic and infrastructural burden on healthcare systems worldwide. The insidious nature of the disease often means that patients remain asymptomatic during the early stages, delaying clinical presentation until irreversible architectural liver damage, such as advanced fibrosis or cirrhosis, has occurred. Furthermore, individuals with progressive stages of the disease face significantly elevated risks of developing hepatocellular carcinoma and severe cardiovascular complications, which

constitute the leading causes of mortality in this patient demographic [2]. Recognizing the critical window for intervention before the onset of irreversible fibrotic remodeling underscores an urgent clinical imperative to develop robust strategies for early detection. Timely identification enables the implementation of lifestyle modifications and emerging pharmacological interventions that can effectively halt or even reverse disease progression, thereby mitigating long-term morbidity and mortality associated with end-stage liver complications [3].

1.2 Limitations of Current Screening Modalities

The current clinical paradigm for diagnosing and staging nonalcoholic fatty liver disease is fraught with diagnostic compromises. Liver biopsy has historically been regarded as the gold standard for diagnosis; however, its widespread application as a screening tool is fundamentally precluded by its invasive nature, high associated costs, risk of severe complications such as hemorrhage, and inherent sampling variability [4]. Consequently, clinical practice heavily relies on non-invasive imaging techniques and conventional circulating biomarkers. Ultrasonography is the most frequently utilized first-line imaging modality due to its wide availability and lack of ionizing radiation. Nevertheless, standard grayscale ultrasound exhibits significant limitations, particularly a profound lack of sensitivity in detecting mild hepatic steatosis and a notable susceptibility to operator dependence [5]. Advanced imaging modalities, including transient elastography and magnetic resonance imaging-derived proton density fat fraction, offer enhanced accuracy for quantifying liver fat and stiffness but are restricted by substantial infrastructural requirements and prohibitive costs, limiting their deployment in resource-constrained primary care settings [6]. Concurrently, traditional serum biomarkers such as alanine aminotransferase and aspartate aminotransferase are routinely measured; however, extensive clinical evidence indicates that these aminotransferases can remain well within normal physiological ranges even in patients with histologically confirmed aggressive steatohepatitis [7]. This diagnostic void highlights the pressing necessity for novel, accessible, and highly sensitive screening biomarkers capable of accurately reflecting the earliest subclinical metabolic and cellular aberrations associated with hepatic lipid accumulation.

1.3 The Promise of Explainable Artificial Intelligence in Multi-Omics

In recent years, the exponential growth of high-throughput molecular technologies has facilitated the comprehensive profiling of biological systems across multiple omic layers. The integration of gut microbiome sequencing with host serum metabolomics offers a particularly promising avenue for biomarker discovery, given the profound interconnectedness of the gut-liver axis in metabolic regulation [8]. However, the translation of complex, high-dimensional multi-omic datasets into actionable clinical diagnostics necessitates advanced computational approaches. Machine learning algorithms, particularly deep learning and ensemble methods, have demonstrated exceptional predictive capabilities in deciphering non-linear patterns within massive biological datasets [9]. Despite these successes, the clinical adoption of such models is frequently hindered by their inherent black-box nature, which obscures the underlying decision-making processes from healthcare practitioners. Explainable artificial intelligence bridges this translational gap by providing methodological frameworks that render complex model predictions interpretable. Techniques that quantify individual feature contributions allow researchers and clinicians to identify precisely which microbial taxa and serum metabolites are driving a positive screening result. By delineating these biologically plausible signatures, explainable artificial intelligence not only bolsters clinical trust and facilitates regulatory compliance but also uncovers novel mechanistic insights into disease pathogenesis, thereby transforming predictive modeling from a purely statistical exercise into a vital engine for biomedical discovery.

2. Background and Literature Review

2.1 The Gut-Liver Axis in Metabolic Dysfunction

The anatomical and physiological connection between the gastrointestinal tract and the liver, commonly referred to as the gut-liver axis, is heavily implicated in the initiation and progression of metabolic liver diseases. The liver receives approximately seventy percent of its blood supply directly from the intestine via the portal vein, establishing it as the first line of defense and the primary metabolic

processor of gut-derived signals, dietary nutrients, and microbial metabolites [10]. Under homeostatic conditions, a robust intestinal epithelial barrier restricts the translocation of pathogenic microorganisms and their structural components into the portal circulation. However, dietary indiscretions, particularly diets high in refined carbohydrates and saturated fats, can induce significant alterations in the composition and function of the gut microbiota, a state known as dysbiosis. This dysbiotic state is frequently characterized by a reduction in microbial diversity and a shift towards an inflammatory microbial profile [11]. Dysbiosis subsequently compromises the integrity of the intestinal tight junctions, leading to increased intestinal permeability, colloquially termed leaky gut syndrome. This compromised barrier facilitates the unrestrained translocation of microbe-associated molecular patterns, most notably lipopolysaccharides derived from the cell walls of Gram-negative bacteria, into the portal and systemic circulation. Upon reaching the liver, these endotoxins bind to Toll-like receptors on the surface of hepatic Kupffer cells, triggering a robust inflammatory cascade characterized by the release of pro-inflammatory cytokines and reactive oxygen species [12]. This continuous low-grade inflammatory state exacerbates hepatic insulin resistance and promotes the dysregulation of lipid metabolism, thereby accelerating the transition from simple steatosis to overt steatohepatitis.

2.2 Microbially Derived Metabolites and Host Lipid Metabolism

Beyond the structural components of bacteria, the vast array of small molecules synthesized or modified by the gut microbiota exerts profound regulatory effects on host metabolic pathways. Short-chain fatty acids, primarily acetate, propionate, and butyrate, are generated through the microbial fermentation of indigestible dietary fibers in the colon. These volatile fatty acids serve not only as the primary energy source for colonocytes but also function as potent signaling molecules upon absorption into the bloodstream [13]. For instance, propionate is known to influence hepatic gluconeogenesis, while butyrate exerts significant anti-inflammatory effects and improves gut barrier function through the modulation of epigenetic mechanisms, such as histone deacetylase inhibition. Studies have consistently demonstrated that patients with metabolic liver disease often exhibit altered profiles of short-chain fatty acid-producing bacteria, resulting in a systemic deficiency of these protective metabolites [14]. Furthermore, the gut microbiota plays an indispensable role in the biotransformation of primary bile acids, which are synthesized from cholesterol in the liver and secreted into the intestine to facilitate dietary fat emulsification. Specific bacterial populations possess bile salt hydrolase activity, which deconjugates primary bile acids, followed by further structural modifications to generate secondary bile acids [15]. The ratio and specific profiles of primary to secondary bile acids critically regulate host lipid and glucose metabolism through their interaction with nuclear receptors, such as the farnesoid X receptor, and G-protein-coupled receptors. Dysregulation of this microbial bile acid metabolism is intricately linked to impaired hepatic lipid clearance, enhanced lipogenesis, and the perpetuation of hepatic steatosis.

2.3 Serum Metabolomics as a Circulatory Readout

Metabolomics, the comprehensive analysis of small-molecule intermediates and products of cellular metabolism within a biological system, provides a high-resolution snapshot of an individual's current physiological or pathological state. Serum, in particular, represents an ideal biofluid for non-invasive diagnostic screening because it acts as a central repository for metabolites originating from various tissues, including the liver, adipose tissue, skeletal muscle, and the gut microbiome [16]. Untargeted mass spectrometry and nuclear magnetic resonance spectroscopy are the primary analytical platforms utilized to capture these intricate metabolic signatures. In the context of hepatic steatosis, extensive metabolomic profiling has identified consistent perturbations in several key metabolic pathways. A hallmark signature involves the dysregulation of lipid metabolism, evidenced by elevated circulating levels of specific lipid classes, such as ceramides, diacylglycerols, and specific species of phosphatidylcholines, which reflect hepatic lipotoxicity and endoplasmic reticulum stress [17]. Additionally, abnormalities in amino acid metabolism are prominent; elevated serum concentrations of branched-chain amino acids, including leucine, isoleucine, and valine, as well as aromatic amino acids, are strongly correlated with insulin resistance and the severity of liver fat accumulation. The accumulation of branched-chain amino acids is thought to result from compromised hepatic catabolism

and altered microbial biosynthesis in the gut [18]. By comprehensively mapping these interconnected metabolomic alterations, researchers can construct a holistic view of the systemic metabolic dysfunction that characterizes the early stages of nonalcoholic fatty liver disease, long before macroscopic organ damage becomes clinically apparent.

2.4 The Imperative for Explainability in Healthcare Predictive Models

The integration of disparate omics layers, such as microbiome composition data and serum metabolomic profiles, results in highly complex, multidimensional datasets that present significant analytical challenges for classical statistical methodologies. Machine learning algorithms excel at navigating this high-dimensional space, identifying subtle, non-linear interactions between variables that collectively serve as robust predictive signatures [19]. However, the most accurate machine learning models, such as deep neural networks and complex tree-based ensembles, inherently operate as opaque systems where the mapping function between input features and diagnostic output is virtually impenetrable to human logic. In the highly regulated domain of clinical medicine, predictive accuracy alone is insufficient for the widespread adoption of a diagnostic tool; clinicians require transparency to ensure that the model's predictions align with established biological principles and do not merely exploit spurious correlations or structural biases within the training data [20]. Explainable artificial intelligence frameworks have been developed to address this critical need. Techniques such as Shapley Additive Explanations provide a mathematically rigorous approach derived from cooperative game theory to allocate the contribution of each individual feature to a specific prediction. By quantifying exactly how much a particular bacterial genus or a specific circulating metabolite influences the probability of a positive screening result, explainable artificial intelligence demystifies the diagnostic process [21]. This transparency is essential not only for establishing trust among healthcare providers and regulatory bodies but also for facilitating the targeted biological validation of the identified multi-omic signatures, ultimately driving the development of novel therapeutic targets.

3. Methodology and Analytical Framework

3.1 Study Cohort and Clinical Data Acquisition

The foundation of this diagnostic framework relies on a meticulously characterized clinical cohort designed to capture the earliest metabolic shifts associated with hepatic lipid accumulation. Participants were prospectively recruited from multiple primary care and gastroenterology clinics over a continuous three-year period. The inclusion criteria were strictly defined to enroll adult individuals between the ages of eighteen and sixty-five who were undergoing routine health evaluations. To ensure the specificity of the nonalcoholic etiology, extensive exclusion criteria were applied. Individuals with a history of significant alcohol consumption, defined as exceeding twenty grams per day for women and thirty grams per day for men, were rigorously excluded from the study [22]. Furthermore, participants presenting with concurrent liver pathologies, including chronic viral hepatitis, autoimmune hepatitis, hemochromatosis, and drug-induced liver injury, were removed from the potential cohort. Recent use of antibiotics, prebiotics, probiotics, or any medications known to significantly alter the gut microbiome or host lipid metabolism within the preceding three months also served as grounds for exclusion to minimize confounding variables [23]. Upon providing written informed consent, each participant underwent a comprehensive clinical assessment. Detailed demographic information, including age, biological sex, ethnic background, and detailed anthropometric measurements such as body mass index and waist circumference, was systematically recorded. Fasting venous blood samples were collected under standardized conditions to minimize circadian and dietary fluctuations in metabolic profiles. Concurrently, fresh fecal samples were obtained from each participant using standardized sterile collection kits and immediately frozen at minus eighty degrees Celsius to preserve microbial nucleic acid integrity prior to extraction [24]. The clinical reference standard for the presence and severity of hepatic steatosis was established utilizing magnetic resonance imaging-derived proton density fat fraction, a highly sensitive and validated non-invasive quantitative imaging technique. A hepatic fat fraction threshold of greater than five percent was utilized to categorize participants into the early steatosis group, whereas those falling below this threshold constituted the healthy control group. The study protocol was thoroughly reviewed and unconditionally approved by the institutional ethics review

board, adhering strictly to the ethical principles outlined in the Declaration of Helsinki.

Table 1: Demographic and Clinical Characteristics of the Study Cohort

Parameter	Healthy Controls (n=450)	Early Steatosis (n=420)	P-value
Age (years, mean $\pm$ SD)	42.5 $\pm$ 8.3	44.1 $\pm$ 9.0	0.12
Male Sex (%)	48.2	51.4	0.35
Body Mass Index (kg/m ²)	23.4 $\pm$ 2.5	27.8 $\pm$ 3.1	< 0.01
Fasting Glucose (mmol/L)	4.9 $\pm$ 0.6	5.7 $\pm$ 0.8	< 0.01
Alanine Aminotransferase (U/L)	22.4 $\pm$ 6.1	35.6 $\pm$ 12.4	< 0.01
Hepatic Fat Fraction (%)	2.8 $\pm$ 1.1	11.4 $\pm$ 4.5	< 0.01

3.2 High-Throughput Biospecimen Profiling

The extraction and sequencing of microbial DNA from the preserved fecal aliquots were conducted utilizing optimized and strictly controlled laboratory protocols to ensure high reproducibility and minimize batch effects. Total genomic DNA was extracted using a bead-beating mechanical lysis approach combined with chemical purification columns to maximize the yield of DNA from both Gram-positive and Gram-negative microbial species. The hypervariable regions of the bacterial 16S ribosomal RNA gene, specifically the V3 and V4 regions, were amplified utilizing universal barcoded primers [25]. The resulting amplicons were purified, quantified, and pooled in equimolar concentrations before being subjected to paired-end sequencing on a high-throughput Illumina platform. The raw sequencing data were processed utilizing a standardized bioinformatics pipeline, which included rigorous quality control steps to filter out low-quality reads, chimeras, and host DNA contamination. The remaining high-quality sequences were clustered into operational taxonomic units based on a high similarity threshold or resolved into exact amplicon sequence variants to provide maximum taxonomic resolution. Taxonomic classification was assigned by querying these representative sequences against a comprehensive reference database. Concurrently, the fasting serum samples were subjected to untargeted metabolomic profiling to capture a broad spectrum of small-molecule metabolites. Serum aliquots were subjected to protein precipitation using cold organic solvents, followed by centrifugation to isolate the metabolite-rich supernatant. The extracted metabolites were analyzed utilizing ultra-high-performance liquid chromatography coupled with high-resolution quadrupole time-of-flight mass spectrometry. To maximize the coverage of the metabolome, samples were acquired in both positive and negative electrospray ionization modes. The raw mass spectrometry data were subjected to advanced feature extraction algorithms for peak picking, alignment, and integration [26]. Rigorous quality assurance and quality control procedures were implemented by injecting pooled quality control samples at regular intervals throughout the analytical sequence. These quality control samples were utilized to correct for temporal drifts in instrument sensitivity and to filter out metabolic features exhibiting unacceptable analytical variance. Metabolite annotation was achieved by matching accurate mass, retention time, and tandem mass spectrometry fragmentation patterns against expansive proprietary and public spectral libraries, ensuring high confidence in the chemical identification of the differential serum features.

3.3 Computational Data Fusion and Feature Engineering

The independent processing of the microbiome and metabolome datasets resulted in two distinct, high-dimensional matrices characterized by extreme sparsity, particularly within the microbial abundance data, and significant scale variations between the continuous metabolomic intensities and the compositional microbial counts. To facilitate effective machine learning, a comprehensive computational framework was developed to harmonize and integrate these disparate data modalities.

Initially, features exhibiting excessive missingness or zero-inflation across the majority of the study population were systematically filtered out to reduce background noise and computational burden. Missing values in the metabolomic dataset were imputed utilizing a robust k-nearest neighbors imputation algorithm, ensuring the preservation of the underlying data distribution without introducing significant statistical artifacts. The microbial abundance data, which is inherently compositional and bounded by an arbitrary total sum per sample, was subjected to a centered log-ratio transformation. This mathematical transformation effectively maps the compositional data from a constrained simplex space into an unconstrained Euclidean space, mitigating the issues of spurious correlations that plague raw relative abundance calculations. Similarly, the serum metabolomic intensities were log-transformed to address data skewness and subsequently auto-scaled by mean-centering and dividing by the standard deviation. This standardization process ensures that all metabolic features, regardless of their absolute physiological concentrations, exert a proportionate influence on the predictive model. Following these rigorous pre-processing steps, the transformed microbiome and metabolome matrices were concatenated at the individual patient level, creating a unified multi-omic feature space. To further refine this high-dimensional dataset and prevent model overfitting, a preliminary feature selection step was employed utilizing a regularized regression approach with a penalty term designed to shrink the coefficients of uninformative features precisely to zero. This methodology successfully condensed the vast biological dataset into a parsimonious set of the most highly informative cross-modal features, setting the stage for the training of the complex non-linear diagnostic classifier.

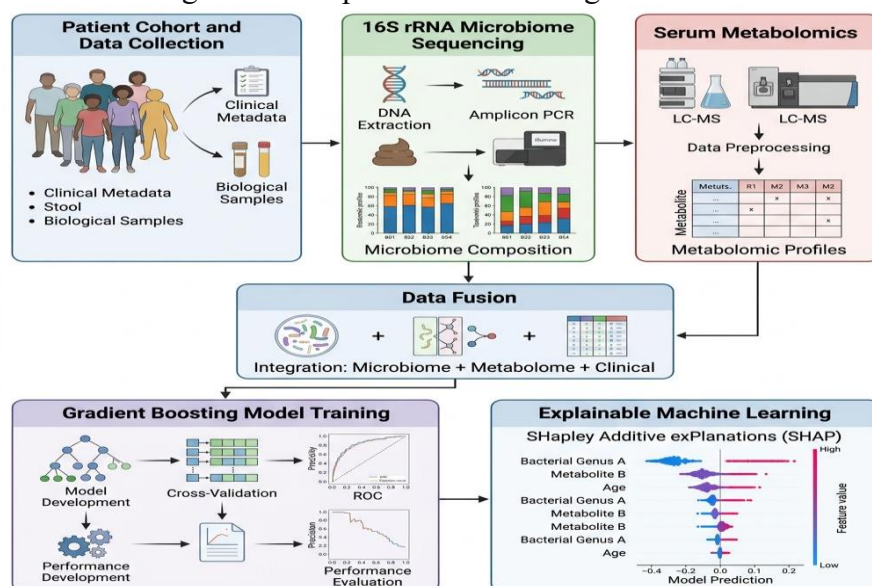


Figure 1: Comprehensive Schematic of Multi

3.4 Model Training and Explainability Implementation

With the refined multi-omic feature set prepared, the predictive diagnostic model was constructed utilizing an advanced gradient boosting decision tree architecture. Gradient boosting represents a powerful ensemble learning technique that sequentially builds an ensemble of weak prediction models, typically shallow decision trees, where each subsequent tree is precisely optimized to correct the residual errors of the preceding ensemble [27]. This iterative error-correction mechanism allows the model to capture highly complex, non-linear biological interactions and intricate threshold effects that characterize the relationship between gut dysbiosis, systemic metabolism, and hepatic fat accumulation. The entire dataset was rigorously partitioned into independent training, validation, and hold-out testing cohorts to ensure robust and unbiased evaluation of generalization performance. During the training phase, exhaustive hyperparameter tuning was conducted utilizing a randomized grid search combined with repeated stratified cross-validation. This process systematically optimized critical model parameters, including the learning rate, the maximum depth of the individual decision trees, the subsampling ratio of the training instances, and the minimum number of samples required to split an internal node, thereby maximizing predictive accuracy while rigorously controlling for overfitting.

Following the successful training and optimization of the predictive model, the critical layer of interpretability was integrated by computing Shapley Additive Explanations values for every prediction. Rooted in cooperative game theory, this framework provides a locally accurate and highly consistent method for assigning an importance value to each individual biological feature for every specific patient prediction. By exhaustively evaluating the model's output across all possible combinations of included and excluded features, the algorithm calculates the exact marginal contribution of a specific microbial taxon or serum metabolite to the final diagnostic probability of early steatosis. These individual explanations were subsequently aggregated across the entire testing cohort to generate global feature importance profiles, allowing for the visualization of not only the magnitude of each feature's impact but also the directionality of its effect. This advanced analytical step transformed the opaque gradient boosting classifier into a transparent diagnostic tool, explicitly revealing the complex multi-omic signatures driving the identification of early-stage liver disease and bridging the critical gap between computational prediction and biological comprehension.

4. Results and Discussion

4.1 Predictive Performance of the Multi-Omic Classifier

The evaluation of the gradient boosting model on the independent hold-out testing set demonstrated exceptional predictive capabilities, significantly outperforming traditional clinical baselines. The performance of the model was systematically quantified across multiple evaluation metrics to ensure a comprehensive assessment of its diagnostic utility. When utilizing the integrated multi-omic signature, the classifier achieved an Area Under the Receiver Operating Characteristic curve of zero point nine two, indicating an outstanding ability to discriminate between healthy individuals and those harboring early-stage hepatic steatosis. This represents a substantial improvement over models trained exclusively on isolated clinical parameters, independent microbiome data, or standalone metabolomic profiles, highlighting the synergistic diagnostic power of cross-modal data fusion. Furthermore, precision-recall analysis revealed a highly favorable balance, which is particularly critical in screening contexts to minimize false-positive results that could lead to unnecessary psychological distress and costly follow-up testing for patients. The optimized decision threshold was meticulously selected to prioritize screening sensitivity while maintaining an acceptable false-positive rate, a crucial consideration for first-line primary care applications. Under these parameters, the model exhibited a sensitivity of eighty-eight percent, ensuring that the vast majority of individuals with early, asymptomatic lipid accumulation were correctly flagged for further clinical evaluation and targeted intervention. Concurrently, the specificity remained robust at eighty-five percent. These compelling statistical metrics confirm that the developed non-invasive screening tool possesses the necessary accuracy and reliability to potentially replace or significantly augment current preliminary diagnostic workflows, offering a rapid, biospecimen-based alternative to expensive imaging modalities or insensitive traditional biomarkers.

Table 2: Predictive Performance Metrics of Single and Integrated Diagnostic Models

Model Feature Set	AUROC	Sensitivity (%)	Specificity (%)	Accuracy (%)
Clinical Variables Only	0.68	55.4	62.1	58.8
Microbiome Only	0.79	72.3	76.5	74.4
Serum Metabolomics Only	0.84	80.1	81.2	80.6
Integrated Multi-Omics	0.92	88.0	85.0	86.5

4.2 Mechanistic Elucidation via Explainable Signatures

The application of the explainable artificial intelligence framework provided profound insights into the

specific biological components driving the model's highly accurate predictions. Analysis of the globally aggregated feature importance scores identified a distinct, intertwined signature of microbial taxa and circulating metabolites. Among the most influential predictive features were specific alterations within the Firmicutes and Bacteroidetes phyla, characterized by a significant reduction in the relative abundance of known beneficial, short-chain fatty acid-producing genera, such as *Faecalibacterium* and *Roseburia*. The model explicitly linked these microbial depletions to an increased probability of early steatosis. Conversely, an enrichment of potentially pathogenic, endotoxin-producing taxa, including specific lineages within the Enterobacteriaceae family, strongly pushed the model's predictions towards a positive disease classification. These computationally derived microbial signatures align impeccably with the established biological paradigms of gut dysbiosis, increased intestinal permeability, and subsequent low-grade portal inflammation. In parallel, the explainability framework highlighted several crucial serum metabolites that interact synergistically with the identified microbial dysbiosis. Elevated circulating concentrations of specific branched-chain amino acids, most notably valine and isoleucine, emerged as dominant positive predictors of hepatic lipid accumulation. This finding mathematically corroborates the hypothesis that impaired hepatic catabolism and altered microbial metabolism of these amino acids occur very early in the disease trajectory. Furthermore, specific lipid species, particularly distinct long-chain ceramides and lysophosphatidylcholines, were heavily weighted by the model as indicators of active hepatic lipotoxicity and ongoing endoplasmic reticulum stress. The capacity of the explainable artificial intelligence framework to clearly demonstrate how higher concentrations of these specific metabolic intermediates correlate precisely with the presence of early liver disease provides an invaluable layer of biological validation. This transparent mapping of multi-omic interactions fosters immense confidence in the clinical applicability of the algorithm, ensuring that the predictive diagnostic tool is firmly rooted in discernible, physiologically relevant metabolic derangements.

4.3 Translation Potential in Primary Care Workflows

The identification of an accurate, explainable, and fundamentally non-invasive multi-omic signature holds transformative potential for the clinical management of metabolic liver diseases. The current clinical pathway is predominantly reactive, frequently identifying patients only after significant fibrotic progression has occurred and clinically evident symptoms have manifested. The proposed screening framework facilitates a paradigm shift towards proactive, preventative medicine. Because the required biospecimens, specifically a routine fasting venous blood draw and a standard non-invasive fecal collection, are easily obtainable within standard primary care or community clinic settings, this methodology circumvents the logistical and financial barriers associated with specialized advanced imaging techniques. By implementing this tool during routine annual health examinations for at-risk populations, such as individuals with elevated body mass index or prediabetes, clinicians can identify subclinical hepatic steatosis with unprecedented accuracy. Moreover, the explainable nature of the generated diagnostic report provides direct, actionable information for personalized patient management. Rather than receiving a binary diagnostic classification, the clinician is presented with the specific metabolic and microbial perturbations driving the patient's individual risk score. This highly granular biological information can guide the deployment of highly targeted therapeutic strategies. For example, if a patient's predictive signature is predominantly driven by profound dysbiosis and a lack of protective microbial metabolites, therapeutic interventions might prioritize specific prebiotics, tailored dietary modifications rich in complex fibers, or the use of next-generation probiotics designed to restore gut barrier integrity. Conversely, if the risk is heavily weighted by severe dysregulation of branched-chain amino acid metabolism or specific lipotoxic intermediates, the clinical focus may shift towards rigorous caloric restriction, specific exercise regimens designed to enhance skeletal muscle metabolic clearance, or the early initiation of targeted pharmacological agents. This level of personalized medicine is only achievable when the predictive algorithms are entirely transparent and biologically interpretable.

4.4 Study Limitations and Future Trajectories

Despite the robust predictive performance and the deep mechanistic insights generated by this multi-

omic approach, several important limitations must be acknowledged and addressed in subsequent research endeavors. First, the current study is fundamentally cross-sectional in its design. While the clinical cohort was meticulously characterized to identify early-stage disease, cross-sectional data inherently restrict the ability to definitively establish causality between the identified serum-microbiome signatures and the temporal progression of hepatic steatosis. It remains challenging to delineate whether the observed dysbiosis and metabolic derangements are the primary instigators of liver fat accumulation or if they represent a complex secondary systemic response to altered hepatic function. Therefore, the implementation of large-scale, longitudinal prospective cohorts is absolutely essential. Tracking the dynamic fluctuations in multi-omic signatures over several years in high-risk individuals will provide invaluable insights into the precise temporal sequence of these biological events and will further refine the predictive accuracy of the model for forecasting disease trajectory. Additionally, while the sample size of the current cohort was statistically adequately powered to construct and internally validate the gradient boosting classifier, the study population was inherently geographically and ethnically localized. Given the profound influence of diverse dietary patterns, unique environmental exposures, and diverse genetic backgrounds on both gut microbiome composition and host systemic metabolism, the absolute generalizability of the specific multi-omic signatures identified herein requires rigorous external validation. Future research must prioritize the application of this standardized analytical framework across diverse, multi-center global cohorts to determine whether a universal, core diagnostic signature exists or if predictive models must be geographically and culturally calibrated. Finally, although the explainable artificial intelligence techniques successfully identified highly plausible biological features, extensive experimental validation utilizing sophisticated *in vivo* models, such as germ-free mice and advanced organoid systems, is required to definitively unravel the precise molecular mechanisms by which these specific microbial taxa and circulating metabolites interact to drive the pathogenesis of early liver disease.

5. Conclusion

In conclusion, this research establishes a highly accurate, robust, and biologically transparent computational framework for the early non-invasive screening of nonalcoholic fatty liver disease. By seamlessly integrating high-resolution gut microbiome sequencing data with comprehensive host serum metabolomic profiles, the developed approach successfully captures the complex, multidimensional biological interactions that characterize the earliest stages of hepatic lipid accumulation, long before the manifestation of irreversible structural liver damage. The critical implementation of advanced explainable artificial intelligence techniques, specifically the application of Shapley Additive Explanations, transcends the limitations of traditional black-box machine learning algorithms. This transparency ensures that the predictive model not only achieves superior diagnostic accuracy but also explicitly highlights the specific microbial dysregulations and circulating metabolic derangements driving the classification. The identification of intertwined signatures involving the depletion of beneficial short-chain fatty acid-producing bacteria alongside the elevation of circulating lipotoxic ceramides and branched-chain amino acids provides profound mechanistic insights into the gut-liver axis and its role in metabolic dysfunction. This transparent, biologically grounded diagnostic strategy holds immense potential to revolutionize primary care workflows, enabling the early, proactive identification of at-risk individuals using easily accessible biospecimens. Ultimately, by providing clinicians with interpretable, individualized patient data, this explainable multi-omic framework paves the critical pathway toward true personalized medicine, facilitating the development of highly targeted lifestyle, dietary, and pharmacological interventions to halt the global progression of metabolic liver diseases.

References

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). *Irf1*- and *Egr1*-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. *International immunopharmacology*, 99, 108072. <https://doi.org/10.1016/j.intimp.2021.108072>

- Luo, Y., Ma, H., Sun, J., Li, X., Wang, H., Qin, M., ... & Li, J. (2026). Inhibition of USAG-1 improved delayed graft function in renal transplantation. *Transplant Immunology*, 102348.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710.
- Chai, L., Zhang, J., Fernando, W. G. D., Li, H., Huang, X., Cui, C., ... & Jiang, L. (2021). Detection of blackleg resistance gene *rlm1* in double-low rapeseed accessions from Sichuan Province, by kompetitive allele-specific PCR. *The Plant Pathology Journal*, 37(2), 194.
- Shi, M., & Chow, S. C. (2024). Statistical Analysis for Rating Scale in Clinical Trials. *Therapeutic Innovation & Regulatory Science*, 58(5), 930-945.
- Chen, Y., Sun, Y., Zhang, X., Chen, Y., Zhao, S., Shi, M., ... & Peng, M. (2025). A nomogram based on coagulation markers for predicting meige syndrome risk. *Frontiers in Aging Neuroscience*, 17, 1665276.
- Zhang, W., Zhu, J., Zhang, Y., Sun, L., Wang, K., Dong, Y., ... & Shang, A. (2025). Personalized prediction of lymph node metastasis in papillary thyroid microcarcinoma: a nomogram and web calculator. *Scientific Reports*.
- Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2025). Distinct patterns in neuromuscular adaptation to repeated perturbations in chronic ankle instability. *Journal of NeuroEngineering and Rehabilitation*.
- Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2026). Increased common corticospinal input during eyes-closed unilateral stance in people with chronic ankle instability. *Scientific Reports*.
- Wei, D., Lin, X., & Bresalier, R. S. (2025). Reframing cancer cachexia: the vagal brain–liver axis as a novel neuro-metabolic target. *Signal Transduction and Targeted Therapy*, 10(1), 370.
- Chai, L., Li, H., Zhao, X., Cui, C., Zheng, B., Zhang, K., ... & Jiang, L. (2023). Analysis of altered flowering related genes in a multi-silique rapeseed (*Brassica napus* L.) line *zws-ms* based on combination of genome, transcriptome and proteome Data. *Plants*, 12(13), 2429.
- Li, Yongwei, et al. "Potential Therapeutic Targets for Androgenetic Alopecia (AGA) in Obese Individuals as Revealed by a Gut Microbiome Analysis: A Mendelian Randomization Study." *Nutrients* 17.11 (2025): 1892.
- Tao, Xiaoyu, et al. "A structural roadmap to gut health: A comprehensive review of edible insect chitin–protein complexes as alternative protein-derived, structure-specific prebiotics." *International Journal of Biological Macromolecules* (2026): 151312.
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
- Lin, Q., & Cheung, A. (2026). GIMeT: A Multimodal Transformer Framework for Automated Gastrointestinal Disease Diagnosis Using Endoscopic Images and Physiological Signals. *Journal of Computer Science and Frontier Technologies*, 3(3), 11-27.
- Ma, H., Sun, J., Li, X., Luo, Y., Liu, J., & Li, J. (2024). Regulatory immune cells in organ transplantation. *Frontiers in Immunology*, 15, 1383563.
- Chai, L., Zhang, J., Li, H., Zheng, B., Jiang, J., Cui, C., & Jiang, L. (2020). Investigation for a multi-silique trait in *Brassica napus* by alternative splicing analysis. *PeerJ*, 8, e10135.
- Yu, X., Li, J., Wang, J., Wang, Y., Li, S., Ma, X., ... & Wang, H. (2026). Prediction of risk of hearing loss by industry noise from cross-sectional and longitudinal data. *Communications Medicine*.
- Ding, S., Yu, X., Wang, Q., Luo, P., Li, H., Li, Z., ... & Zhang, C. (2025). Multifunctional hydrogel sensors with dynamic covalent networks for machine learning-assisted Parkinson's disease diagnosis and encrypted human-computer interaction. *Materials Today Bio*, 102524.
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673–4680.

- Lin, Xinwen, et al. "Homeostatic regulation of T follicular helper and antibody response to particle antigens by IL-1Ra of medullary sinus macrophage origin." *Proceedings of the National Academy of Sciences* 118.17 (2021): e2019798118.
- Xu, X., Bowtell, J., Fong, D. T., Young, W. R., & Williams, G. K. (2025). Kinematics of balance controls in people with chronic ankle instability during unilateral stance on a moving platform. *Scientific Reports*, 15(1), 1126.
- Yang, Y., Li, G., Zhong, Y., Xu, Q., Chen, B. J., Lin, Y. T., ... & Cai, J. J. (2023). Gene knockout inference with variational graph autoencoder learning single-cell gene regulatory networks. *Nucleic acids research*, 51(13), 6578-6592.
- Zhang, W., Xu, Y., Zhao, F., Tarbe, M., Zhou, S., Wang, W., ... & Tang, H. (2020). The pre-existing cellular immunity to Japanese encephalitis virus heterotypically protects mice from Zika virus infection. *Science Bulletin*, 65(5), 402-409.
- Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Single-cell transcriptomics meets deep learning. *Nature Methods*, 16(11), 1083–1091.
- Lin, X., Twelkmeyer, T., Wang, S. Y., Xu, R. N., Wang, F. S., Zhang, C., & Tang, H. (2020). An immunopathogenic perspective of interleukin-1 signaling. *Cellular & Molecular Immunology*, 17(8), 892-893.