American Journal of Psychiatric Rehabilitation

Expert Opinion Article

Integrating Multi-Omics Data and AI For Early Detection And Personalized Treatment of Colorectal Cancer

Gowthamm Mandala[1*]

[1*]Research Student, Purdue University, Gmandala@purdue.edu

Abstract

Colorectal cancer (CRC) is the fourth leading cause of cancer-related morbidity and mortality worldwide. More than half of the CRC cases are due to preventable causes. Aberrant gut microbiota has been found to play an important role in the early phase of CRC by causing chronic inflammation. However, this residual risk combined with the lack of patient-friendly methods can limit early detection, intervention success, and post-intervention monitoring. This creates an urgent need for additional innovative point-of-care tests that could distinguish early preclinical cancers from precancers, both of which could be associated with immune dysbiosis. Artificial intelligence (AI) tools are expected to assist with these challenges, as vehicles for these assays have come to the forefront, with promising results in other epigenetic tests. In this review, we focus on innovative colorectal cancer risk biomaterials to detect these early events, highlighting the potential of artificial intelligence (AI) to aid in data integration and develop better tools, ultimately leading to patient-tailored treatments for better colorectal cancer outcomes. As synergistic material platforms, carbon nanopipettes (CNPs) are then discussed, along with circuit nanosensors, with a focused discussion on early CRC disease detection.

Keywords: Colorectal Cancer, Cancer Morbidity, Cancer Mortality, Preventable Causes, Gut Microbiota, Chronic Inflammation, Early Detection, Point-of-Care Tests, Immune Dysbiosis, Artificial Intelligence, Epigenetic Tests, Risk Biomaterials, Data Integration, Patient-Tailored Treatments, Disease Outcomes, Carbon Nanopipettes, Circuit Nanosensors, Innovative Diagnostics, CRC Monitoring, Intervention Success.

1. Introduction

In recent decades, the changes in eating habits and lifestyle have led to an increase in the morbidity of colorectal cancer (CRC). Patients in the early stage of the disease usually have no obvious clinical symptoms and lack the effective performance of the potential screening methods, resulting in up to 40% of patients having developed into more advanced stages at the time of initial diagnosis. However, patients in different stages have different prognoses. Contrary to the 90% five-year survival rate in stage 0 CRC, the rates in stages II and III are merely 60% and 10%, respectively. Therefore, accurate early diagnosis is vital in improving prognosis. Currently, there are two main early screening and diagnosis methods: fecal occult blood test (FOBT) and colonoscopy. However, both methods have their limitations. For FOBT, the sensitivity is low and it can only be used for the early stage of CRC and cancer recurrence monitoring. To some extent, colonoscopy may not be accepted by the general population due to its invasiveness and possible complications, and its poor compliance may also lead to the neglect of early clinical symptoms.

CRC is a multi-step process involving the accumulation of different genetic and epigenetic factors. These factors affect different molecular levels such as DNA, RNA, microRNA, protein, and metabolites in the tumor initiation, invasion, and metastasis stages. Directed targeted therapy uses the specific drug action mechanism to directly counter or inhibit tumor growth. However, due to the complexity of driving factors in different stages of cancer, personalized therapy should be designed for different patients by studying the genomes of different patients. Therefore, the early detection and individualized therapy for CRC based on the multi-omics data become a powerful tool. Recent research found that big data technology could help the diagnosis and treatment of cancer by promoting the analysis of multi-omics data of patients. However, currently, the data are not well integrated, and we lack digital means to quantitatively evaluate the weight of both single omics and multi-omics features. Taking advantage of artificial intelligence can help us find key changes and evaluate correlations in multi-omics data.

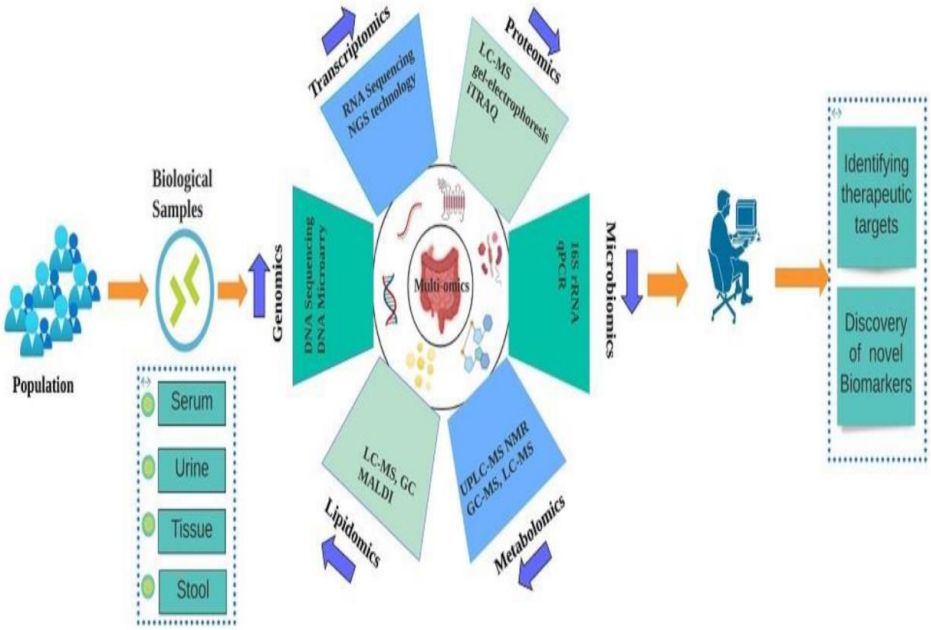

Fig 1 : Multi-Omics Approaches in Colorectal Cancer Screening and Diagnosis

This will promote multi-omics data integration and discover a possible mechanism of precision therapy in the early stage of CRC.

This review provides innovative thinking for the integrative analysis of multi-omics data and the potential applications of artificial intelligence for early detection and personalized treatment of CRC. We first discuss the different modes of single omics and multi-omics data and explain their innovations and difficulties. Then, the specific integrative strategies of multi-omics data are outlined, followed by a discussion on the potential power of artificial intelligence in furthering the exploration of the Landau paradigm, which proposes both single omics and multi-omics fit the Weighted Network Model. This new thinking provides future directions for early detection and personalized treatment of CRC.

1.1. Fundamental Concepts

The constellation of genes expressed in a cell determines its functions. Changes in the expressiveness and the levels of gene expression, or mutations and epigenetic changes, can lead to a vast range of disorders, including the formation of malignant tumors. The potential use of gene expression patterns for diagnosis is investigated when data from a specific disease becomes available, and candidate genes are sought for this application. Instead of relying on data from a single test tissue to discover gene expression patterns that discriminate between different disease states, 'multi-omics' involves the integration of information from various levels of gene expression and sequence data. It enables screening for gene expression patterns in normal tissue from individuals who later develop some specific disease. By conducting a variety of measurements on a range of tissues in different environments, opportunities for early detection and personalized treatment are increased.

The comprehensive mapping of all the genes and their transcripts involved in the biological functions of an organism is known as the transcriptome. Genome-wide transcriptome profiling can be completed after extracting the total RNA from a given tissue, converting it to complementary DNA using reverse transcription, and isolating each transcript. In microarrays, the cDNA is hybridized to a grid of different sequence-specific oligonucleotide probes. In contrast, next-generation sequencing reads the cDNAs directly to create a high-resolution map of the transcriptome. With RNA-seq data, it can be discovered not only what specific genes are expressed in a cell or tissue but also which ones are differentially regulated under conditions such as stress or disease. Sequencing analysis allows the relative levels of spliced isoforms or the abundance of novel RNA sequences to be determined. Furthermore, without having to design a custom set of microarray probes or in need of an annotated gene model, RNA-seq can be used to discover novel transcripts.

1.2. Contextual Framework

The clinical context for developing diagnostic or early screening tools is given by two scenarios. Firstly, in patients with a hereditary or genetic disposition, a first-degree family history of colorectal cancer or polyposis, a high expected lifetime risk of developing the disease, or having multiple polyps or advanced adenomas in previous clinical or imaging examinations are identified. Secondly, these symptoms can occur spontaneously without former

symptoms in the patient years before, for example, the development of adenomas or the onset of cancer. Only the second type of patient for whom no pre-warning is available would benefit from non-invasive screening.

The clinical objective of systems analytical research for cancer diagnosis is to identify all potentially relevant patient-specific target entities that provide significant information about the progression and stage of the disease and the interaction of the disease with the organism, with potential relevance for therapeutic options in the context of multi-omics data for the design of a plasma-based cancer diagnosis. This is different from ongoing clinical research, where the aim of in vitro detection with defined sensitivity and selectivity often uses predefined biomarkers and their controls in case vs. control settings of defined patients to compare a novel approach with gold standard methodologies.

Equation 1 : Multi-Omics Feature Integration for Risk Prediction

$$R_c = \sum_{i=1}^{N} \alpha_i \cdot O_i + \sum_{j=1}^{M} \beta_j \cdot C_j$$

where:
R_c = Colorectal Cancer Risk Score,
O_i = Omics Feature i (e.g., genomic, transcriptomic, proteomic),
C_j = Clinical Feature j (e.g., age, lifestyle factors),
α_i, β_j = Weight Coefficients,
N, M = Number of Omics and Clinical Features.

2. Understanding Colorectal Cancer

Colorectal cancer (CRC) is the third most common cancer in men and the second in women. Colorectal cancer has a 90% cure rate if detected at an early stage. However, the 5-year survival rate of late-stage patients decreases to 20%. Therefore, developing early detection technologies is of great importance. In recent years, there have been many examples of multi-omics approaches revealing the pathological molecular mechanisms of CRC. Although these studies have substantially extended our knowledge of the mechanisms of the functional changes associated with CRC, the molecular basis underlying the early diagnosis of CRC has only been limitedly explored so far.

Colorectal cancer (CRC) is the third most common cancer in men and the second in women. Both men and women are affected almost equally in terms of the incidence of non-hereditary CRC. Currently, there are about a million new cases detected worldwide. Colorectal cancer has a 90% cure rate if detected at an early stage. However, the 5-year survival rate of late-stage patients decreases to 20%. Therefore, developing early detection technologies is of great importance. In recent years, there have been many examples of multi-omics approaches revealing the pathological molecular mechanisms of CRC. Although these studies have substantially extended our knowledge of the mechanisms of the functional changes associated with CRC, the molecular basis underlying the early diagnosis of CRC has only been limitedly explored so far.

2.1. Epidemiology and Risk Factors

Colorectal cancer (CRC) is the fourth leading cause of cancer worldwide, with one and a half million new cases being diagnosed each year. The third in terms of incidence in men and the second in women, CRC shows large differences in incidence and mortality between populations, with a high proportion of associated mortality in some countries. Its two major components are colon cancer and rectal cancer. The risk of developing CRC during an individual's life is 5%, both for males and females. While the risk is slightly lower for developing colon cancer, it is much higher for rectal cancer. There are significant gender and race differences in CRC incidence and mortality, perhaps due to lifestyle or environmental factors. Some well-studied risk factors for CRC are high consumption of red and processed meats, low consumption of fruits and vegetables, low physical activity, overweight, and smoking and alcohol consumption.

Cancer development is a very complex process. This multifactorial and stochastic character is underlined by the growing number of potential disease agents known to date, such as epidemiologic or environmental agents. Individual-specific susceptibility is due to the combination of common genetic risk factors and rare genetic mutations. The presence of polygenic and/or sporadic risk factors, besides genetics, can be a determinant in disease development. Epigenetic and environmental factors can contribute to a person's overall susceptibility to cancer, but exposure to certain agents leads to a higher risk for some individuals. Some compatibility with our body, associated with particular lifestyles, contribute to genetic changes that can be reverted before a disease manifests. Signals identifying those agents in the environment that induce neoplasms via distinct pathways should be generated and need to precede cancer development in susceptible individuals. Treatment modalities and prognostic methods are highly enriched by early detection; thus, the development of new methods has widely been considered important. Early detection is a key strategy for reducing CRC, which depends on continuous control and early treatment to prevent cancer development in patients with adenomatous polyps.

2.2. Pathophysiology

Intestinal stem cells (ISCs) are the primary epithelial stem cells present in all of the longitudinal crypts. Several abundant genetic mutations, both in oncogenes and tumor suppressor genes, are recognized as determinants of colorectal cancer. The failure of the Wnt signaling pathway promotes colon cancer development, and the mutation in APC is the most prevalent cause of sporadic colonic adenoma and cancer. APC has a function of degrading β-catenin in the Wnt signaling pathway, and if polyposis occurs, the APC mutation results in the accumulation of free β-catenin in the cell nucleus to activate the Wnt signaling pathway and stimulate cell proliferation. β-catenin translocation to the nucleus induces a genomic state change by altering the expression of over 300 Wnt target genes, especially the genes associated with EMT, stemness, and cancer development. Because 80% of the primary colonic adenomas and cancers have the APC truncation mutation, a truncated APC and c-Myc pathway have been expressed for the initiation of malignant polyps and cancer.

BMP signaling is another pathway that suppresses Wnt-dependent self-renewal, and it plays a significant role in maintaining the stem cell niche by controlling the gradient of BMP activity from the bottom to the top of the crypt. Upon midcrypt localization, BMP4, acting as a differentiation-driving force, induces apoptosis in the stem and transit-amplifier (TA) cells, and the depletion of the BMP gradient causes overproliferation or cancer. On the contrary, overactivation of the BMP pathway results from the inactivation mutation of BMPR2, BMPR1A, SMAD1, and SMAD4, thus leading to the blocking of stem differentiation to form the colon cancer stem cell (CSC) to stimulate cancer development. Conversely, if the BMP pathway is overactivated, the stem differentiation of crypts is stimulated to suppress cancer progression. TGFβR2 controls the development and homeostasis of colonic crypts. The loss of function of TGFβR2 in the TGFβ signaling pathway causes colonic polyp formation and aggravates the WNT pathway to promote the progression of colonic polyps. TGFBR2 functions as a tumor suppressor gene that can promote TA cell formation by triggering increased cell cycling for the promotion of cellular differentiation and maintenance of the crypt diploid stem niche. The loss-of-heterozygosity mutation of TGFBR2 causes predisposing cancer phenotypes with an elevated cancer risk or predisposition. On the other hand, the TGF-β receptor can regulate cell motility for chemotaxis and cell-matrix adhesion, exacerbating the development of cancer through mutations in TGFBR2 that are associated with increased migration and invasiveness resulting from EMT by the increased activity of K-Ras. In sum, TGFβ signaling plays a dual role during cancer initiation and progression by repressing the initiation and promoting the malignant progression to CSCs in late-stage tumors. The downregulation of the BMP and TGFBR signaling pathways changes the microenvironment to suppress the process of differentiation of colonic crypts and promote the process of tumor initiation and growth.

2.3. Current Diagnostic Methods

Screening for CRC can be done in several ways. These include fecal immunochemical testing for CRC, fecal testing and colonoscopy, flexible sigmoidoscopy, CT colonography, and stool DNA testing as test options, and double contrast barium enema or visual inspection of the colon by sigmoidoscopy or colonoscopy. Tests were both patient-directed and in-office. The strengths and weaknesses of different screening methods and their use in specific situations have been evaluated. FIT has more limited sensitivity and specificity than colonoscopy, but it has a higher sensitivity for detecting cancer and a similar or slightly higher sensitivity for detecting advanced adenomas. Nearly half of adult Americans are not getting tested at the recommended age. Inadequate information or access to these various tests could be the reason why testing rates are low.

To reduce the burden of diagnostic and time-consuming lab procedures, there have been growing efforts to develop less invasive and more efficient diagnostic methods. The main advantage of DNA methylation is the greater stability of the plasma DNA, leading to the potential for better sample handling and storage. The current diagnostic technique for hereditary CRC syndromes is based on the score, which evaluates as a modifier mutation. Another method involves immunohistochemistry to evaluate the loss of proteins and serves as a quick and cost-effective test. These tests, however, only evaluate one or two proteins and require labor-intensive analysis to give a reliable and specific molecular diagnosis on a large scale. The sum of these has improved the clinical approach in CRC diagnosis but has also increased therapy value. Because most clinical tests are based on the evaluation of just a single characteristic, the use of multiple tests can improve the accuracy of tumor staging.

3. Overview of Multi-Omics Approaches

Recent advances in high-throughput screening and sequencing technologies facilitate the concurrent measurement of multiple types of molecular data, such as DNA methylation, chromatin accessibility, histone modifications, gene expression, and protein expression, allowing researchers to gain unprecedented access to a holistic picture of a complex biological system. However, previous research has produced only a fragmented understanding of the molecular mechanisms fueling

CRC development because it investigated such mechanisms using individual molecular components and did not fully elucidate the connections or coordination between them. Single-omics data analysis of CRC only provides partial information on disease-related mechanisms, and omics-based methylation data analysis partially characterized CRC EMT, tumor heterogeneity, and the sample heterogeneity of mCRC. Integrated analysis of DNA methylation and RNA expression entirely reveals the association between DNA methylation and function. Integrating diverse omics data may provide important insights into these multi-level molecular mechanisms, go beyond reductionism, and facilitate the identification and interpretation of complex disease-related mechanisms. Multi-omics data integration-based analysis of the mRNA pool only captures part of the gene expression regulation relationship, and multi-omics data integration-based analysis of the histone modification pool and chromatin accessibility pool partially characterized CRC cell type heterogeneity. Therefore, investigating molecular mechanisms with a multi-omics approach is of critical importance for obtaining an in-depth view of the progression of CRC.

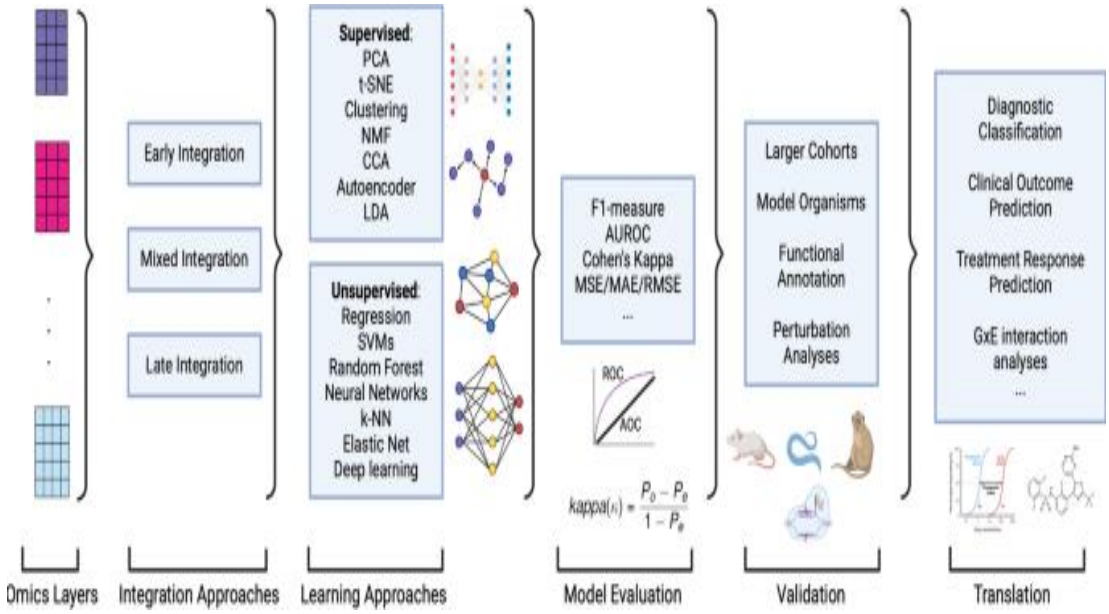

Fig 2 : Multi-omics approaches for understanding gene-environment

3.1. Genomics

Colon cancer is characterized by gene mutations in KRAS, BRAF, and various mismatch repair genes like MLH1, MSH2, MSH6, PMS2, or a DNA repair gene. Besides gene-related mutations, other genetic variants of single nucleotide polymorphisms may also affect colorectal cancer risk in genes such as PIK3CA, BCL2, ATM, SEC61G, MMP1, MMP2, MMP3, MMP7, ORMDL3, DSC, DSC1, TSLP, LEMD2, SEMA6A, and SEMG1. Initially, the roles of these gene mutations, variants, and gene expressions were mainly uncovered by using case-control study-based approaches. In addition, miRNA also plays a crucial role in early detection and genomic surveillance of colorectal cancer, and in the clinical phenotype and development of adenomas or different stages of the tumor.

The dysregulation of miRNA causes the alternation of oncogenes, tumor suppressor genes, and other cancer-related genes to promote oncogenic proliferation, clonal expansion, and immune invasion.

By integrating information such as data, metadata, epigenomic peaks, the active transcription signal, expression, mRNA, and RNA-binding protein motifs, the expression profile of individual exons within a gene can be obtained. Some exon-isoforms will result in the alteration of the open reading frame of a gene, consequently affecting the interaction with its downstream targets or premature termination codons. Such exon-isoform–gene structure and protein binding interaction heterogeneity may give rise to gene-specific functions and contribute to wall-peeling identification, drug discovery, customized drug targeting, and a better understanding of the underlying molecular mechanism of late-recurrence patients. Single-cell RNA sequencing of peripheral blood mononuclear cells and transcript serum exome sequencing were applied to investigate the early detection of colorectal cancer by the change in gene expression. Additionally, the lymphocyte–monocyte ratio in peripheral blood and the detection of mutation in serum are potential biomarkers for early diagnosis, monitoring of tumor recurrence, and predicting the prognosis of colorectal cancer.

3.2. Transcriptomics

Profiling technologies at the RNA level have facilitated the human transcriptome resources, which are important to understanding how genes behave in the context of cellular homeostasis and diseases. Genes encode messenger RNAs (mRNAs) - the tiny copy from DNA that serves as the blueprint for proteins to accomplish cellular functions. In addition to the coding of mRNA, genes may generate long non-coding RNA, micro RNA, and other small noncoding RNA, which also have unique roles and functions in cellular biology. Classification of different cellular mRNA expression states helps characterize functional states and, thus, has been one of the major focuses in biology. Genetic alterations such as mutations and intimate functions, including cell-type specific signal networks and drug responses, can be revealed by alternative expression of transcripts. RNA is one of the diagnostic markers according to the cancer hallmark and its association with somatic mutations. The cancer transcriptome landscape and transcriptome diversity across the common types of cancers have been widely characterized.

Transcriptomics has contributed to tumor-initiating cell thesis as well. For solid cancers, the spatial and temporal transcriptome heterogeneity in primary and metastatic sites and the cell compositions within tumors have been characterized based on carefully collected cohorts, as well as public pathological sections. A comprehensive atlas of cancer cell types within the tumor microenvironment has been built from advanced bulk RNA-seq, single-cell RNA-seq, spatially resolved RNA profiling, and in situ hybridization techniques. Transcriptome variability from diverse disease states is a goal of transcriptome diagnostics. With a large number of well-organized gene signatures and disease labeling by therapeutic responses, profiling technologies have reshaped disease taxonomy and therapeutic strategies. In summary, transcriptomics can reveal cancer pathophysiology and has enormous promise in personalized treatments for improved patient survival.

3.3. Proteomics

Proteomics provides the most direct evidence of physiological changes within individuals by studying the expressions, modifications, localization, and interactions of proteins. Detected at their functional level, their high diversity, and their representation of the actual occurrence of a disease, proteins are considered the best and most direct targets for clinical diagnosis and prognosis. Various challenges in proteomics, such as low protein abundance, large dynamic range, and high complexity, led researchers to adopt multiple statistical models, prediction models, and AI algorithms to analyze, interpret, visualize, investigate, and manage the resulting fragment masses and peak lists. Among the data-dependent acquisition strategy and data-independent acquisition strategy, the targeted MS acquisition strategy showed the most robust protein quantification for colon cancer tissue and adjacent tissue. Overcoming the challenges of serum background protein suppression and the unknown fragmentation of the target ions, a detection system utilizing human serum absorption to concentrate the analytes, the microchip electrophoresis pre-separation to suppress the high-abundance proteins, and the adopted photonic molecule-based fragmentation to acquire the information of amino acid residue oxidation showed superior sensitivity and specificity in detecting oxidized proteins at diluted serum state. Cost-effective and high-throughput protein crystallization screening displayed a 30% increase in successful protein crystallization when it used the machine learning method to estimate potential crystallization conditions from sparse and pulsed ultraviolet spectroscopy data under various phase diagrams compared to traditional visual observation. Analyzing the time-course protein S-glutathionylation data to explore the potential regulators of proteins with similar S-glutathionylation dynamics also showed the advantages of machine learning. With its high-efficiency velocity-based diagnosis for imaging mass spectrometry detected colorectal cancer tissues, a series of machine learning-based analysis algorithms opened up a new era for the broad application of metabolomics technologies containing mass spectrometry.

3.4. Metabolomics

Early detection and personalized treatment of colorectal cancer (CRC) have been highlighted to improve clinical outcomes. With the rapid development of large-scale genomic, transcriptomic, epigenetic, and metabolomic techniques in biomedicine and artificial intelligence, the integration of multi-omics data can reveal the comprehensive molecular and cellular characteristics of CRC and provide us with the possibility to develop early detection and personalized treatment strategies for patients. These strategies enable better profiling of identical and fraternal twins with CRC compared to monozygotic twins, which indicates the importance of targeting the molecular mechanisms of individual patients with the same stage of colorectal cancer rather than general molecular characteristics. In this study, the recent progress of integrating multi-omics data to obtain the specific molecular and cellular characteristics of CRC patients, including the integration of multi-omics to understand the tumor microenvironment and the advanced artificial

intelligence for deriving other types of omics data for studies, has been reviewed thoroughly.

Metabolomics, which measures the concentrations of low molecular weight compounds generated by metabolic systems such as cancer metabolic phenotypes, offers therapeutic targets for CRC. These compounds are usually referred to as metabolites and are considered the end products of gene expression. In CRC, metabolomics can produce genetic processes activated before the appearance of morphologically detectable changes and support screening populations for early-stage disease with high benefit-risk ratios. Recent technical advances allow for conducting untargeted investigations of metabolites using liquid and gas chromatography coupled with mass spectrometry. With the rapid development of metabolomics, global metabolic assessment has become functional for the pharmaceutical identification of the early stages of CRC. The number of studies focused on the qualitative and quantitative estimation of the metabolome to determine potential discrimination between patients at different stages has increased recently.

4. Artificial Intelligence in Healthcare

Artificial intelligence (AI) and machine learning (ML) have a growing role in the field of healthcare, with the potential to shape future healthcare systems. Although there are still significant challenges, the improved interpretation of biological data to enhance decision-making for patients is a major application, and it is anticipated that these approaches will aid in making personalized medicine a reality. Many AI-based methods have been applied to healthcare to facilitate the increasingly large, diverse, and publicly available health-related datasets. The rapidly changing medical knowledge is also a good environment for AI development. Yet limitations do exist concerning data privacy, security, cost, and ethical guidelines. AI-driven models can indeed find biomarkers for early detection and precise treatment planning of colorectal cancer patients more effectively, and here is a summary of the best updates on AI tools.

Treatment strategies based on clinical and molecular characteristics for patients with colorectal cancer (CRC) are needed to improve the therapeutic efficacy of personalized treatments. Due to multiple levels of organization and pathways of human physiological and pathophysiological states, genomics, transcriptomics, epigenomics, proteomics, metabolomics, and microbiomes have always been widely used as biomarkers to reveal various aspects of biological systems. The rapidly growing number of multi-omics data provides an unprecedented opportunity to explore complex relationships between diverse human complex multi-omics data by using artificial intelligence (AI). Such large-scale data offers the potential to decipher cancer heterogeneity and underlying regulatory mechanisms, discover clinically actionable cancer subtypes, and uncover patient-specific or disease-specific potential biomarkers and therapeutic targets. AI could effectively exploit complementary information contained in these large amounts of high-dimensional data to infer the complex relationship between different layers, as well as handle multimodal fusion representing the partial observations from multi-layered interactions. With integrative analysis of multi-omics data, it is possible to construct comprehensive models of the cancer system, analyze pre-neoplastic stages, and move toward the goal of finding potential therapeutic solutions and biomarkers in the context of systems medicine. Therefore, due to the potential impact on the early prognosis and treatment of CRC patients, it is critical to develop effective AI-driven panomics approaches for recognizing multi-omics detailed information related to cancer initiation, metastasis, and drug resistance.

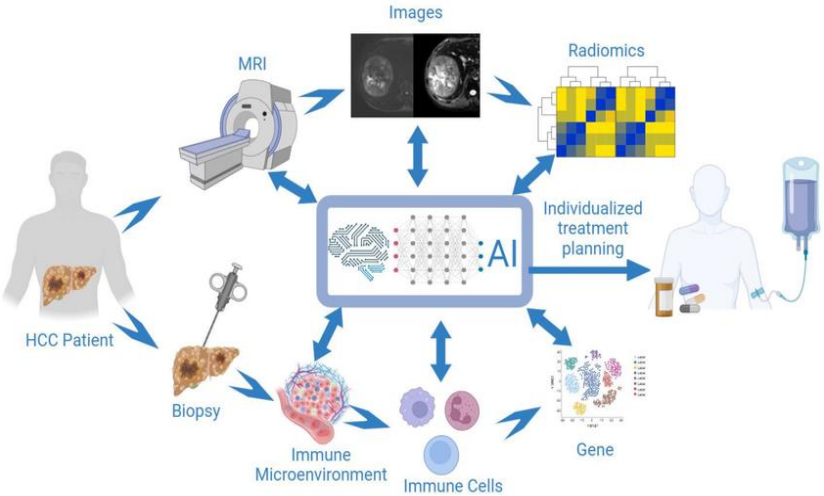

Fig 3 : Artificial Intelligence (AI) and Multi-Omics

4.1. Machine Learning Techniques

Artificial intelligence (AI) models contribute greatly to the multi-omics analysis of CRC by integrating transcriptomics and clinical information, displaying high performance in the prediction of cancer patients, treatment effects, drug sensitivity, survival, relapse, and so forth. We reviewed various machine learning methods including general and cancer type-specific predictors, feature selection and hierarchical layer-specific predictors, database-based predictors, network-based methods, single-omics-based models, personalized training set models, models for drug sensitivity, survival, and relapse predictors, logical representation of specific gene expression outcomes, pipeline-based models, and visual methods towards addressing the transcriptomic and clinical information to solve specific problems in CRC. Furthermore, deep learning methods including convolutional neural networks, recurrent neural networks, residual networks, hierarchical deep learning, advanced deep learning, deep learning for cancer risk prediction, deep learning with gene signals, transfer learning for improved prediction, and autoencoders used in feature extraction and phenotype prediction over large-scale transcriptomic data, taken either separately or together with clinical data, were investigated for the development of advanced AI techniques for newly integrated biological information in CRC.

4.2. Deep Learning Applications

Deep learning-based single-cell DNA sequencing algorithms, instead of computationally determining the copy number of gene expression, use Print-C information at the cell level directly. Deep learning architectures have also been employed for the denoising of chemical probes to infer potentially hidden physical signals and serve data analysis and validation. Yet another novel single-DL model for the imputation of spatially resolved transcriptomics, termed SeSPA, was supported by a large-scale single-cell spatial transcriptome dataset built on tissue that is home to over 40 different cell types, demonstrating the ability to impute spatially resolved transcriptomics and estimate cell types based on spot-based molecular data. However, the recent integration of multimodal human-single cell data reveals tissue-wide patient-specific differences in RNA isoforms.

Deep learning in cancer detection has gained a new way. Pre-trained models provide a groundbreaking strategy that trains a model on an image-rich task and reuses the pre-trained model for other tasks. At the time of diagnosis, transfer learning can be used for non-invasive subtyping of colorectal cancer. Then, inaccurate radiological findings were demonstrated, in which transfer learning outperformed CNN. Deep learning has been used in the prognostic outcome model development of colorectal cancer in different ways. On the other hand, three prognostic classification models for colorectal cancer are loaded with single-cell RNA sequencing data and supervised. Single-cell consensus clustering, in some others. Combining individual layer sequencing from the medical student dataset allows us to provide the next identified properties.

Equation 2 : AI-Driven Biomarker-Based Early Detection Model

$$P_d = \frac{1}{1 + e^{-\left(\sum_{k=1}^{L} \gamma_k \cdot B_k\right)}}$$

where:

P_d = Probability of Disease Presence,

B_k = Biomarker Expression Level k,

γ_k = Weight Coefficient for Biomarker k,

L = Number of Biomarkers.

4.3. Natural Language Processing

Natural language processing methods are used to analyze unstructured data from different sources, including electronic health record data such as clinical notes and pathology notes. Electronic health records contain important clinical data of patients, and the information is stored in an unstructured or semistructured format. Text mining techniques have been widely applied to extract useful information from unstructured text. Currently, different machine learning models, including rule-based systems, are employed as a basis for clinical information extraction. A bag-of-words model could be utilized to convert the words into a numerical value for further modeling. Topic models could further capture the hidden structure of the unstructured text data. The generative ability of recurrent neural networks and deep generative models could all be models for potentially producing better performance in clinical document classification and relation extraction. The major barriers to deep learning in clinical NLP include the scarcity of labeled data, data imbalance, and non-negligible noise. Practical methods and strategies have been proposed to solve these challenges, but more efforts are still needed.

5. Integrating Multi-Omics Data

Despite substantial progress in computational tools and technologies, understanding disease systems on multi-omics is still challenging. Until now, molecular efforts have generally focused on the identification of individual and direct interactions between

biomolecules within one type of omics data. However, the cellular system is a large and complex network, containing multi-layer interactions between DNA, RNA, proteins, and metabolites. For example, genome mutations lead to changes in gene products such as transcriptome, translatome, proteome, and metabolome, which are responsible for fundamental properties. Transcriptome changes determine protein features, translating to altered interaction networks and the identification of cellular substrates and phenotype outcomes. Moreover, single omics data provide an incomplete view of the molecular and cellular heterogeneity of individuals, failing to accurately understand the onset, progression, and clinical contradictions of diseases. A large-scale comprehensive integration of multi-omics would provide unprecedented insights into a systemic view of cellular functions, alterations, interactions, and heterogeneity of individuals under different physiological and pathological states.

Multi-omics data, in combination with AI, can provide deep insights into disease systems. Multi-omics are datasets of all key biomolecules for an individual, including genomic, transcriptomic, proteomic, metabolomic data, and more. It represents an individual as a network model, constraining the analysis at the systems level. The biomedical phenotype can be informed based on the interpretability of associated reference data, while the multi-omics profiles hold the prediction of high-level effects of drugs or disease treatments. If and when single types of omics data resolve ambiguous or contradictory diagnoses, multi-omics data filling in for each other would resolve conflicts more effectively, thereby increasing the diagnostic yield without a need for additional, often costly tests, and avoiding excessive invasiveness, pain, and suffering for individuals. With the increasing availability of multi-omics profiles, and clinical and microbiome data linking phenotypes and clinical outcomes in local and global populations, multi-omics can coordinate actions for risk predictions, tumor subtyping, and optimization of individual, even personalized treatments. Only with the generation of new knowledge in association with clinical symptoms and multi-omics would a hospital use these datasets for better-quality diagnostics.

5.1. Data Collection and Preprocessing

The stages of colorectal cancer (CRC) consist of normal, adenoma, early stage adenocarcinoma, lymph node and hematogenous spreading, and advanced cancer. The early detection and radical treatment of CRC are the key to improving the survival rate and quality of life of patients, among which the application of AI technology and blood multi-omics testing are of concern to scholars. The application of AI and blood multi-omics tests in early diagnosis and personalized treatment mainly includes biomarker discovery, early screening, grading, and treatment response evaluation.

The clinical data and multi-omics data of CRC were collected and preprocessed, including the clinical data such as age, sex, and treatment response of the patients and the multi-omics data such as DNA methylation, copy number variation, mRNA expression, long non-coding RNA, microRNA, and exosome data. First of all, we can obtain the patients' data mainly through various databases. Among them, one database mainly includes clinical and multi-omics data of 14 tumor types, which contain basic patient and tumor information and integrate gene expression, microRNA, DNA methylation, copy number variation, proteins, long non-coding RNAs, and other relevant data. Besides, another dataset is an important supplementary dataset, covering seven cancer types, including proteomics, genomics, transcriptomics, and pathology data, focusing on identifying candidate protein markers for cancer diagnosis and treatment.

5.2. Data Integration Techniques

Here, we outline recent data integration techniques. Briefly, the data mining of massive amounts of heterogeneous multi-omics data is divided into phenotype-oriented data mining and association-oriented data mining. Phenotype-oriented data mining aggregates related data for phenotypes of interest via known characteristics of various data domains or directly uses multi-domain data to train the model to solve the problem of interest. Association-oriented data mining combines data in a principled way and automatically initiates the analysis of associations between different data modalities. In essence, this technique uncovers potential cooperating omics biomarkers for the phenotype of known characteristics. This information is then used to interpret the biological and therapeutic effects of intervention in a specific individual and explore the progression of cancer.

In recent years, strategies for jointly analyzing genetic and molecular phenotypic traits have been studied. Individual studies with multitrait analytics examine the relationship between latent traits. This research identified gene modules that are functionally conserved across several species and suggested that this function is conserved throughout cancer development. Dataset partitioning predicts molecular traits in the heart and blood. To extend this research, we develop a pipeline, including experiment-level multi-domain integrative partitioning of structured datasets. We identify the association between experiment activities and global DNA methylome conformers in two separate studies, leading to the identification of candidate genes that may be regulated by DNA methylation and, importantly, inhibition of the heart failure target. Through conserved regulatory elements, we identify

common regulatory determinants of gene expression and translation.

5.3. Challenges in Integration

Despite the increasing availability of multi-omics data from the same set of patient samples, their use in basic research, clinical applications, translational research, and drug discovery remains modest. These many opportunities are not matched by equivalent progress, especially in machine learning analyses and in terms of smart data innovation. Both the challenges and the benefits to be gained are substantial. Essential strategies for the routine application of integrated multi-omics data are to develop testable, actionable hypotheses about gene function and regulation and to provide causal inferences based on the structure of the data while considering the multimodality and large size of omics data.

The challenges include the large dimensionality, low replication, highly imbalanced labels, high heterogeneity, abundance of noise, relations between sources and types of omics and technologies, also compared with retrospective genomics and transcriptomics results in terms of mapping the relations between genes, transcription factors, non-coding RNAs, variants of the two sources of DNA, RNA, proteins, metabolites, epi transcriptome, epigenome, quantitative differences, regulatory motifs, co-expression, co-methylation, pairwise relationships, joint or independent analysis. To enhance data integration, we require the science behind the quantity and quality of multi-omics tissue samples for patient subgroups undergoing different treatments and patient outcomes. This necessitates further smart big data innovation in terms of the digitization of cell biobank repositories and electronic health records, and state-of-the-art science in terms of tumor biology and knowledge about the tumor microenvironment across a spectrum of data modalities and histological characteristics, architectural patterns, tissue microenvironment, and each class of malignancy. We describe these smart data strategies and require machine learning model interpretability for a potential multi-omics cancer gene integration multitask learning system, not just on a computational level but on a precision level.

6. AI Models for Early Detection

Once the predictive model of all stages of CRC from the multi-omics data is available, its principle can be extended to the AI model for the early detection of CRC. Although stool DNA testing and colonoscopy are current clinical strategies for the early detection of CRC, a noninvasive early detection method for CRC would still be desirable due to the underperformance of the above tests, patient intolerance, availability, accessibility, and cost. Importantly, being successful in the early detection of early-stage CRC should start with the understanding of the underlying mechanism and the identification of multi-omics markers and the AI model of early-stage CRC. Besides, the AI model should have high sensitivity and specificity to accurately detect early-stage CRC and should be comprehensively tested and confirmed using a wide array of multi-omics signatures. In addition, it should be thoroughly validated using stool samples from patients with adenomas, and CRC, and matched normal samples. Upon validation of the AI model, it can be integrated into a sensitive, specific, and cost-effective kit for the early detection of CRC using stool or serum samples. With the kit, a reader can be developed to accurately read the AI model analysis. Its users can easily interpret and implement the AI model and even detect underlying mutations for early-stage cancer cells, allowing personalized therapeutic interventions.

6.1. Predictive Modeling

A central theme in early detection and precision medicine is building predictive models that can predict the future development of diseases or the progression of diseases. Accurate predictions of complex questions like these require consideration of not only clinical risk factors and single-omic biomarkers but also multi-omics and environmental/lifestyle information collected in our Colon Cancer Family Registry. While clinicians may use clinical risk models to help inform further diagnostic testing, they rarely use additional information for an individual patient. Developing and translating accurate research models that demonstrate broadly useful performance in clinical settings and decision-making has faced several obstacles.

There are relatively few reports that incorporate more complex omic data, and among those, relatively few can make meaningful predictions within the clinical setting. A few models can use gene expression of the primary tumor to predict the recurrence of that tumor. While the recurrence of the primary tumor is certainly important, the pressing question for patients is whether subsequent primary tumors will occur; gene expression of the primary tumor cannot address that question. A recent summary identified such risk models, but none have been recommended for clinical practice because of limitations such as small sample size, acting on populations of patients that are not clinically important, using data that cannot be obtained in the clinical setting, and so on. For a real patient with newly diagnosed colorectal cancer, omics risk models are not part of standard care because they are not yet considered ready for clinical translation. Crucial for omic-based risk models to be used in prospective and retrospective clinical studies is a

validated demonstration of their broad utility for the patients they are designed to serve.

6.2. Risk Stratification

Colorectal cancer (CRC) is experienced as a preventable and treatable disease, with a low incidence at the early stages and a high curative ratio. Due to the lack of overt symptoms and the use of invasive tests, most CRC cases are incidentally diagnosed at a late stage. Therefore, early detection is key to curing CRC. Patients with early-stage disease have a high cure rate after receiving radical surgery or endoscopic surgical treatment. Given the favorable prognosis in patients classified with early-stage colorectal cancer, preoperative evaluations are required to develop effective, individualized treatment recommendations to provide early, effective, and personalized treatment strategies to individual patients. In the early stages of colorectal cancer, patients with low risks could directly receive minimally invasive surgeries. Patients with intermediate or high risks of distant metastasis have to undergo imaging tests before receiving radical surgery. Preoperative risk stratification is essential for diagnosis and the tailored management of colorectal cancer. Evidence now supports that proposed risk stratification provides a prognostic characteristic that colorectal patients need. Accordingly, we turn genomic characteristics into general clinical practice. Preoperative risk stratification allows the surgeon to develop an individual treatment plan, which maximizes the benefit to the patient, avoids the risk of complications, and effectively prevents the occurrence of potential adverse events.

6.3. Screening Protocols

Different countries have different national CRC screening guidelines, mainly because of the limited information available on individual colorectal cancer risks, such as genetic variants, environmental factors, or coexisting diseases. As we now understand, the European guidelines are representative. They recommend setting up general population-based screening, using the fecal immunochemical test at ages between 50 and 74 (the usual recommendation is a two-year interval). It also suggests that other clinical factors should be constantly taken into account when evaluating a patient (e.g., individual data/risk, age). It stresses that symptom-driven testing may be considered outside the screening program for patients at higher risk for CRC due to the number of false-positive test results and the potentially higher rate of missed cancer.

However, we think these screening protocols are not enough. Based on the multi-omics differences between benign and malignant tumors and the large number of disease-causing pathways, more efficient and effective screening algorithms should include more considerations.

7. Personalized Treatment Approaches

Emerging molecular insights have opened up new ways to treat patients with colorectal cancer by suggesting a more personalized approach. It is now clear that, although in the past colorectal cancers were thought of as a single type of disease entity, diversity exists in patients, driven by mutations and the gut microbiome within both healthy and cancerous tissues. The idea of a drug matched to patients' genes, and then mapped against the genes in their tumors, seems a very reasonable approach to the effective treatment of cancer, avoiding unnecessary toxicity and side effects. Simply put, the likelihood of success of a therapy can be maximized by personalizing treatments. In the last decades, the overall survival of patients with colorectal cancer has increased and their quality of life improved. This is in part due to the new drugs that have been introduced. In the coming years, personalized medicine will add additional leaders to support patients.

In addition to the use of DNA-based sequencing for personalized medicine, gene expression data have been used for decision-making. The first companion diagnostic assay for a drug approved was based on the expression of the gene ERBB2. Here, we explore the literature concerning potential future achievements in the personalization of treatments for colorectal cancer. We focus on the application of personalized medicine considering the genomic profile and the microbiome status, not only in more advanced cases but also in the setting of curative-intent treatment for metastasis, as well as in the treatment of all patients in general.

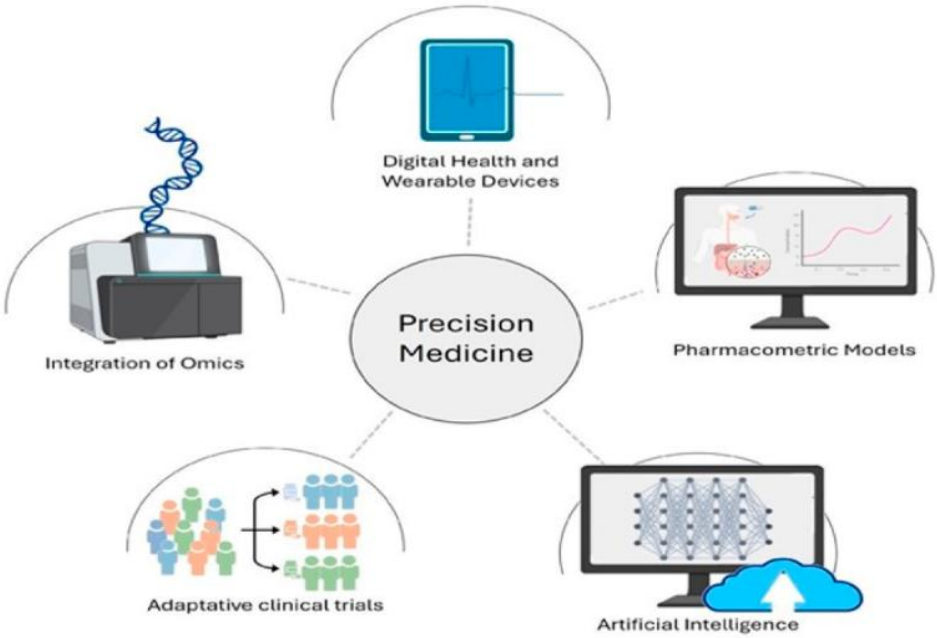

Fig 4 : Revolutionizing Personalized Medicine

7.1. Targeted Therapies

The molecular biology and genetics of CRCs have been extensively characterized. This knowledge has resulted in a new classification of CRC with clinical and therapeutic implications, offering significant promise of personalized therapeutic interventions. The distinction between right- and left-sided colonic tumors, as well as the tumors' biomolecular profiling, impacts treatment options and the efficacy of therapies using anti-epidermal growth factor receptor-targeted monoclonal antibodies. These issues are potential tools for molecular diagnosis and should provide evidence to support personalized medicine in CRC. In CRC, the multikinase inhibitors, or trifluridine and tipiracil, or regorafenib, offer better options for patients who have previously received all other available therapies, including systemic chemotherapy with or without anti-EGFR/VEGF agents.

Inhibition of the programmed death-1 immune checkpoint is proving to be clinically effective in increasing the overall survival of patients with metastatic CRC. Certain therapies have been approved for patients with certain cancers with microsatellite instability-high or deficient DNA mismatch repair solid tumors that have progressed following prior treatment. In addition, other therapies have also been approved for patients with microsatellite instability-high or deficient DNA mismatch repair metastatic CRC that has progressed following the presence of fluoropyrimidine, oxaliplatin, and irinotecan. It is known that immune checkpoint therapies can achieve durable responses in patients with microsatellite instability-high tumors. However, it is also important to highlight that patients with microsatellite-low and microsatellite-stable tumors could respond either from single-agent checkpoint blockade or combination immunotherapy. The biological rationale for the use of checkpoint inhibitors and other immunotherapies in patients with microsatellite-stable and microsatellite-low tumors is already being explored and could be, shortly, an important option for patients who are affected by overlapping survival and toxicity.

7.2. Immunotherapy

Although immunotherapies have revolutionized cancer treatment, not every patient derives long-lasting benefits. Many colorectal cancers, particularly in the microsatellite stable biotype, have relatively non-immunoreactive phenotypes, and the general efficacy of existing immunotherapies has not been as successful in microsatellite stable colorectal cancer as microsatellite instability-high subtypes. Our multi-omics data and machine learning models have been successfully employed to explore the specific and common factors underpinning the clinical success of existing immunotherapies and the great potential of novel personalized immunotherapy approaches for colorectal cancer patients. Our project is at an early stage, but the very strong cancer-specific studies have been our main breakthrough in the field. In addition to our cancer inflammation project, there was a part related to imprinting N6-methyladenosine into RNA at play in a shared framework for feature reduction in some omics-related projects. A dimension reduction analysis that unveiled the patient-relevant incontinent features of RNA was performed. These features are closely related to clinical features and pathways in the domain of immunoreaction. Drug

candidates that target these genes have been successfully identified, opening up the possibility for an alternative therapeutic strategy for personalized cancer immunotherapy. These potential targets are closely linked to immune checkpoints and tumor immunoreaction, which provide very strong insights for personalized treatment of colorectal tumors.

7.3. Combination Therapies

Another way of addressing the limitations of monotherapies in colorectal cancer is through combination therapies. These have the added potential benefit of reducing toxicity while increasing efficiency. Traditional monotherapy-based cancer treatments have been challenged by the development of resistance. Combination therapies composed of different or similar types of anticancer agents have long been used as a strategy for combating drug resistance. Furthermore, multiple therapeutic compounds belonging to different drugs or the accumulation of similar drugs address different molecular targets to strengthen and consolidate the antitumor efficacy of single drugs while weakening the generality and attenuation of drug resistance in colorectal cancer therapy.

Several known drugs in colorectal cancer integrate the above ideas. Recent research has combined the known effective inhibitors to form a powerful drug combination that was effective against the primary colorectal cancer cells carrying the BRAF mutation. It was indicated that the combination of multiple inhibitors partially repressed the MEK/ERK pathway compared to a single inhibitor in these cells, accompanied by the inhibition of multiple other signaling pathways. The cell-intrinsic resistance of colorectal cancers to the chemotherapeutic agent was also studied by the drug combination regimens. These drug regimens combined targeted the pathways and offered a new potential method to improve chemotherapy.

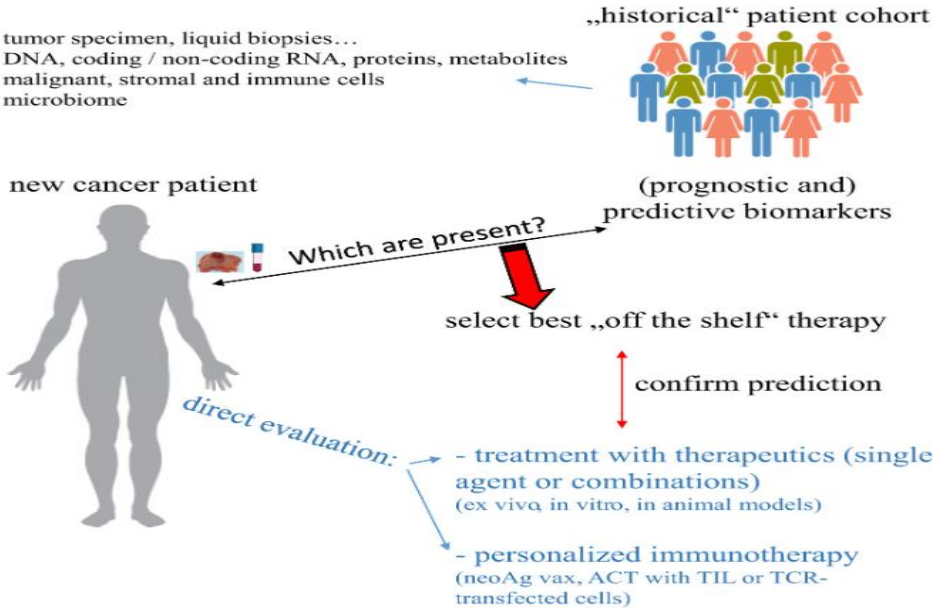

Fig 5 : Combination of multiple omics techniques for a personalized therapy or treatment selection

Multiscale Integrative Genomic Analysis was performed to recognize gene signatures associated with drug resistance in treated patients, which also indicated the interactions between the axis and autophagy activity in driving resistance to therapy, offering opportunities for targeted therapies adjunct with treatment, such as inhibitors that could enhance sensitivity in these patients. KRAS mutant cells upon strong signaling inhibitor combinations can dramatically shut down the feedback signaling reactivation and lead to death during treatment with dual inhibitors. This further illustrated that brutal inhibition of feedback reactivation could be the best solution to overcome drug resistance. These studies suggested that the combination of multi-target inhibitors presents the possibility of adding promising drug candidates in anticancer regimens by coordinating drug combinations and repurposing drug actions, with the same logic to develop new therapeutic strategies according to specific resistant pathways for colorectal cancer patients.

8. Case Studies

In the first example, to identify the genes contributing to the molecular classification, they applied a multi-modal autoencoder to find subtypes of colorectal cancer based on seven types of expression data. They searched for sets of genes that truly distinguish each subtype and derived a network-based method to identify genes that are informative concerning the disease. They identified reduced subnetworks of 15, 6, 7, and 22 genes that distinguish COADREAD classes 2, 3, 4, and 5, which correlate with patient survival. They also presented

a new way to visualize a high-dimensional subspace identified by an autoencoder, carefully examined in terms of unplanned subgroup analyses in the context of a mouse colon tumor microarray study. They hope that their approach can provide a useful tool to help guide future discovery research in this difficult-to-treat disease. In the second example, researchers designed an interpretable deep learning algorithm to predict 5-year overall survival in patients with non-metastatic colorectal cancer using gene expression, clinical, and treatment data. They used an encoded clinical vector consisting of baseline patient information—personal, demographic, and health information—encoded treatment features, continuous gene expression data, and a binary variable to incorporate biological subtypes and highlight circulating DNA mutation and methylation features. They transformed biological features into 2D feature maps that visualize how the model combines information, captures co-occurrence, and understands co-expressed gene patterns that are biologically meaningful. Their approach thus capitalizes on artificial intelligence to extract information from a large dataset, leading to models that are highly predictive, and describing and visualizing relationships between the feature sets. Their findings have implications not only for risk assessment of colorectal cancer patients but also for making personalized predictions of the oncological outcome using multi-omics molecular profiling.

8.1. Successful Integrations

The clinical management of colorectal cancer (CRC) can be enhanced by integrating multiple types of data derived from high-throughput omics experiments. This often requires the use of supervised, semi-supervised or transfer learning methods, which we summarize here. Some of the most successful tools for integrating multi-omics data to study pathways and diseases are those for the detection of differentially matched gene modules and disease-associated modules or regulator genes. We summarize successful approaches for the integration of multi-omics data in the context of two real-life examples, which involve metaproteomics and metabolomics profiles, or metabolic genes, metabolites, and metatranscriptomic data from the context of CRC. These approaches pinpoint similarities and differences in diverse CRC data from distinct populations and allow the effect of individual bacteria to be identified.

The increasing amount and heterogeneity of multi-omics data that is available for omics applications in colorectal cancer is pushing the development of taxonomy-based prognosis, driver discovery frameworks, or Metagenome-Wide Association Studies. Finally, integrating multi-omics data in the study of colorectal cancer often involves the use of state-of-the-art supervised, semi-supervised, or transfer machine learning methods. After some final remarks, we end by providing a database of gene signatures for the diagnosis and prognosis of colorectal cancer, to be used by the interested reader, with complete instructions on how to obtain the gene expression signatures.

Equation 3 : Personalized Treatment Optimization Using AI

$$T^* = \arg\max_{T} \sum_{t=1}^{T} (E_t - S_t - C_t)$$

where:

T^* = Optimal Treatment Plan,

E_t = Expected Treatment Effectiveness at Time t,

S_t = Side Effect Impact at Time t,

C_t = Treatment Cost at Time t,

T = Total Treatment Duration.

8.2. Failures and Lessons Learned

Despite the key successes of multi-omics data and AI, there remain notable failures and lessons learned. About multi-omics data, relatively few models have been validated through clinical trials, as opposed to current genomics models in common cancer diagnostic kits. Additionally, the small sample sizes do not cover the highly heterogeneous genetic features of patients. Existing models also largely depend on in vitro constructed vectors, while human cancer cells exist in a three-dimensional nature. On the other hand, few AI models have been validated by validation cohorts. There is still a lack of AI models that consistently take into account not only gene expression and methylation sequencing but also mutational status, to isolate mutation-specific rules for personalized precision medicine.

Importantly, relatively few efforts have been devoted to applying random omics at the cellular image level to cancer research and treatment. Currently, the most widely used techniques combine pre-affixed biomarkers, which are used to validate the precision of omics data and/or to guide active therapy. Such techniques have always presented difficulties when combining different omics in an integrative manner since each application introduces stochastic or context-specific issues. Data output techniques, including feature selection and synthetic generation,

often produce over-optimized models, leading to reproducibility issues. Furthermore, explanation techniques that produce improved feature interpretability of AI models are not easily and/or generally applicable, particularly to omics data.

9. Ethical Considerations

Early screening, detection, and personalization of treatment are primarily implemented with the completion of the omics of patients and progressed with AI algorithms, which may prompt ethical considerations such as data ethics, algorithm interpretability, transparency, and fairness. The generation of large-scale multi-omics data involves significant ethical issues centering on protection, permissions of DNA and genomic data, as well as the protection of patients in such research. The substantial initial investment in the formation of large-scale data for initial research for certain types of populations is likely to be paid off by subsequent health applications catering more to wealthy patients. Dependence on life-saving medical care or attention to such economic factors may result in the socialization of multimillion or multibillion-dollar data benefiting only a few richest taxpayers.

There are ethical concerns that can be raised from the employment of AI-based methods to generate and examine multi-omics data for early detection and precision treatment planning of colorectal cancer. Data ethics involves the rigorous obtaining of informed consent for multi-omics data generation and sharing. It largely depends on the fulfillment of the three principal tenets of research ethics, namely respect for persons, beneficence, and justice. The interpretability of deep learning is central to assessing if the results would be meaningful for casual end users, especially the patients, and the knowledge of whether the algorithm is applicable in a given patient's case, as well as the trust and final acceptability of the AI-based approach by the public, are integral components of the design of a clinically applicable algorithm to tackle colorectal cancer-related medical issues, keeping the welfare of patients, providers, and the public in mind.

9.1. Data Privacy

Technologies and disciplines contributing to precision medicine, such as genomics and data analytics, have also provided tools so the medical community can treat patients with greater personalization. This evolution has made it increasingly important to address the potential for misuse and misunderstanding of the increasingly large and complex datasets that come from individual biorepositories. For example, when multiple classes of disease are represented by various human tissues, there is always a risk that the most informative personalized medical tools can be used for discrimination or bias. An example is the possibility of deriving significant racial discrimination from factors related to virus spreading and immunity, both from bystander exposure and parent and embryo exposure. Because personalized, "race-blind" tools looking at single mutations may compound the dismissal of evolutionary histories overrepresented in other known infection-related molecular factors, such interventions should ideally be disadvantaged in the framework of single ancestry-drug interactions.

After subject identifiers, such as genotype and family history, have been replaced by extended phenotypic profiles, unique meta-information originating from affected patients' hometowns, workplaces, travel destinations, or disaster locations shall be excluded from the biorepository. Other data elements unique to one user, including patient-generated health records, lifestyle and coding information, or the specific detail in medication or other therapy databases, must be eradicated, de-identified, or hidden to ensure that re-identification attacks with statistical or other vulnerabilities are rendered increasingly unfeasible. With this multitude of personal data diversity, biorepositories remain a treasure trove of resources for life science research, optimization, and discovery to understand the genes' regulatory connections and the biochemical machinery in diseases such as cancer.

9.2. Bias in AI Algorithms

Bias in the process of identifying these biomarkers and the risks that come inherently with developing these models by integrating multi-omics sources has implications for healthcare systemic disparities in minority groups and women. Bias exists in AI algorithms, particularly in those embedded within various healthcare systems and institutions. Machine learning algorithms learn from data, and when biases exist within this data, they might be extrapolated and used by algorithms to make decisions that are biased and discriminatory toward individuals or groups of individuals. In other words, the inequalities of historical data are pervasive; algorithms do exactly what they have been trained to do. These biases show up along demographic markers such as gender, race, and age and, in the context of the healthcare space, can lead to disparities in treatment. Indeed, the models used to generate important healthcare conclusions and formulate treatment plans might be systematically receiving unproven variations of care based on socio-demographic parameters.

Bias in AI has been the subject of large controversy, and research groups and ethics boards alike are constantly considering this concept and how to protect against it. This is due to severe impacts on people's lives; imagine healthcare models systematically giving male patients preference over females, adding them to more medical tests, redistributing personalized treatments, or

recommending riskier surgeries, for example. Such is the problem with training AI models using biased datasets. The issue of bias extends far beyond historical data; from the formulation and expression of AI models' inherent function, the predictions differ significantly as the demographic and socioeconomic information of the patient changes. Infractions in the training and model fitting process could lead to potentially catastrophic results and legal issues for the institutions using the tools.

9.3. Informed Consent

As for the matters directly related to this research, we will obtain written consent for participation and ethical research from the research agency that was approved by the Bioethics and Safety Act. We will inform you in writing that the donor is acknowledged and used as an experimental subject in the research report. We will also inform you in writing that the stored samples will be used for the research and that the results and the developed technology will be used for commercial purposes. If the object of this information does not agree, the stored samples will be discarded. In addition, this study was exempted from the requirement for informed consent by the approved Institutional Review/Animal Care Board.

As for the clinical information provided by omics data, the clinical information data are shared worldwide among researchers and utilized for co-studies. The datasets used in the current study are available from the corresponding author upon reasonable request. Furthermore, the clinical information was analyzed and selected in this study to present the statistical significance of machine learning models, and the patients' information in the datasets was de-identified and clinically protected, according to the guidelines of the institution that approved the ethics. Therefore, the need for informed consent was waived. The study protocol conforms to the ethical guidelines and was approved by the Institutional Review Board.

10. Future Directions

The development of machine learning methods and AI can help build more advanced precision medicine strategies in oncology by integrating multi-omics data with the vast amount of experimental data available. One such system-level in silico predictive model is the Darwin-PC. Darwin-PC is an example of combined predictive computational systems biology and biostatistics-based deep learning methodologies. The Darwin-PC predicted PC scores can be used to test if consensus disease subtypes exhibit diversity in the proliferation program and to correlate the PC scores in human disease with known diagnostic tests. The generation of PC-predicted gene regulatory networks presents an actionable framework on gene function and the dynamical process in which tissue subtypes and cell populations emerge and cooperate to perform broader functions. Tracking the direction of the disease subtype with multi-omics data and the PC-predicted subnetworks will enrich the outcomes for diagnostics and translational drug discovery to treat underlying diseases.

In addition to the development of Darwin-PC, Darwin-PC extends the cell and tissue lineage trajectories by disentangling the potential heterogeneity within single-cell cognate lineages. This extensibility is useful for more accurate multi-scale analyses in human tissues and disease modeling. It also provides opportunities for personalized medicine by incorporating patient genotype and additional diagnostic parameters for more precise disease subclass definition. Over the next decade, combination therapies of drugs that target the dysfunctional tissue type-specific oncology program and the tissue microenvironment will prove more effective than single drug agents. The integration of Darwin-PC-predicted patient subnetworks with the RNA-seq data of patient-derived cancer cell lines that have varying responses to drugs is a promising direction to search for tissue-specific network signatures that can predict treatment responses in the future.

10.1. Technological Advancements

The majority of multi-omics studies in cancer research have focused on the integration of omics data, such as miRNA and mRNA expression data. Although higher-order multilevel molecular interaction programs are possible, avoiding negative and non-supervised relations, some interesting opportunities and future potential research directions arise with the integration of larger cohorts of multi-omics data. As single-omics studies may generate unreliable and often incomplete portraits of living systems, the reason for using multi-omics data in molecular interaction analysis is the fact that each molecule is only capable of performing exhaustive operations in small parameter spaces. Keywords are: bioinformatics, biomedical data analysis, cancer, data analysis, gene interaction, inference, integration, machine learning, molecular level, multi-omics, omics, pre-clinical, translational. In this chapter, we present a survey of methods that utilize multi-omics data for modeling gene interactions and the inference of interaction networks that facilitate such studies. We postulate that the need for large-scale data integration of genomic and other sources, for prediction, diagnosis, treatment, and the generation of biological insight with clinical potential, benefits from bioinformatics tools implementing recent developments in machine learning and deep learning strategies and models. Metagenomic analyses largely contribute to our understanding of individual microbiomes and are starting to provide insights into inter-individual

microbiome differences and structure. In the context of molecular-level studies, we propose a fair comparison framework for downstream analyses of multi-omics data. We have also designed and developed a multi-omics data integration framework, which enables the integrative analysis and comparison of multiple unique multi-omics data across independent cohorts in specific contexts.

10.2. Potential Research Areas

Here, we describe some areas for the integration of multi-omics data and the use of AI in CRC for early detection and precision medicine. Blood biomarkers are important for the early detection of CRC. Many optimistically select unique biomarkers for CRC detection, but usually none are FDA-approved. The urgent need to develop non-invasive blood biomarkers remains. Utilizing multi-omics data from the same CRC patient may also identify more robust blood biomarkers for early detection of CRC. However, a multimodal biological pattern heavily relies on AI approaches for high accuracy in prediction. Further developing interpretable AI approaches will enhance prediction reliability and the utility of biomarkers by producing more meaningful insights. Prognostic biomarkers could easily guide cancer treatment. Based on a tumorigenesis model, mutation-cluster counts should be included in prognostic biomarkers. Similar to the need for blood biomarkers, finding the right cluster is very important. The transcriptomic, epigenomic, and proteomic features have more associated gene functions. Thus, integrative multi-omics efforts may further refine prognostic biological signatures by using molecular functions. Early-stage treatment of CRC and the considerable clinical heterogeneity are not included in reaching the aim. As finding synergistic-additive genes plays a role together, integrative multi-omics data is crucial. Towards this, both other tools and human intelligence will play a role. According to the clinical therapeutic data, regulatory mechanisms and pathways are easily established and are crucial for the development of CRC. Towards this, statistical combinations and more comprehensive experimental designs from different layers of the data connected by accurate computational models will shed light on the use of pathways.

10.3. Collaboration Across Disciplines

What is required to solve more challenging and open problems, such as early detection of colorectal cancer, is more collaborative work - not just between researchers with complementary expertise but also harnessing resources and trust from the public, government, industry, and members of different disciplines and cultures. This research has been conducted with a combination of knowledge and expertise of researchers from different fields, including those in data modeling and analysis, wet lab biological experiments, and clinical validation and translation. Computing not only enabled significant data-driven advances to be made by the practical application of supervised learning on large-scale multi-omics data but also facilitated scientists of various disciplines to easily and systematically answer complex biological questions with a high level of accuracy and generalizability. Our current aim is to combine explanations and predictions from AI with evidence from wet-lab biology that can be applied in the real world to realistically measure a range of relevant clinical outcomes. It is also important to attract statistically driven research and number-oriented machine learning researchers to work on medically important research questions that have been traditionally addressed using traditional testing or rule-based stratification. AI could help gain fresh insights and discoveries in elucidating complex multifactorial etiology and pathogenesis in human diseases and become an additional option in the toolbox of experimental and observational epidemiology and developed science.

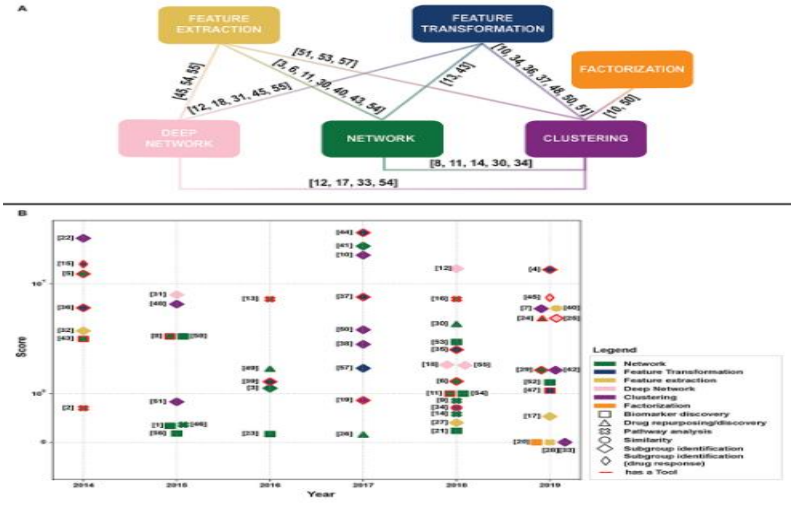

Fig 6 : Integrated Multi-Omics Analyses in Oncology

11. Conclusion

Colorectal cancer (CRC) is one of the most preventable cancers. Early detection of the disease and timely personalized treatments could reduce the hazard of lethal loss and burden on society. With the technological advances in decoding the genome and microbiome, the complex etiology of CRC has been explored. Multi-omics data, including genomics, transcriptomics, proteomics, metabolomics, and metagenomics, have revealed the molecular signatures and microbiome alterations of CRC disease. This chapter focuses on the application of the selected multi-omics data in the CRC field. We have shown how to integrate data and employ artificial intelligence to early detect and treat patients with CRC effectively. This chapter reviews the integrity of multi-omics data sources described and analyzed and the algorithm behavior illustrated as examples. It is also important to note that the advancement of multi-omics data exploration is affected by sample size, study design, supervised or unsupervised algorithm, model, and feature selection processes, etc. The samples tend to be hard to obtain and may have missing value features with a low mutation rate from sample collection and processing problems, sample bias, and financial constraints. However, emerging scientific technologies have driven the cost and time of sequencing down and provided portability and precision improvement that are increasingly accessible to clinicians. Scientists could design an in-depth and longitudinal follow-up study to discover the ethics for advanced clinical implementations.

11. References

[1] Polineni, T. N. S., & Seenu, A. (2025). The New Frontier of Healthcare and Industry: Subash's Expertise in Big Data and Cloud Computing for Enhanced Operational Efficiency. *Cuestiones de Fisioterapia*, *54*(2), 271-283.

[2] Maguluri, K. K., Ganti, V. K. A. T., Yasmeen, Z., & Pandugula, C. (2025, January). Progressive GAN Framework for Realistic Chest X-Ray Synthesis and Data Augmentation. In *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)* (pp. 755-760). IEEE.

[3] Koppolu, H. K. R. Deep Learning and Agentic AI for Automated Payment Fraud Detection: Enhancing Merchant Services Through Predictive Intelligence.

[4] Nampalli, R. C. R., & Adusupalli, B. (2024). AI-Driven Neural Networks for Real-Time Passenger Flow Optimization in High-Speed Rail Networks. *Nanotechnology Perceptions*, 334-348.

[5] Chakilam, C. (2022). Generative AI-Driven Frameworks for Streamlining Patient Education and Treatment Logistics in Complex Healthcare Ecosystems. *Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2*, *3719*.

[6] Sriram, H. K. (2023). Harnessing AI Neural Networks and Generative AI for Advanced Customer Engagement: Insights into Loyalty Programs, Marketing Automation, and Real-Time Analytics. *Educational Administration: Theory and Practice*, *29*(4), 4361-4374.

[7] Burugulla, J. K. R. (2025). Enhancing Credit and Charge Card Risk Assessment Through Generative AI and Big Data Analytics: A Novel Approach to Fraud Detection and Consumer Spending Patterns. *Cuestiones de Fisioterapia*, *54*(4), 964-972.

[8] Chava K. Dynamic Neural Architectures and AI-Augmented Platforms for Personalized Direct-to-Practitioner Healthcare Engagements. J Neonatal Surg [Internet]. 2025Feb.24 [cited 2025Mar.24];14(4S):501-10. Available from: https://www.jneonatalsurg.com/index.php/jns/article/view/1824

[9] Challa, K. (2024). Neural Networks in Inclusive Financial Systems: Generative AI for Bridging the Gap Between Technology and Socioeconomic Equity. *MSW Management Journal*, *34*(2), 749-763.

[10] Sondinti, K., & Reddy, L. (2025). The Future of Customer Engagement in Retail Banking: Exploring the Potential of Augmented Reality and Immersive Technologies. *Available at SSRN 5136025*.

[11] Malempati, M., & Rani, P. S. Autonomous AI Ecosystems for Seamless Digital Transactions: Exploring Neural Network-Enhanced Predictive Payment Models.

[12] Pallav Kumar Kaulwar. (2023). Tax Optimization and Compliance in Global Business Operations: Analyzing the Challenges and Opportunities of International Taxation Policies and Transfer Pricing. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 150-181.

[13] Vankayalapati, R. K. (2025). Architectural foundations of hybrid cloud. *The Synergy Between Public and Private Clouds in Hybrid Infrastructure Models: Real-World Case Studies and Best Practices*, 17.

[14] Nuka, S. T. (2025). Leveraging AI and Generative AI for Medical Device Innovation: Enhancing Custom Product Development and Patient Specific Solutions. *Journal of Neonatal Surgery*, *14*(4s).

[15] Rao Suura S. Agentic AI Systems in Organ Health Management: Early Detection of Rejection in Transplant Patients. J Neonatal Surg [Internet]. 2025Feb.24 [cited 2025Mar.24];14(4S):490-50.

[16] Kannan, S. (2025). Transforming Community Engagement with Generative AI: Harnessing Machine Learning and Neural Networks for Hunger Alleviation and Global Food Security. *Cuestiones de Fisioterapia*, *54*(4), 953-963.

[17] Srinivas Kalisetty, D. A. S. Leveraging Artificial Intelligence and Machine Learning for Predictive Bid Analysis in Supply Chain Management: A Data-Driven Approach to Optimize Procurement Strategies.

[18] Challa, S. R. Diversification in Investment Portfolios: Evaluating the Performance of Mutual Funds, ETFs, and Fixed Income Securities in Volatile Markets.

[19] Vamsee Pamisetty. (2023). Intelligent Financial Governance: The Role of AI and Machine Learning in Enhancing Fiscal Impact Analysis and Budget Forecasting for Government Entities. Journal for ReAttach Therapy and Developmental Diversities, 6(10s(2), 1785–1796. https://doi.org/10.53555/jrtdd.v6i10s(2).3480

[20] Komaragiri, V. B. (2022). AI-Driven Maintenance Algorithms For Intelligent Network Systems: Leveraging Neural Networks To Predict And Optimize Performance In Dynamic Environments. *Migration Letters*, *19*, 1949-1964.

[21] Annapareddy, V. N., & Rani, P. S. AI and ML Applications in RealTime Energy Monitoring and Optimization for Residential Solar Power Systems.